Joint Biomarker and Survival Prediction via Concept-Conditioned Multimodal Slot Factorization

Zuqi Huang[1,2*], Liang Mao[1*], Wentao Li[2], Tianxiang Dong[2], Mengxin Tian[3], Zhaoqing Tang[3], and Zhongyu Li[1(✉)]

[1] School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
zhongyuli@sjtu.edu.cn
[2] School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[3] Gastrointestinal Surgery and the Gastric Cancer Center, Zhongshan Hospital, Fudan University, Shanghai, China

Abstract. In clinical oncology, biomarker status prediction and survival risk estimation are closely related but distinct objectives. While joint multi-task learning benefits from exploiting correlations among tasks, most existing methods remain unimodal and neglect task-specific heterogeneity, limiting joint modeling of local biomarker cues and global survival patterns using multimodal clinical context. To address these limitations, we propose PathoSlot, a concept-conditioned slot factorization framework that factorizes shared multimodal evidence into clinically anchored slot representations for joint biomarker and survival prediction. Specifically, PathoSlot factorizes fused features from whole slide images (WSIs) and pathology reports into a set of task-aligned slots through structured slot factorization. The factorized slots enable hierarchical prediction, with biomarker status predicted from slots encoding local diagnostic morphology and survival risk predicted from the slot capturing global prognostic patterns. To stabilize slot–task alignment, we further anchor each slot to its corresponding clinical prior via concept conditioning, constraining inter-task interference and enhancing interpretability and robustness across centers. Extensive experiments on internal and external cohorts demonstrate that PathoSlot consistently outperforms state-of-the-art methods on both biomarker and survival prediction tasks. Code is available at https://github.com/jtneuron/PathoSlot.

Keywords: Multimodal · Joint Learning · Slot Factorization · Biomarker Prediction · Survival Prediction.

1 Introduction

In contemporary clinical oncology, biomarker prediction and survival analysis play closely related but distinct roles in personalized treatment planning

* Equal contribution.

[19,18]. Clinically established biomarkers are routinely assessed for their association with disease progression and survival outcomes, making biomarker status an integral component of prognostic risk stratification [12]. However, jointly modeling biomarker status and survival outcomes is nontrivial due to their distinct representational and information requirements. Biomarker prediction is primarily driven by fine-grained morphological patterns within localized tumor regions, whereas survival analysis requires modeling global tissue architecture and patient-level prognostic context that is typically distributed across additional modalities (*e.g.*, pathology reports) [7]. This heterogeneity in granularity and modality complicates unified multi-task learning within a single predictive framework.

Deep learning–based multiple instance learning has been well explored for weakly supervised slide-level prediction from hematoxylin and eosin (H&E) WSIs [8,22,13]. However, reliance on unimodal histopathology constrains the ability to model the multifaceted determinants of disease, particularly for survival prediction [9]. Recent multimodal studies extend WSI-based modeling by incorporating additional clinical modalities through cross-modal alignment pretraining [11,3,15] and modality-aware learning that integrates non-image modalities via structured fusion or evidence aggregation of complementary signals [5,23,6]. Representative examples include TITAN [3], which aligns WSIs with pathology reports during pretraining to derive slide-level representations for diagnosis or prognosis prediction, and MSCPT [6], which enhances vision–language WSI classification by multi-scale and context-focused prompt tuning.

Joint multi-task learning optimizes multiple clinical tasks within a unified framework, allowing supervision from correlated targets to provide complementary training signals. Compared with separate training of each task, this paradigm exploits inter-target correlations and has been empirically shown to improve predictive performance and generalization [17,4]. Despite these advances, joint biomarker and survival prediction still faces several key problems: 1) Multimodal integration in joint multi-task learning remains underexplored, as most existing approaches are unimodal and fail to exploit report-derived clinical context, which is critical for survival prediction requiring holistic patient-level integration; 2) Existing methods often neglect the aforementioned task-granularity heterogeneity and lack explicit mechanisms to bridge the fundamental discrepancy between locally driven biomarker prediction and globally governed survival estimation.

