H^2-Surv: Hierarchical Hyperbolic Multimodal Learning for Survival Prediction

Supplementary Material

1. Rationale

1.1. Ablation experiments on hyperparameters.

In this section, we conducted ablation experiments on hyperparameters λ, β, focusing on the effects of the regularization loss $\mathcal{L}_{\text{Hd}}$, $\mathcal{L}_{\text{Hyent}}$ and $\mathcal{L}_{\text{ordinal}}$. We used the concordance index (C-index) as the evaluation metric, and the specific results across five TCGA datasets (BRCA, BLCA, UCEC, LUAD, GBMLGG) are summarized in Table 1.

As shown in the table, we observe that small values of λ (e.g., 0.001) result in under-regularization and suboptimal performance across all datasets. Increasing λ gradually improves performance, and the best overall results are achieved when $\lambda = 0.01$, $\beta = 0.1$. Under this setting, our model obtains the highest C-index across all datasets: 0.763 (BRCA), 0.701 (BLCA), 0.760 (UCEC), 0.700 (LUAD), and 0.859 (GBMLGG). This configuration balances the contribution of survival ranking constraints and censoring information, enhancing the model's ability to learn robust representations.

We also observe performance degradation when λ becomes too large (e.g., $\lambda = 0.5$), likely due to over-penalization from regularization terms that dominate the optimization. Therefore, we selected $\lambda = 0.01$, $\beta = 0.1$ as the optimal hyperparameter combination for all subsequent experiments.

Table 1. Ablation study of hyperparameters λ, β on C-index across datasets.

λ	β	BRCA	BLCA	UCEC	LUAD	GBM
0.001	0.01	0.640	0.618	0.568	0.598	0.609
0.005	0.05	0.669	0.670	0.669	0.626	0.741
0.01	**0.1**	0.763	0.701	0.760	0.700	0.859
0.05	0.1	0.749	0.699	0.701	0.691	0.824
0.1	0.5	0.662	0.690	0.656	0.624	0.690
0.5	0.8	0.649	0.611	0.599	0.601	0.630

1.2. T-test Analysis.

The T-test is a statistical analysis method used to determine if there is a significant difference between the means of two groups. In the T-test analysis, the t-value quantifies the size of the difference relative to the variability within two groups, while the p-value is the probability that the results from our results occurred by chance. As the results shown in Fig. 1, our framework can distinguish between different groups with good p-values and t-values, especially on BLCA, and UCEC datasets.

Besides, we report the p-values of pairwise log-rank tests across five TCGA datasets (BLCA, BRCA, GBMLGG, LUAD, and UCEC) to evaluate the statistical significance of survival curve separability achieved by different methods. As shown in Table 2, our proposed method consistently yields significantly lower p-values compared to existing baselines, indicating stronger survival stratification capability. Particularly, our method achieves the lowest p-value on BLCA ($3.2e^{-15}$), GBMLGG ($1.4e^{-43}$), and LUAD ($8.2e^{-22}$), and highly competitive results on the remaining datasets. These results confirm the robustness and effectiveness of our model in capturing prognostic signals across diverse cancer types.

Table 2. P-values of pairwise significance tests (log-rank test) across five TCGA datasets. Lower is better.

Methods	**BLCA**	**BRCA**	**GBMLGG**	**LUAD**	**UCEC**
SNN	$1.7e^{-6}$	$1.5e^{-2}$	$1.1e^{-27}$	$1.1e^{-3}$	$9.8e^{-4}$
TransMIL	$5.6e^{-2}$	$2.4e^{-3}$	$2.5e^{-25}$	$3.8e^{-2}$	$9.8e^{-2}$
MCAT	$7.8e^{-6}$	$7.5e^{-3}$	$1.6e^{-19}$	$6.9e^{-5}$	$4.8e^{-3}$
SurvPath	$8.2e^{-6}$	$1.4e^{-3}$	$1.0e^{-29}$	$2.2e^{-4}$	$1.4e^{-5}$
MOTCat	$2.9e^{-7}$	$4.9e^{-4}$	$3.4e^{-30}$	$1.1e^{-5}$	$3.8e^{-7}$
CMTA	$2.0e^{-8}$	$3.1e^{-3}$	$1.8e^{-33}$	$9.6e^{-7}$	$1.7e^{-3}$
CGM	$5.7e^{-11}$	$2.2e^{-7}$	$9.8e^{-32}$	$1.1e^{-4}$	$3.6e^{-4}$
Ours	$\mathbf{3.2e^{-15}}$	$\mathbf{1.5e^{-5}}$	$\mathbf{1.4e^{-43}}$	$8.2e^{-22}$	$\mathbf{3.9e^{-10}}$

