ConVL: Interpretable Concept-Guided Vision-Language MIL for Survival Analysis in Whole Slide Images

Junjian Li[1], Hulin Kuang[1,*], Jin Liu[1], Hailin Yue[1], Mengshen He[1], Xinyu Li[1], and Jianxin Wang[1,*]

[1] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, Hunan, China
hulinkuang@csu.edu.cn, jxwang@mail.csu.edu.cn

Abstract. Whole Slide Images (WSIs) pose significant challenges for survival modeling due to their complex tissue heterogeneity. While Multiple Instance Learning (MIL) shows promise, it lacks semantic transparency, failing to identify the underlying pathological concepts driving predictions. Moreover, although vision-language models introduce semantic priors, they rely on static text prompts that remain invariant during inference across all patients, failing to capture the morphological variations of phenotypes. To bridge this gap, we propose ConVL, a novel, interpretable concept-guided vision-language MIL framework that explicitly grounds survival prediction in explicit pathological knowledge. ConVL introduces three novel mechanisms: (1) Instance Conditioned Prompt Learning, which dynamically adapts concept prototypes to slide-specific visual contexts, enhancing robustness to tissue heterogeneity; (2) a Synergistic Dual-Stream Mechanism combining data-driven visual features with a knowledge-driven Concept Stream, utilizing Concept-Guided Cross Attention to quantify pathological attributes; and (3) a Distribution Alignment Loss that enforces consistency between the visual feature manifold and the concept-based prediction manifold. We evaluate ConVL across nine real-world cancer cohorts from TCGA. ConVL not only achieves state-of-the-art survival prediction performance but also provides fine-grained, concept-driven interpretability. Code is available at https://github.com/junjianli106/ConVL.

Keywords: Computational Pathology · Whole Slide Images · Survival Analysis · Vision-Language Model.

1 Introduction

Deep learning has shown strong potential in a broad range of medical image analysis tasks [2,25,23,8,26,9]. Whole Slide Images (WSIs) are the clinical gold standard for cancer diagnosis, encapsulating rich prognostic information that spans

* Corresponding author.

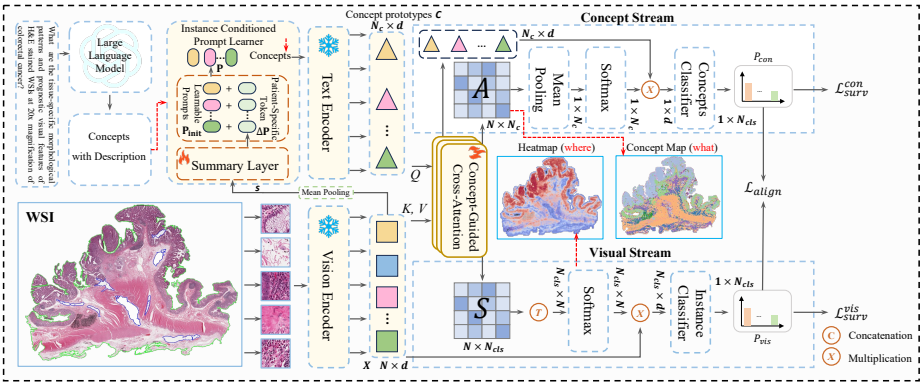

Fig. 1: Overview of the ConVL for WSI survival prediction. ConVL integrates Instance Conditioned Prompt Learning for dynamic concept adaptation, a Synergistic Dual-Stream Mechanism for coupling visual context with semantic evidence, and a Distribution Alignment Loss for dual-branch consistency.

cellular morphology, tissue architecture, and the tumor microenvironment. However, the gigapixel resolution of WSIs, coupled with the scarcity of fine-grained annotations and significant intra-class heterogeneity, poses severe challenges for computational survival modeling. Tumors of the same subtype often exhibit substantial morphological variations across different patients, which complicates the extraction of robust prognostic features [21].