In this paper, we propose PathoSlot, a concept-conditioned slot factorization framework for joint biomarker and survival prediction. Specifically, PathoSlot applies competitive reweighting to joint WSI and pathology report representations to emphasize information-dense cues. The reweighted representations are softly factorized into task-aligned slots within a unified multimodal latent space, yielding slot embeddings dedicated to biomarker prediction and survival estimation. This structured design encourages biomarker-oriented slots to focus on diagnostic-level local morphological evidence while directing survival-oriented slots to capture prognostic-level global risk patterns. To maintain consistent slot–task alignment, we incorporate concept conditioning that anchors each slot

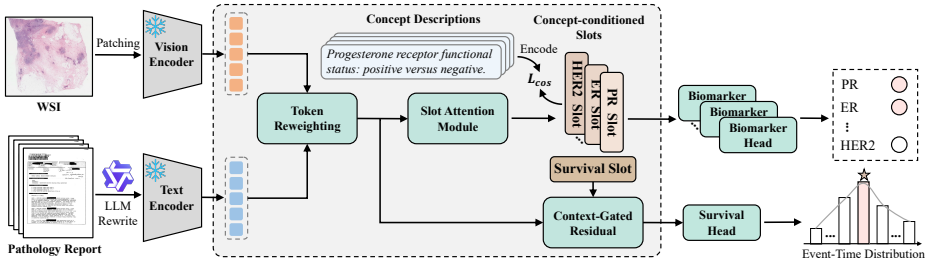

Fig. 1: Overview of the PathoSlot framework.

to its corresponding clinical prior, enhancing interpretability and cross-center generalization. Moreover, a context-gated residual treats the survival slot as a control variable to selectively inject global multimodal context into its updates under patient-level survival supervision. Evaluations on internal and external cohorts demonstrate that PathoSlot achieves superior performance over state-of-the-art methods in both biomarker and survival prediction.

2 Method

Figure 1 presents the PathoSlot framework for joint biomarker and survival prediction. PathoSlot reweights multimodal tokens and factorizes them into concept-conditioned, task-aligned slots that support hierarchical prediction, from local diagnostic biomarker status prediction to global prognostic risk estimation. We detail each component below.

Multimodal Token Construction and Reweighting. To leverage complementary clinical information from heterogeneous modalities, we construct multimodal tokens from paired WSIs and pathology reports using a multimodal slide-level foundation model (FM), TITAN [3]. Rather than relying on patch-level feature extraction followed by a separately designed slide-level aggregator, we extract modality-specific tokens directly from the pretrained slide-level FM before the final pooling stage, preserving fine-grained yet globally contextualized representations aligned with whole-slide clinical semantics. For the WSI modality, visual tokens $F_v \in \mathbb{R}^{N_v \times D}$ are extracted from the pretrained vision encoder. In parallel, pathology reports are rewritten into standardized and diagnostically focused textual descriptions using Qwen2.5-32B-Instruct [20] and are then encoded into textual tokens $F_t \in \mathbb{R}^{N_t \times D}$ by the TITAN text encoder. Prompt-based rewriting removes sensitive identifiers unrelated to pathological diagnosis and excludes outcome-revealing information to prevent label leakage, with all rewritten reports reviewed by three experienced pathologists with over ten years of clinical practice to ensure diagnostic fidelity. Visual and textual tokens are subsequently concatenated to form $F_{\text{raw}} \in \mathbb{R}^{(N_v+N_t)\times D}$, consistent with recent findings that simple concatenation serves as an effective multimodal fusion strategy [21].

To further distinguish informative tokens from non-discriminative visual or auxiliary textual ones without disrupting the representation space learned by the FM, we introduce a simple yet effective token reweighting mechanism. Specifically, a linear projection is applied to F_{raw} to produce token-wise importance scores, which are normalized across the token dimension with Softmax and used to rescale the original token features, yielding the reweighted representations $\hat{F}$:

$$\hat{F} = \text{Softmax}(\text{Linear}(F_{\text{raw}})) \odot F_{\text{raw}} \tag{1}$$

Concept-Conditioned Slot Factorization. Motivated by object-centric learning [10,14], which decomposes visual inputs into distinct latent entities via competitive slot interactions, we extend this paradigm to computational pathology for multimodal joint prediction. We treat each clinical prediction target as an entity inferred from shared multimodal evidence and softly factorize the reweighted token set into a compact set of slot embeddings. Here, factorization means supervised allocation of shared WSI and report evidence to endpoint-aligned slots rather than independent clinical factor discovery. All slots attend to the same token pool, enabling endpoint-specific weighting without isolated task paths. However, standard slot attention [10] lacks explicit endpoint semantics for slot identities, leading to ambiguous slots that may weaken interpretability and drift across tasks or centers toward different clinical evidence patterns. Such variability is undesirable when clinical reasoning requires consistent conceptual alignment across domains.

