Bridging Radiology and Pathology Foundation Models via Concept-Based Multimodal Co-Adaptation

Yihang Chen[1], Yanyan Huang[1], Fuying Wang[2], Maximus Yeung[1], Yuming Jiang[3], Shujun Wang[4], Lequan Yu[1]

[1]The University of Hong Kong, [2]Stanford University, [3]Wake Forest University School of Medicine, [4]The Hong Kong Polytechnic University

ICLR

Introduction

➢Background and Motivation

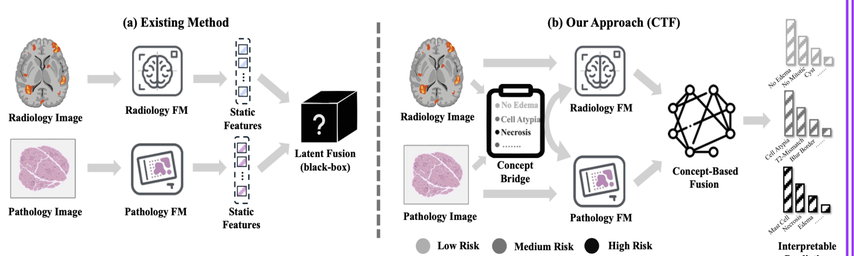
- **Problem:** Clinical workflows rely on joint diagnosis from radiology and pathology. These two modalities provide complementary information, e.g., tumor boundaries from CT/MRI and cellular-level malignancy from whole-slide images (WSIs).
- Current foundation models are typically designed for a single modality, limiting cross-modal synergy for patient-level prediction.

➢ Limitations of Existing Methods:

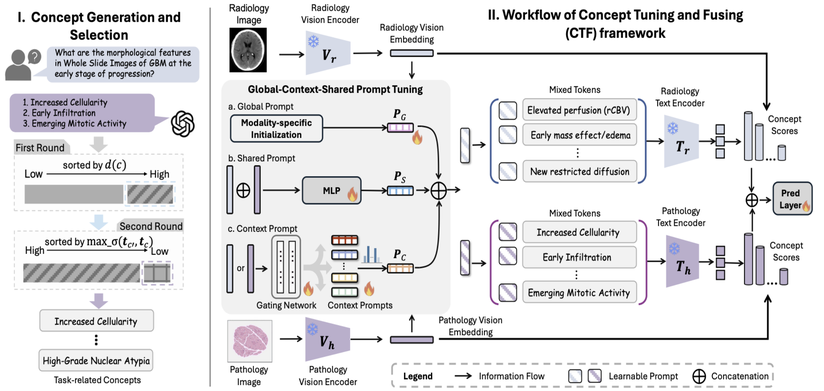
- Existing approaches use frozen VLMs as static feature extractors with simple fusion (e.g., concatenation), limiting adaptability to downstream tasks and interplay between modalities.
- Full fine-tuning of large VLMs is computationally expensive and often domain-confined, weakening knowledge transfer across domains.
- Both approaches result in "black-box" systems lacking clinical interpretability.

➢ Contributions:

- We propose CTF, a novel framework using medically-relevant concepts as a dynamic *and* interpretable bridge to fuse distinct medical VLMs.
- We introduce the Global-Context-Shared Prompt (GCSP), a parameter-efficient cross-domain conditioning.
- Extensive experiments on four datasets show our model's superior performance in survival analysis and cancer grading, with strong interpretability and clinically plausible concept-level reasoning.



Method



➢Prognostic Concept Selection

- We generate medical concepts via LLMs and select a compact subset using Mutual Information (MI) + diversity maximization.

➢ Cross-Domain Concept Co-Adaptation (GCSP)

- Global prompt (P_G) for overall downstream task adaptation.
- Context (P_C) for cross-domain guidance via MoE-style gating.
- Shared (P_S) for patient-level synergy from concatenated features.

➢ Fusion and Interpretable Prediction

- **Concept Scores:** Per-domain concept alignment scores (cosine similarities) are concatenated and fed into the prediction head.

➢ Dataset Description

- **TCGA-LGG:** Lower-Grade Glioma (173 cases) for 3-way cancer grading (w/ TCGA-GBM as GBMLGG) and survival prediction.
- **TCGA-GBM:** Glioblastoma Multiforme (186 cases) for 3-way cancer grading (as GBMLGG) and survival prediction.
- **Center1-GC:** In-house Gastric tumor (683 cases) for 5-way cancer grading and survival prediction.
- **Center2-CHS:** In-house Chondrosarcoma (76 cases) for 5-way cancer grading.