Multiple Instance Learning (MIL) has become the dominant paradigm for WSI analysis by aggregating features from local patches into slide-level representations [1,16,7,3,12,5,10,24,19,20,4]. Attention-based MIL models [1,22] focus on localizing discriminative regions, **answering where the model looks**, while Transformer-based approaches [16,17] capture broader contextual dependencies. Despite these advances, existing MIL methods face a critical interpretability bottleneck. They rely on purely visual statistical correlations without explicit semantic guidance. Consequently, they can highlight high-attention regions but fail to explain the underlying pathological rationale. **They cannot answer what morphological phenotype is present.** This absence of semantic anchors makes it difficult to capture consistent prognostic patterns across morphological variations, limiting model generalization when facing highly heterogeneous samples.

Recent advances in Vision-Language Models (VLMs) have inspired prompt-based WSI analysis [15,13]. For instance, VLSA [11] utilizes language-encoded prognostic priors to guide MIL aggregation. However, these methods primarily rely on learned static prompts that remain invariant during inference across all patients. Such static representations fail to adapt to the significant morphological heterogeneity inherent in cancer tissue, where the visual manifestation of the same concept (e.g., 'tumor infiltration') can vary drastically between slides.

Consequently, applying such static prompts often fails to accurately capture patient-specific pathological nuances, inevitably leading to feature misalignment.

To address these limitations, we propose ConVL, an concept-guided vision-language MIL framework that deeply couples pathological domain knowledge with visual features. Specifically, ConVL employs a dual-branch architecture integrated with three novel mechanisms. First, Instance Conditioned Prompt Learning generates patient-specific context offsets to adaptively refine concept prototypes to mitigate tissue heterogeneity. Second, a Synergistic Dual-Stream Mechanism combining data-driven visual features with a knowledge-driven Concept Stream. The visual branch functions as a data-driven feature aggregator to capture latent prognostic cues, while the concept branch explicitly assesses the histological burden of pathological phenotypes. Third, a Distribution Alignment Loss ensures consistent prediction patterns between the visual and concept branches. Experiments on nine real-world TCGA cancer cohorts demonstrate that ConVL not only delivers competitive prognostic performance but also provides intelligible explanations by explicitly associating tissue regions with specific pathological concepts.

2 Methodology

2.1 Overview of ConVL

To address the semantic gap and tissue heterogeneity in WSI-based survival modeling, we propose ConVL, an interpretable concept-guided vision-language MIL framework. As shown in Fig. 1, ConVL employs a synergistic dual-branch architecture, comprising a Visual Branch that aggregates visual features using standard attention, and a Concept Branch that grounds predictions in learnable pathological prototypes to achieve semantic evidence localization.

2.2 WSI Preprocessing and Concept Construction

For each patient, the WSIs are segmented to remove background and tiled into non-overlapping patches at 20× magnification, forming a bag of instances $\mathbf{X} = \{x_n\}_{n=1}^{N} \in \mathbb{R}^{N \times d}$. Each patch x_n is a d-dimensional feature embedding extracted via the image encoder of a pre-trained pathology-specific VLM (e.g., CONCH).

To establish a prognostic semantic prior, we prompt a Large Language Model (e.g., Gemini 3 Pro) with a domain-specific inquiry: 'What are the tissue-specific morphological patterns and prognostic visual features of H&E stained WSIs at 20× magnification for [Cancer Type]?' This process yields a set of pathological concepts describing histological evidence (e.g., 'solid sheets of tumor cells'). These texts serve as the foundational semantic descriptions for our prompt learning module. The full list of concepts is available in our public repository.

2.3 Instance Conditioned Prompt Learning

Prior prompt-based MIL methods (e.g., VLSA) rely on fixed and static concept prototypes shared across all patients. While providing a robust semantic prior,

these static prompts struggle with severe tissue heterogeneity, as visual manifestations of the same pathological concept vary substantially across slides. Consequently, fixed prototypes fail to capture patient-specific morphological variations, leading to domain misalignment during inference. To address this, we introduce Instance Conditioned Prompt Learning (ICPL), which dynamically refines concept prototypes using global slide-level context. By generating tailored concepts for each WSI, even slides from the same diagnostic class retrieve distinct concept prototypes. This enables fine-grained adaptation, effectively shifting the prompt paradigm from static priors to patient-specific knowledge.