To stabilize slot semantics and enforce consistent slot-to-task correspondence, we introduce concept-conditioned slot initialization. For each task, we define a short concept description that precisely states the target semantics. For example, for PR status, "Progesterone receptor functional status: positive versus negative." We encode these descriptions using the TITAN text encoder to obtain concept embeddings $P \in \mathbb{R}^{M \times D}$, where M is the number of tasks, equal to the number of slots. We then initialize slots by injecting task-specific concept information into learnable slot initialization vectors, using a gated modulation that controls the injection strength per slot,

$$S^{(0)} = S_{\text{init}} + \sigma(G) \odot \text{LN}\big(\text{Linear}(P)\big), \tag{2}$$

where $S_{\text{init}} \in \mathbb{R}^{M \times D}$ denotes learnable slot initialization vectors, $\text{Linear}(\cdot)$ and LayerNorm $\text{LN}(\cdot)$ project the concept embeddings into the slot space, and $G \in \mathbb{R}^{M \times 1}$ is a learnable gating parameter. The sigmoid gate $\sigma(G) \in (0, 1)$ adaptively controls the strength of concept injection for each slot, providing a clinical prior while retaining the flexibility of data-driven slot learning. Unlike event-centric SlotSPE [24], which discovers sparse patient-specific prognostic events, PathoSlot factorizes evidence along predefined supervised clinical endpoints, using one slot per endpoint for stable readouts without post-hoc event-to-endpoint matching.

Given the reweighted multimodal tokens $\hat{F} = \{\hat{F}_n\}_{n=1}^{N_v+N_t}$ and slots $S = \{S_m\}_{m=1}^{M}$ with $\hat{F}_n \in \mathbb{R}^D$ and $S_m \in \mathbb{R}^D$, slot attention is applied for T iterations so that slots competitively explain the input tokens in a differentiated manner.

At iteration t, attention weights $A_{n,m}^{(t)}$ represent the soft assignment of token $\hat{F}_n$ to slot $S_m^{(t-1)}$ and are computed with a Softmax over the slot dimension,

$$A_{n,m}^{(t)} = \frac{\exp\left(\frac{1}{\sqrt{D}}\, q\left(S_m^{(t-1)}\right)^\top k\left(\hat{F}_n\right)\right)}{\sum_{j=1}^{M} \exp\left(\frac{1}{\sqrt{D}}\, q\left(S_j^{(t-1)}\right)^\top k\left(\hat{F}_n\right)\right)} \tag{3}$$

where $q(\cdot)$ and $k(\cdot)$ are linear projections into the query and key spaces. Slots are updated using an attention-weighted mean of values, followed by GRU-based refinement [2] to improve stability and reduce collapse. The update vector $U_m^{(t)}$ is computed as

$$U_m^{(t)} = \frac{\sum_{n=1}^{N_v+N_t} (A_{n,m}^{(t)} + \epsilon) \cdot v(\hat{F}_n)}{\sum_{n=1}^{N_v+N_t} (A_{n,m}^{(t)} + \epsilon)}, \tag{4}$$

where $v(\cdot)$ denotes the value projection and ϵ is a small constant for numerical stability. Slot refinement is then performed by

$$\tilde{S}_m^{(t)} = \mathrm{GRU}\left(U_m^{(t)}, S_m^{(t-1)}\right), \quad S_m^{(t)} = \tilde{S}_m^{(t)} + \mathrm{MLP}\left(\mathrm{LN}(\tilde{S}_m^{(t)})\right), \tag{5}$$

where MLP denotes a multilayer perceptron. After T iterations, $S_m^{(T)}$ serves as the representation of the corresponding task-aligned slot. To maintain endpoint-slot alignment during training, we add a cosine consistency loss $\mathcal{L}_{\cos}$ that aligns each slot with its concept embedding via cosine similarity.

Slot-Based Biomarker and Survival Prediction. Biomarker prediction is primarily driven by localized morphological patterns, whereas survival analysis requires modeling global tissue architecture and patient-level prognostic context. We therefore adopt a diagnostic-to-prognostic hierarchy rather than treating survival as a peer biomarker. For biomarker prediction, each concept-conditioned biomarker slot S_{bio} provides a focused representation of task-relevant local morphology and is directly connected to an MLP classifier for prediction.