Experimental Results

➢Quantitative Results

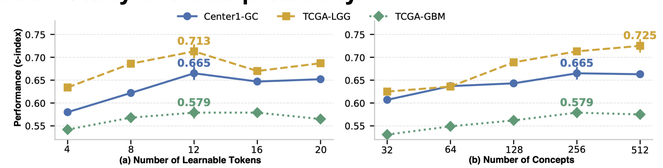
- **Survival Prediction (C-index)**

Model	TCGA-LGG	TCGA-GBM	Center1-GC
Unimodal Baselines			
Radiology-Only	0.598 ± 0.128	0.477 ± 0.055	0.614 ± 0.052
ABMIL (Ilse et al., 2018)	0.669 ± 0.101	0.480 ± 0.093	0.590 ± 0.030
CLAM (Lu et al., 2021)	0.689 ± 0.108	0.497 ± 0.068	0.631 ± 0.060
TransMIL (Shao et al., 2021)	0.682 ± 0.121	0.503 ± 0.055	0.613 ± 0.066
ACMIL (Zhang et al., 2024b)	0.678 ± 0.142	0.519 ± 0.057	0.628 ± 0.083
Multimodal Latent Fusion Baselines			
Concat-Fusion	0.674 ± 0.112	0.515 ± 0.070	0.626 ± 0.048
Cross-Attention	0.685 ± 0.108	0.527 ± 0.068	0.631 ± 0.060
MOTCAT (Xu & Chen, 2023)	0.571 ± 0.080	0.563 ± 0.108	0.622 ± 0.040
PIBD (Zhang et al., 2024a)	0.687 ± 0.123	0.531 ± 0.061	0.638 ± 0.058
Multimodal Adaptive Fusion Baseline			
M4Survive (Lee et al., 2025)	0.709 ± 0.112	0.545 ± 0.072	0.642 ± 0.065
CTF (Ours)	**0.713 ± 0.103**	**0.579 ± 0.063**	**0.665 ± 0.061**

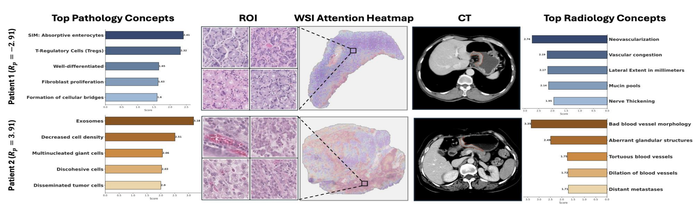
- **Cancer Grading (AUC / ACC)**

Model	TCGA-GBMLGG (3-way)		Center2-CHS (5-way)		Center1-GC (5-way)	
	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
Unimodal Baselines						
Radiology-Only	0.776 ± 0.059	0.624 ± 0.064	0.679 ± 0.069	0.429 ± 0.091	0.595 ± 0.087	0.341 ± 0.080
ABMIL (Ilse et al., 2018)	0.855 ± 0.050	0.667 ± 0.076	0.770 ± 0.092	0.493 ± 0.138	0.609 ± 0.063	0.384 ± 0.053
CLAM (Lu et al., 2021)	0.860 ± 0.048	0.681 ± 0.070	0.775 ± 0.089	0.512 ± 0.130	0.628 ± 0.055	0.390 ± 0.051
TransMIL (Shao et al., 2021)	0.864 ± 0.050	0.684 ± 0.068	0.781 ± 0.085	0.518 ± 0.128	0.625 ± 0.058	0.388 ± 0.049
ACMIL (Zhang et al., 2024b)	0.853 ± 0.046	0.680 ± 0.069	0.779 ± 0.046	0.515 ± 0.129	0.619 ± 0.062	0.389 ± 0.054
Multimodal Latent Fusion Baselines						
Concat-Fusion	0.858 ± 0.038	0.687 ± 0.062	0.805 ± 0.075	0.535 ± 0.115	0.629 ± 0.051	0.391 ± 0.048
Cross-Attention	0.868 ± 0.030	0.695 ± 0.059	0.817 ± 0.071	0.581 ± 0.110	0.635 ± 0.048	0.394 ± 0.049
MOTCAT (Xu & Chen, 2023)	0.865 ± 0.025	0.657 ± 0.053	0.826 ± 0.078	0.612 ± 0.120	0.641 ± 0.050	0.390 ± 0.052
Multimodal Adaptive Fusion Baseline						
M4Survive (Lee et al., 2025)	0.861 ± 0.031	0.691 ± 0.061	0.830 ± 0.075	0.626 ± 0.115	0.649 ± 0.052	0.390 ± 0.051
CTF (Ours)	**0.903 ± 0.028**	**0.718 ± 0.063**	**0.854 ± 0.081**	**0.698 ± 0.164**	**0.660 ± 0.049**	**0.401 ± 0.057**

➢ Ablation Study and Interpretability



- **Sensitivity:** Stable performance across prompt lengths (L=4–20). K=256 concepts yields optimal results, with fewer lacking coverage and more introducing noise.



- **Interpretability:** Low-risk patients activate benign concepts ("Well-differentiated") and high-risk patients highlight aggressive features ("Disseminated tumor cells").