Specifically, we first introduce a set of Learnable Prompt Vectors $\mathbf{P}_{init} \in \mathbb{R}^{L\times d}$. To inject patient-specific information, we then derive a global contextual representation $\mathbf{s} \in \mathbb{R}^d$ by aggregating all instance embeddings within a bag. We employ a Summary Layer $\mathcal{M}$ (acting as a transformation bridge) to map the visual summary $\mathbf{s} = \frac{1}{N}\sum_{i=1}^{N} x_i$ into the prompt embedding space. The transformation is formulated as:

$$\Delta\mathbf{P} = \text{Reshape}(\mathbf{W}_2(\sigma(\mathbf{W}_1\mathbf{s}))), \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{d}{2}\times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{(L\cdot d)\times\frac{d}{2}}$ are learnable projection matrices, σ denotes the ReLU activation, and $\Delta\mathbf{P} \in \mathbb{R}^{L\times d}$ represents the patient-specific tokens. These tokens serve as dynamic offsets that capture the unique morphological characteristics of the current patient. The prompt prototypes are then formed by:

$$\mathbf{P} = \mathbf{P}_{init} + \Delta\mathbf{P}. \tag{2}$$

Finally, each text concept T_i is tokenized and mapped into a sequence of word embeddings e_i via the embedding layer of the VLM text encoder. We then concatenate the dynamic prompt prototypes P with the concept embeddings e_i, formulating the input text sequence as $[P, e_i]$. Here, each concept description provides the pathological semantic content, while the patient-specific tokens generated from the WSI-level summary inject slide-level visual context. The concatenated sequence is then fed into the subsequent Transformer layers of the VLM text encoder, producing a dynamic concept prototype $C_i \in \mathbb{R}^d$ that is explicitly grounded in the specific visual context of the patient.

2.4 Synergistic Dual-Stream Survival Prediction

To leverage both holistic visual context and fine-grained semantic evidence, ConVL employs a synergistic dual-stream architecture. We introduce a Concept-Guided Cross-Attention (CGA) module to align the bag of instance embeddings $\mathbf{X} \in \mathbb{R}^{N\times d}$ with patient-specific concept prototypes $\mathbf{C} \in \mathbb{R}^{N_c\times d}$. Formally, we treat $\mathbf{C}$ as queries and $\mathbf{X}$ as keys and values; the CGA module computes a concept–instance affinity matrix $\mathbf{A} \in \mathbb{R}^{N\times N_c}$ via scaled dot-product attention. Each entry $\mathbf{A}_{j,i}$ quantifies the activation of the i-th concept at the j-th patch. From $\mathbf{A}$, we obtain patch-level class logits $\mathbf{S} \in \mathbb{R}^{N\times N_{cls}}$ by a linear projection over the concept dimension (N_{cls} is number of survival risk bins).

Visual Stream. Unlike standard MIL approaches that derive attention weights solely from visual features, this stream leverages the concept-guided patch-class logits $\mathbf{S}$ to direct instance aggregation. Specifically, per-class attention over the patches is obtained by applying a Softmax function along the instance dimension, denoted as $\boldsymbol{\Theta} = \text{Softmax}(\mathbf{S}^\top)$. The bag-level visual representation $\mathbf{h}_{vis}$ is then generated by weighting the instance embeddings $\mathbf{X}$ with $\boldsymbol{\Theta}$, followed by mean pooling across the class dimension:

$$\mathbf{h}_{vis} = \text{Mean}\big(\boldsymbol{\Theta}\mathbf{X}\big) = \text{Mean}\big(\text{Softmax}(\mathbf{S}^\top)\mathbf{X}\big), \quad \text{and} \quad P_{vis} = f_{cls}^{vis}(\mathbf{h}_{vis}). \tag{3}$$

Here, the Mean operator computes the average across the N_{cls} class-conditioned aggregations. Thus, $\mathbf{h}_{vis}$ aggregates patch features weighted by class-wise semantic priors, capturing holistic prognostic cues in relevant regions beyond the predefined concept set.