For survival analysis, the survival slot S_{surv} serves not merely as a factorized representation, but as a control variable that regulates the selective incorporation of global multimodal context. To this end, we introduce a context-gated residual to aggregate C_{global} from $\hat{F}$ by global average pooling and adaptively inject it into the survival slot:

$$\hat{S}_{\text{surv}} = \mathrm{LN}\big(S_{\text{surv}} + \sigma\big(\mathrm{Linear}(S_{\text{surv}})\big) \odot C_{\text{global}}\big)\,. \tag{6}$$

The enriched representation $\hat{S}_{\text{surv}}$ is fed into the survival head to predict the risk score.

Training Objectives. PathoSlot is trained end-to-end in a multi-task setting for joint biomarker and survival prediction. For biomarker prediction, following [22], we use a masked binary cross-entropy loss $\mathcal{L}_{\text{bio}}$ that computes loss terms only for observed labels to accommodate incomplete annotations. For survival

modeling, continuous time is discretized into non-overlapping intervals and optimized with a discrete-time negative log-likelihood loss $\mathcal{L}_{\text{surv}}$. A cosine consistency regularizer $\mathcal{L}_{\cos}$ further enforces stable slot–task alignment during training. The overall objective is a weighted sum of these loss terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{bio}}\mathcal{L}_{\text{bio}} + \lambda_{\text{surv}}\mathcal{L}_{\text{surv}} + \lambda_{\cos}\mathcal{L}_{\cos}, \tag{7}$$

where λ_{bio} and λ_{surv} balance the biomarker and survival losses, respectively, while $\lambda_{\cos}$ modulates the consistency regularization without dominating the supervised task losses.

3 Experiments and Results

Datasets. Experiments were conducted on three multi-center cohorts: the public TCGA-BRCA dataset ($N = 1098$), a gastric cancer (GC) cohort from Zhongshan Hospital (GC-Internal, $N = 971$), and a GC cohort from Xiamen Hospital (GC-External, $N = 170$). GC-Internal and GC-External were collected under ethics approval Nos. B2021-785R and B2023-080R, respectively. All GC data were de-identified before analysis. We performed patient-level multi-biomarker prediction, targeting ER, PR, and HER2 status in TCGA-BRCA, as well as HER2 and MMR status in both GC cohorts. For prognostic modeling across all datasets, overall survival (OS) is used as the event of interest, following [5].

Implementation Details. All baselines were implemented with hyperparameters consistent with their original publications for fair comparison, with only task heads and losses adapted to the biomarker-survival endpoint setting. PathoSlot was trained for 20 epochs using the AdamW optimizer with a learning rate of 3×10^{-4}, weight decay of 1×10^{-5}, batch size 1, and a cosine learning rate scheduler. Slot factorization used $T = 3$ iterations. For TCGA-BRCA, $M = 4$ slots were used for ER, PR, HER2, and survival prediction, whereas for GC-Internal and GC-External, $M = 3$ slots were used for HER2, MMR, and survival. The loss weights λ_{bio}, λ_{surv}, and $\lambda_{\cos}$ were set to 1, 1, and 0.01, respectively. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU.

Internal Comparison. We conduct patient-level 5-fold cross-validation and report mean performance. For biomarker prediction, we use the area under the receiver operating characteristic curve (AUC) and balanced accuracy (bACC), while OS prediction is evaluated by the concordance index (C-Index). Table 1 compares PathoSlot with 13 state-of-the-art WSI-only and WSI-report methods on TCGA-BRCA and GC-Internal. For methods using patch-level features, representations are extracted with CONCH v1.5 [11], consistent with TITAN [3]. For patch-level foundation models, we follow the widely adopted evaluation protocol by integrating patch embeddings into ABMIL [8] and fine-tuning, while for slide-level foundation models we follow the original papers by fine-tuning pooled slide embeddings with task-specific heads. Overall, PathoSlot exhibits the strongest performance across both datasets. On TCGA-BRCA, PathoSlot achieved the best ER, PR, and HER2 results under both AUC and bACC and the highest survival C-index of 0.703, significantly improving over the strongest

Table 1: Internal validation results on TCGA-BRCA and GC-Internal.