Concept Stream. This stream interprets the WSI by evaluating concept prevalence. Specifically, we apply global average pooling to the affinity matrix $\mathbf{A}$ across the instance dimension, yielding a concept distribution $\mathbf{v} \in \mathbb{R}^{N_c}$. Here, each element $v_i = \frac{1}{N}\sum_{j=1}^{N} \mathbf{A}_{j,i}$ explicitly quantifies the global burden of the i-th concept across the entire tissue. The slide-level concept representation is then formulated by weighting the concept prototype embeddings with this distribution:

$$\mathbf{h}_{con} = \mathbf{Z}^\top \mathbf{v}, \quad \text{and} \quad P_{con} = f_{cls}^{con}(\mathbf{h}_{con}), \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{N_c \times d}$ denotes the concept prototype embeddings. Consequently, $\mathbf{h}_{con}$ characterizes the WSI as a weighted combination of clinical concepts, effectively grounding the final survival prediction in explicit phenotypic evidence.

2.5 Distribution Alignment Loss

To ensure semantic consistency between modalities, we introduce a Distribution Alignment Loss (DAL) as a mutual regularization constraint. This mechanism explicitly aligns the interpretable concept-based predictions with the holistic visual representations, enforcing a robust consensus between the two distinct feature manifolds. Formally, since the model predicts survival probabilities across discrete time intervals, we compute the alignment loss using the symmetric Kullback-Leibler (KL) divergence between the predictive distributions P_{vis} and P_{con}:

$$\mathcal{L}_{align} = \text{KL}(P_{vis} || P_{con}) + \text{KL}(P_{con} || P_{vis}). \tag{5}$$

The overall training objective is formulated as a weighted sum of the task-specific losses and the alignment constraint:

$$\mathcal{L} = \alpha \mathcal{L}_{surv}^{vis} + (1-\alpha)\mathcal{L}_{surv}^{con} + \beta \mathcal{L}_{align}, \tag{6}$$

where $\mathcal{L}_{surv}^{vis}$ and $\mathcal{L}_{surv}^{con}$ denote the discrete-time Negative Log-Likelihood (NLL) losses for the visual and concept branches, respectively. The hyperparameter β controls the strength of the alignment regularization. Crucially, $\mathcal{L}_{align}$ enforces a

Table 1: Survival prediction performance results (C-index) of different methods. Best results are highlighted in **bold**, and second-best results are underlined.