Type	Method	TCGA-BRCA							GC-Internal				
		ER		PR		HER2		Surv	MMR		HER2		Surv
		AUC	bACC	AUC	bACC	AUC	bACC	C-Index	AUC	bACC	AUC	bACC	C-Index
WSI-only	DAMLN [22]	0.756	0.678	0.720	0.655	0.542	0.515	0.517	0.553	0.527	0.635	0.586	0.517
	Joint-MTL-CPath [4]	0.556	0.524	0.571	0.561	0.584	0.548	0.502	0.572	0.545	0.545	0.543	0.554
	HistMIMT [17]	0.903	0.841	0.824	0.758	0.712	0.674	0.613	0.612	0.587	0.640	0.594	0.588
	TDA-MIL [13]	0.880	0.824	0.807	0.740	0.662	0.633	0.577	0.575	0.569	0.712	0.635	0.597
	MAMMOTH [16]	0.895	0.844	0.829	0.773	0.719	0.675	0.609	0.560	0.570	0.676	0.630	0.623
	UNI2-h [1]	0.910	0.830	0.847	0.786	0.691	0.595	0.617	0.650	0.572	0.708	0.587	0.604
	Virchow2 [26]	0.875	0.798	0.808	0.743	0.634	0.569	0.610	0.589	0.554	0.716	0.576	0.592
WSI & Report	TITAN [3]	0.908	0.855	0.832	0.761	0.714	0.675	0.609	0.588	0.579	0.703	0.623	0.664
	PRISM [15]	0.869	0.810	0.812	0.782	0.691	0.667	0.609	0.558	0.543	0.656	0.595	0.605
	PathM3 [25]	0.892	0.813	0.808	0.743	0.659	0.616	0.599	0.635	0.607	0.710	0.604	0.582
	QPMIL-VL [5]	0.699	0.502	0.679	0.507	0.619	0.500	0.618	0.637	0.510	0.709	0.552	0.610
	MSCPT [6]	0.830	0.717	0.762	0.669	0.533	0.496	0.521	0.607	0.500	0.687	0.528	0.529
	MOC [23]	0.862	0.769	0.784	0.686	0.609	0.500	0.562	0.571	0.500	0.526	0.500	0.591
	PathoSlot (Ours)	**0.923**	**0.875**	**0.859**	**0.788**	**0.737**	**0.705**	**0.703**	**0.655**	**0.621**	**0.720**	**0.685**	**0.685**

Table 2: External validation results on GC-External.

Type	Method	GC-External				
		MMR		HER2		Surv
		AUC	bACC	AUC	bACC	C-Index
WSI-only	DAMLN [22]	0.549	0.519	0.586	0.548	0.508
	Joint-MTL-CPath [4]	0.582	0.571	0.529	0.524	0.587
	HistMIMT [17]	0.659	0.609	0.613	0.520	0.646
	TDA-MIL [13]	0.675	0.625	**0.687**	0.576	0.568
	MAMMOTH [16]	0.579	0.440	0.643	0.586	0.591
	UNI2-h [1]	0.611	0.539	0.627	0.503	0.613
	Virchow2 [26]	0.569	0.513	0.561	0.503	0.631
WSI & Report	TITAN [3]	0.650	0.614	0.640	0.570	0.661
	PRISM [15]	0.641	0.570	0.625	0.601	0.634
	PathM3 [25]	0.629	0.589	0.607	0.543	0.571
	QPMIL-VL [5]	0.608	0.503	0.639	0.541	0.608
	MSCPT [6]	0.614	0.523	0.671	0.537	0.531
	MOC [23]	0.537	0.500	0.584	0.500	0.563
	PathoSlot w/o CCond.	0.674	0.578	0.626	0.567	0.668
	PathoSlot	**0.687**	**0.626**	0.685	**0.624**	**0.673**

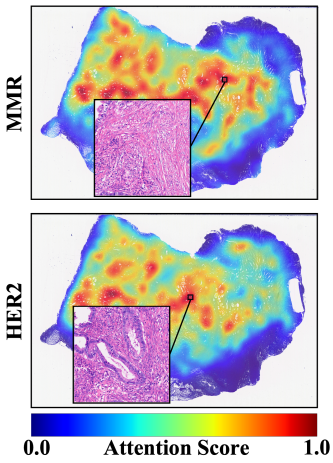

Fig. 2: Slot attention maps for MMR and HER2 on GC-External.

baselines in ER-AUC (+0.013, $p = 0.019$), HER2-bACC (+0.030, $p = 0.007$), and C-index (+0.085, $p < 0.001$). On GC-Internal, it again leads MMR and HER2 prediction and achieves the top survival C-index of 0.685, surpassing the second-best method at 0.664. Notably, PathoSlot consistently outperforms two representative WSI-only joint multi-class learning methods (Joint-MTL-CPath [4] and HistMIMT [17]), highlighting the benefit of explicitly incorporating multimodal information and structured endpoint-aligned slot factorization for joint biomarker and survival prediction.

External Comparison. As demonstrated in Table 2, we evaluate the models obtained from GC-Internal 5-fold validation on the independent GC-External cohort and report the averaged performance. PathoSlot preserves strong per-

Table 3: Ablation study on TCGA-BRCA for each component of PathoSlot. SA, CCond., CGR and TR denote Slot Attention, Concept Conditioning, Context-Gated Residual and Token Reweighting.

SA	CCond.	CGR	TR	AUC_{ER}	AUC_{PR}	AUC_{HER2}	C-Index
✗	✗	✗	✗	0.870	0.794	0.672	0.638
✓	✗	✗	✗	0.905	0.811	0.708	0.641
✓	✓	✗	✗	0.911	0.825	0.706	0.648
✓	✓	✓	✗	0.915	0.841	0.722	0.685
✓	✓	✓	✓	**0.923**	**0.859**	**0.737**	**0.703**

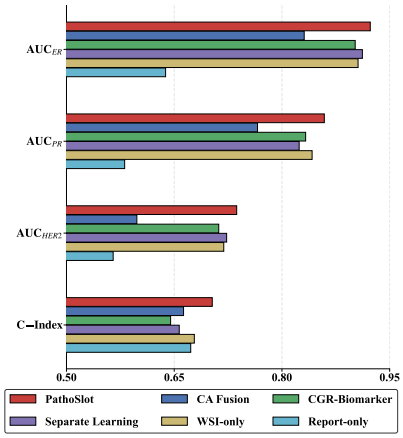

Fig. 3: Ablation study on TCGA-BRCA under different settings of PathoSlot.

formance under this cross-center setting, achieving the best survival prediction on GC-External and maintaining competitive biomarker prediction, indicating robust generalization beyond the training domain. In contrast, removing concept conditioning (w/o CCond., Table 2) consistently degrades external performance, particularly for biomarker prediction, underscoring its importance for cross-center robustness. Anchoring each slot to a predefined clinical concept preserves stable slot semantics and task alignment, facilitating reliable generalization. Fig. 2 further supports cross-center generalization with interpretable attention evidence, showing that the MMR slot emphasizes lymphocyte-infiltrated stroma whereas the HER2 slot attends to tumor epithelium, consistent with the expected tissue contexts for each biomarker.

Ablation Study. We perform an ablation study on TCGA-BRCA to assess the effect of slot attention (SA), concept conditioning (CCond.), context-gated residual (CGR), and token reweighting (TR). As shown in Table 3, progressively incorporating the proposed modules improves overall performance, with the full model achieving the best results across all biomarker AUCs and C-Index. Fig. 3 further compares PathoSlot with alternative settings. Replacing token reweighting with cross-attention (CA) fusion degrades performance, indicating that token reweighting emphasizes informative tokens without perturbing the embedding space. Notably, joint learning consistently surpasses separate learning, indicating that cross-task supervision improves inductive transfer and regularization through coordinated optimization of task-specific slots. Introducing CGR to biomarker slots reduces accuracy, implying that biomarker prediction benefits from localized representations, whereas survival prediction requires global context. Finally, single-modality variants perform worse than the multimodal model, confirming the value of integrating complementary modalities.

4 Conclusion

In this work, we propose PathoSlot, a concept-conditioned slot factorization framework for joint biomarker prediction and survival modeling from multimodal pathology data. PathoSlot factorizes shared WSI-report evidence into task-aligned slots, enabling localized biomarker prediction and global survival estimation within a unified framework. Concept conditioning stabilizes slot semantics and enhances interpretability and robustness across centers. Evaluations on internal and external cohorts demonstrate that PathoSlot consistently surpasses state-of-the-art methods on both biomarker and survival prediction tasks.