Methods	BLCA	BRCA	CO&RE	LUSC	KIRC	SKCM	LIHC	SARC	UCEC	Average
ABMIL (18' ICML)	$0.575_{0.068}$	$0.563_{0.079}$	$0.650_{0.038}$	$0.556_{0.030}$	$0.703_{0.041}$	$0.581_{0.036}$	$0.682_{0.034}$	$0.565_{0.082}$	$0.717_{0.103}$	0.621
TransMIL (21' NeurIPS)	$0.564_{0.070}$	$0.607_{0.071}$	$0.668_{0.083}$	$0.517_{0.026}$	$0.711_{0.047}$	$0.547_{0.033}$	$0.685_{0.046}$	$0.595_{0.065}$	$0.710_{0.066}$	0.623
ILRA (23' ICLR)	$0.597_{0.026}$	$0.590_{0.081}$	$0.645_{0.078}$	$0.541_{0.067}$	$0.698_{0.046}$	$0.524_{0.044}$	$0.660_{0.032}$	$0.525_{0.054}$	$0.712_{0.067}$	0.610
PseMix (24' TMI)	$0.531_{0.055}$	$0.561_{0.116}$	$0.669_{0.049}$	$0.509_{0.057}$	$0.719_{0.041}$	$0.584_{0.028}$	$0.670_{0.017}$	$0.513_{0.089}$	$0.686_{0.062}$	0.605
RRTMIL (24' CVPR)	$0.579_{0.018}$	$0.577_{0.120}$	$0.698_{0.007}$	$0.520_{0.037}$	$0.689_{0.041}$	$0.581_{0.029}$	$0.679_{0.050}$	$0.550_{0.041}$	$0.675_{0.080}$	0.616
WiKG (24' CVPR)	$0.599_{0.031}$	$0.534_{0.064}$	$0.670_{0.068}$	$0.544_{0.018}$	$0.718_{0.044}$	$0.576_{0.022}$	$0.642_{0.029}$	$0.575_{0.032}$	$0.707_{0.085}$	0.618
ACMIL (24' ECCV)	$0.590_{0.047}$	$\underline{0.613_{0.052}}$	$0.676_{0.042}$	$0.543_{0.014}$	$0.714_{0.055}$	$0.570_{0.027}$	$0.688_{0.043}$	$0.484_{0.026}$	$\underline{0.731_{0.066}}$	0.623
MiCo (25' MICCAI)	$0.600_{0.055}$	$0.607_{0.072}$	$\mathbf{0.714}_{0.069}$	$0.552_{0.048}$	$\underline{0.738_{0.046}}$	$0.574_{0.022}$	$\mathbf{0.690}_{0.031}$	$0.529_{0.069}$	$0.722_{0.087}$	$\underline{0.636}$
TopMIL (23' NeurIPS)	$0.604_{0.045}$	$0.532_{0.046}$	$0.696_{0.040}$	$\underline{0.572_{0.023}}$	$0.733_{0.046}$	$\underline{0.601_{0.021}}$	$0.675_{0.029}$	$0.575_{0.020}$	$0.720_{0.092}$	0.634
VLSA (25' ICLR)	$\underline{0.625_{0.020}}$	$0.581_{0.051}$	$0.683_{0.055}$	$0.531_{0.026}$	$0.712_{0.057}$	$0.547_{0.029}$	$0.657_{0.028}$	$\underline{0.596_{0.015}}$	$0.714_{0.075}$	0.627
ConVL (Ours)	$\mathbf{0.629}_{0.021}$	$\mathbf{0.618}_{0.076}$	$\underline{0.709_{0.051}}$	$\mathbf{0.579}_{0.024}$	$\mathbf{0.746}_{0.052}$	$\mathbf{0.612}_{0.025}$	$\underline{0.688_{0.037}}$	$\mathbf{0.608}_{0.059}$	$\mathbf{0.734}_{0.080}$	**0.658**

bidirectional consensus: it constrains the explainable concept predictions to remain faithful to the robust visual evidence, while simultaneously regularizing the visual stream with structured semantic priors. This synergy effectively mitigates modality-specific overfitting and ensures that the final prognosis is supported by consistent cross-modal reasoning.

3 Experiments

3.1 Datasets and Implementation Details

We evaluate ConVL on nine TCGA cohorts for survival prediction: BLCA, BRCA, CO&RE, LUSC, KIRC, SKCM, LIHC, SARC, and UCEC. We compare ConVL against representative MIL baselines and state-of-the-art survival models, including ABMIL [1], TransMIL [16], ILRA [18], PseMix [12], RRT-MIL [17], WiKG [3], ACMIL [22], and MiCo [6], as well as recent VLM-based MIL methods such as TopMIL [14] and VLSA [11]. For feature extraction, we follow CONCH [13] to generate 448×448 patches at 20× magnification. We evaluate our method using a 4-fold cross-validation scheme, randomly splitting the data into 60% training, 15% validation, and 25% test sets for each fold. For fair comparison, all baselines utilized the same CONCH features network and data splits. We train for 200 epochs with batch size 1, learning rate 2e-4, and Ranger optimizer. Early stopping is applied with patience 8 based on validation loss. Following prior work [6], the number of survival risk bins N_{cls} is set to 4. Furthermore, we set the alignment weight $\beta = 0.1$ and prompt length $L = 16$, while the number of concepts N_c is tailored per dataset. All experiments are conducted on a single 3080Ti GPU.

3.2 Results and Discussion

Comparison with State-of-the-Art Methods. As summarized in Table 1, ConVL establishes a new state-of-the-art benchmark in WSI-based survival prediction, achieving the highest average C-index of 0.658. This performance represents a comprehensive improvement of 2.2% over the strongest competitor,

Table 2: Ablation study for ConVL.