Acknowledgments. This work was sponsored by Shanghai Pujiang Program under Grant 25PJA074.

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

References

1. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: A multimodal whole-slide foundation model for pathology. Nature medicine pp. 1–13 (2025)
4. El Nahhas, O.S., Wölflein, G., Ligero, M., Lenz, T., van Treeck, M., Khader, F., Truhn, D., Kather, J.N.: Joint multi-task learning improves weakly-supervised biomarker prediction in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 254–262. Springer (2024)
5. Gou, J., Ji, L., Liu, P., Ye, M.: Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3158–3166 (2025)
6. Han, M., Qu, L., Yang, D., Zhang, X., Wang, X., Zhang, L.: Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. IEEE Transactions on Medical Imaging (2025)
7. Hou, J., Zhang, R., Xie, Y., Li, C., Qin, W.: Multimodal deep learning for cancer prognosis prediction with clinical information prompts integration. npj Digital Medicine (2025)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

9. Jonnagaddala, J., Shulajkovska, M., Gradišek, A., Jue, T.R., Zhou, Q., Guo, Y., Chayeb, J.M.E., Li, R., Lipkova, J., Kather, J.N., et al.: Multimodal analysis of whole slide images in colorectal cancer. npj Digital Medicine **8**(1), 719 (2025)
10. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in neural information processing systems **33**, 11525–11538 (2020)
11. 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)
12. McShane, L.M., Altman, D.G., Sauerbrei, W., Taube, S.E., Gion, M., Clark, G.M.: Reporting recommendations for tumor marker prognostic studies (remark). Breast cancer research and treatment **100**(2), 229–235 (2006)
13. Reisenbüchler, D., Deng, R., Matek, C., Feuerhake, F., Merhof, D.: Top-down attention-based multiple instance learning for whole slide image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 651–660. Springer (2025)
14. Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.J., He, T., Zhang, Z., Schölkopf, B., Brox, T., et al.: Bridging the gap to real-world object-centric learning. arXiv preprint arXiv:2209.14860 (2022)
15. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., et al.: Prism: A multimodal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024)
16. Shao, D., Runevic, J., Chen, R.J., Williamson, D.F., Kim, A., Song, A.H., Mahmood, F.: Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=S5Io33pc78
17. Shao, W., Shi, H., Liu, J., Zuo, Y., Sun, L., Xia, T., Chen, W., Wan, P., Sheng, J., Zhu, Q., et al.: Multi-instance multi-task learning for joint clinical outcome and genomic profile predictions from the histopathological images. IEEE transactions on medical imaging **43**(6), 2266–2278 (2024)
18. Skrede, O.J., De Raedt, S., Kleppe, A., Hveem, T.S., Liestøl, K., Maddison, J., Askautrud, H.A., Pradhan, M., Nesheim, J.A., Albregtsen, F., et al.: Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet **395**(10221), 350–360 (2020)
19. Subik, K., Lee, J.F., Baxter, L., Strzepek, T., Costello, D., Crowley, P., Xing, L., Hung, M.C., Bonfiglio, T., Hicks, D.G., et al.: The expression patterns of er, pr, her2, ck5/6, egfr, ki-67 and ar by immunohistochemical analysis in breast cancer cell lines. Breast cancer: basic and clinical research **4**, 117822341000400004 (2010)
20. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
21. Tripathi, A., Waqas, A., Schabath, M.B., Yilmaz, Y., Rasool, G.: Honeybee: enabling scalable multimodal ai in oncology through foundation model-driven embeddings. npj Digital Medicine **8**(1), 622 (2025)
22. Wang, M., Wang, T., Cong, F., Lu, C., Xu, H.: Double-tier attention based multi-label learning network for predicting biomarkers from whole slide images of breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 91–101. Springer (2024)
23. Xiang, T., Li, Y., Zhang, Q., Li, X.: Moc: Meta-optimized classifier for few-shot whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 423–432. Springer (2025)

24. Zhang, Y., Nanbo, L., Yang, C., Schmidhuber, J., Gao, X.: Structural prognostic event modeling for multimodal cancer survival analysis. arXiv preprint arXiv:2512.01116 (2025)
25. Zhou, Q., Zhong, W., Guo, Y., Xiao, M., Ma, H., Huang, J.: Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 373–383. Springer (2024)
26. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)