Methods	BLCA	BRCA	CO&RE	LUSC	KIRC	SKCM	LIHC	SARC	UCEC	Average
w/o DAL	$\underline{0.617_{0.054}}$	$\underline{0.617_{0.072}}$	$0.670_{0.080}$	$0.554_{0.017}$	$\underline{0.742_{0.045}}$	$\underline{0.609_{0.024}}$	$0.682_{0.041}$	$\underline{0.564_{0.073}}$	$0.710_{0.077}$	$\underline{0.641}$
w/o CGA	$0.599_{0.018}$	$0.590_{0.080}$	$\underline{0.681_{0.081}}$	$\underline{0.576_{0.028}}$	$0.733_{0.041}$	$0.602_{0.020}$	$\mathbf{0.697}_{0.031}$	$0.550_{0.091}$	$\underline{0.723_{0.075}}$	0.639
w/o ICPL	$0.550_{0.118}$	$0.607_{0.076}$	$0.680_{0.040}$	$0.554_{0.018}$	$0.735_{0.038}$	$0.606_{0.022}$	$0.685_{0.026}$	$0.545_{0.025}$	$0.717_{0.066}$	0.631
ConVL	$\mathbf{0.629}_{0.021}$	$\mathbf{0.618}_{0.076}$	$\mathbf{0.709}_{0.051}$	$\mathbf{0.579}_{0.024}$	$\mathbf{0.746}_{0.052}$	$\mathbf{0.612}_{0.025}$	$\underline{0.688_{0.037}}$	$\mathbf{0.608}_{0.059}$	$\mathbf{0.734}_{0.080}$	**0.658**

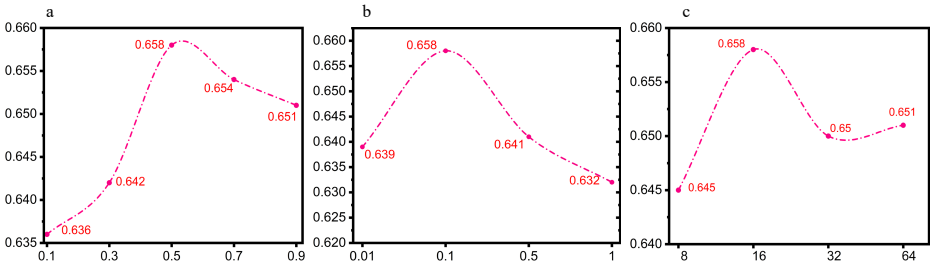

Fig. 2: Hyperparameter sensitivity analysis on the average C-Index across nine datasets. We illustrate the impact of (a) the concept fusion weight α, (b) the distribution alignment weight β, and (c) the prompt length L. The red markers denote the mean C-Index.

MiCo (0.636), and outperforms recent baselines such as VLSA (0.627). Notably, ConVL demonstrates remarkable robustness and generalizability, ranking first in seven out of the nine evaluated cohorts. The performance gains are particularly pronounced in the SARC and SKCM cohorts, where our ConVL outperforms the second-best approaches by margins of 1.2% and 1.1%, respectively. Furthermore, even in cohorts where ConVL does not secure the top rank (i.e., CO&RE and LIHC), it remains highly competitive, trailing the leading methods by negligible margins of 0.005 and 0.002.

Ablation Study. To validate the contribution of each component, we conduct an ablation study as shown in Table 2. The results demonstrate that all proposed modules are essential for optimal performance. The most significant degradation occurs when replacing the Instance Conditioned Prompt Learning (w/o ICPL) with static prompts, which causes the average C-index to drop from 0.658 to 0.631. This confirms the critical role of our adaptive mechanism in handling tissue heterogeneity. Similarly, substituting the Concept-Guided Cross-Attention with standard Gated Attention (w/o CGA) or removing the Distribution Alignment Loss (w/o DAL) leads to performance declines to 0.639 and 0.641, respectively. These findings verify that semantic-aware evidence retrieval and dual-branch consistency are indispensable for robust survival prediction.

Hyperparameter Analysis. As illustrated in Fig. 2, we conduct sensitivity analyses on three key hyperparameters. The model achieves optimal performance with a distribution alignment weight of $\beta = 0.1$, a prototype count of $L = 16$, and a balanced concept fusion weight of $\alpha = 0.5$. Deviations from these values lead to performance degradation, which validates the necessity of moderate alignment

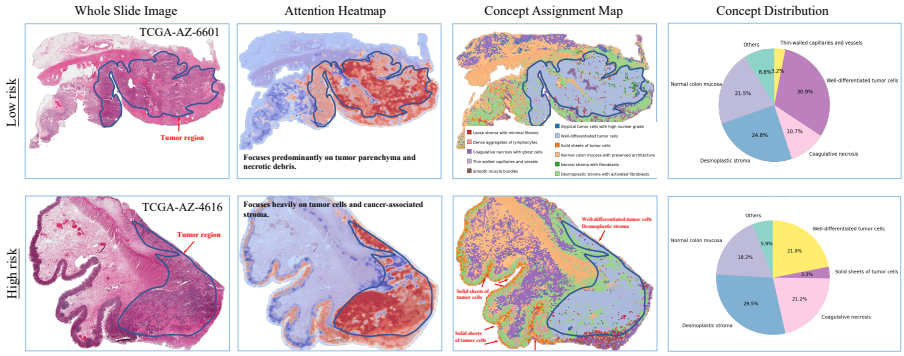

Fig. 3: Interpretability Analysis. Visual comparison of ConVL against standard attention mechanisms. Columns from left to right display the Original WSI, standard Attention Heatmap, Concept Assignment Map, and the quantified Concept Distribution pie chart. The visualizations contrast a Low-Risk patient (top row) characterized by well-differentiated cells with a High-Risk patient (bottom row) driven by malignant desmoplastic stroma.

regularization, sufficient context capacity, and the equal contribution of visual and textual modalities.

Interpretability Analysis. To validate the semantic grounding of ConVL, Fig. 3 compares the original WSIs and standard Attention Heatmaps alongside our Concept Assignment Maps and slide-level Concept Distributions. Unlike standard heatmaps that only indicate where the model looks, these visualizations explicitly reveal what phenotypes drive the prediction, with the distribution pie charts quantifying the overall morphological burden across the entire WSI. In the High-Risk case (bottom row), the model identifies tumor cells intermingled with *Desmoplastic stroma* (light green). The concept distribution (Column 4) corroborates this by assigning the highest importance (29.5%) to this malignant stroma. As confirmed by an expert pathologist, this signifies a desmoplastic reaction at the invasive front, which is a hallmark of aggressive tumors and a primary high-risk indicator. Conversely, the Low-Risk case (top row) is dominated by *Well-differentiated tumor cells* (30.9%) and *Normal colon mucosa* (21.5%). Here, the desmoplastic stroma burden is notably reduced, outweighed by preserved architectures and localized *Coagulative necrosis* (10.7%). This reflects a confined, expansile growth pattern rather than a highly infiltrative one. Furthermore, this localized necrosis resembles dirty necrosis, a morphological feature often associated with high microsatellite instability and favorable outcomes.

4 Conclusion

In this paper, we present ConVL, a novel, interpretable concept-guided vision-language framework designed to overcome severe intra-class heterogeneity and

the lack of semantic transparency in WSI-based survival analysis. To shift from static semantic priors to patient-specific knowledge, we introduce Instance Conditioned Prompt Learning, which dynamically adapts pathological concepts to unique slide-level contexts. This is coupled with a Concept-Guided Cross Attention module to explicitly retrieve and quantify discriminative semantic evidence. Furthermore, a Distribution Alignment Loss ensures robust consensus between visual and concept-driven prediction manifolds. Extensive evaluations across nine TCGA cancer cohorts demonstrate that ConVL not only establishes a new SOTA in prognostic accuracy but also delivers fine-grained, clinically actionable interpretability.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (Nos. U25A20449 and U24A20256), and by the Key R&D Program of Xinjiang Uygur Autonomous Region under Grant No. 2024B03039-3. This work was also partially carried out using computing resources at the High Performance Computing Center of Central South University.

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

References

1. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
2. Kui, X., Xiao, Q., Li, Q., Ji, Z., Zhang, J., Zou, B.: Flip distribution alignment vae for multi-phase mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 208–218. Springer (2025)
3. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11323–11332 (2024)
4. Li, J., Kuang, H., Liu, J., Yue, H., He, M., Wang, J.: Universal-to-specific: Dynamic knowledge-guided multiple instance learning for few-shot whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26614–26623 (2026)
5. Li, J., Kuang, H., Liu, J., Yue, H., Wang, J.: Ca2cl: Cluster-aware adversarial contrastive learning for pathological image analysis. IEEE Journal of Biomedical and Health Informatics (2025)
6. Li, J., Liu, J., Kuang, H., Yue, H., He, M., Wang, J.: Mico: Multiple instance learning with context-aware clustering for whole slide image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376–385. Springer (2025)
7. Li, J., Liu, J., Yue, H., Cheng, J., Kuang, H., Bai, H., Wang, Y., Wang, J.: Darc: Deep adaptive regularized clustering for histopathological image classification. Medical image analysis **80**, 102521 (2022)
8. Lin, J., Zhu, C., Kneuertz, P.J., Bai, Y., Xue, Y.: When models learn to ask why: Adaptive causal reasoning for trustworthy medical vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5556–5568 (2026)

9. Liu, M., Cai, C., Chen, D., Li, J., Li, J., Ming, W., Xu, J.: Data-and knowledge-driven multimodal learning in computational pathology: A comprehensive survey. EngMedicine **3**(2), 100132 (2026)
10. Liu, M., Cai, C., Li, J., Xu, P., Li, J., Ma, J., Xu, J.: Murrenet: Modeling holistic multimodal interactions between histopathology and genomic profiles for survival prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 396–406. Springer (2025)
11. Liu, P., Ji, L., Gou, J., Fu, B., Ye, M.: Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In: The Thirteenth International Conference on Learning Representations (2025)
12. Liu, P., Ji, L., Zhang, X., Ye, F.: Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. IEEE Transactions on Medical Imaging **43**(5), 1841–1852 (2024)
13. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
14. Qu, L., Fu, K., Wang, M., Song, Z., et al.: 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**, 67551–67564 (2023)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
16. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
17. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., 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)
18. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)
19. Zhang, D., Ge, J., Liu, J., Wang, C., Gong, T., Gao, Z., Li, C.: Stadis: Stability distance to detecting out-of-distribution data in computational pathology. Medical Image Analysis p. 103774 (2025)
20. Zhang, D., Gong, Z., Pang, X., Liu, J., Lu, J., Cui, H., Ge, J., Zeng, Z., Yi, K., Li, Y., et al.: Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21078–21088 (2026)
21. Zhang, Q., Lou, Y., Yang, J., Wang, J., Feng, J., Zhao, Y., Wang, L., Huang, X., Fu, Q., Ye, M., et al.: Integrated multiomic analysis reveals comprehensive tumour heterogeneity and novel immunophenotypic classification in hepatocellular carcinomas. Gut **68**(11) (2019)
22. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European conference on computer vision. pp. 125–143. Springer (2024)
23. Zhu, C., Lin, Y., Chen, S., Wang, Y., Lin, J.: Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13916–13924 (2026)

24. Zhu, C., Lin, Y., Shao, J., Lin, J., Wang, Y.: Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9196–9205 (2025)
25. Zhu, C., Shao, J., Lin, J., Wang, Y., Wang, J., Tang, J., Li, K.: fmri2ges: Co-speech gesture reconstruction from fmri signal with dual brain decoding alignment. IEEE Transactions on Circuits and Systems for Video Technology (2025)
26. Zhu, C., Zeng, J., Jiang, J., Lin, J., Wang, Y.: Medsynapse-v: Bridging visual perception and clinical intuition via latent memory evolution. arXiv preprint arXiv:2604.26283 (2026)