1.3. Efficiency comparison

Table 3 presents a comparison of our method with several existing multimodal survival prediction models in terms of GPU memory consumption (GB), training time, and inference time (in minutes). While models such as SurvPath and MCAT demonstrate low computational overhead, they sacrifice performance and lack fine-grained multimodal integration. In contrast to existing approaches, our method adopts a hyperbolic fusion strategy that enables more expressive modeling of multimodal interactions, particularly under hierarchical constraints. While this leads to moderately higher training resource requirements (GPU memory: 12.98 GB), the inference time remains comparable to MOTCat (10.34 vs. 38.35 min). Importantly, our method achieves significantly better predictive performance (see Section 4 of the original paper), demonstrating that the added training complexity translates into meaningful survival prediction gains—a favorable trade-off in clinical survival modeling scenarios where inference efficiency and accuracy are both critical.

Table 3. Efficiency comparison with baselines in terms of GPU memory (GB), training time, and test time (minutes).

Methods	GPU memory ↓	Training time ↓	Test time ↓
MCAT [2]	4.06	13.78	9.69
MOTCat [4]	3.09	39.28	38.35
CMTA [7]	19.70	33.17	28.14
PIBD [6]	23.06	59.71	23.49
Ours(ES)	12.56	34.22	11.85
Ours(HS)	12.98	36.36	10.34

1.4. Robustness to Missing Modalities.

Previous multimodal survival prediction methods typically assume that all modalities are available during both training and inference. However, in real-world clinical settings, multi-omics data are often incomplete due to technical limitations or sample quality issues. To assess robustness under missing-modality conditions, we simulate two levels of modality absence—60% and 100%—by randomly masking either the WSI or genomic branch during inference. As shown in Table 4, our model consistently achieves the highest C-index values across all datasets and missing levels. Under complete modality absence (100%), our approach outperforms DisPro (CVPR25) [5] by up to 0.047 on CPTAC-LUAD. Even under partial missing conditions (60%), H^2-Surv maintains clear advantages over all baselines, demonstrating strong resilience to modality dropout. These improvements stem from the pathway-level bridge and the TOCL module, which jointly enhance cross-modal alignment and temporal survival ranking.

Table 4. C-index performance under missing-modality conditions on three cancer cohorts. ✗ is the missing modality, and ✓ means the modality is available.

Missing Rate	WSI	Genomics	Model	TCGA-LUAD	CPTAC-LUAD
100%	✗	✓	SNNTrans	0.638 ± 0.022	0.602 ± 0.018
			DisPro (CVPR25)	0.646 ± 0.038	0.634 ± 0.012
			Ours	0.676 ± 0.050	0.681 ± 0.031
100%	✓	✗	DFTD-MIL	0.647 ± 0.051	0.626 ± 0.019
			DisPro (CVPR25)	0.639 ± 0.061	0.631 ± 0.014
			Ours	0.681 ± 0.057	0.690 ± 0.017
60%	✗	✓	DisPro (CVPR25)	0.673 ± 0.013	0.664 ± 0.039
			Ours	0.684 ± 0.028	0.686 ± 0.010
60%	✓	✗	DisPro (CVPR25)	0.668 ± 0.044	0.682 ± 0.028
			Ours	0.693 ± 0.036	0.693 ± 0.024

1.5. Sensitivity to Curvature Hyperparameter

The curvature c is learned via a *softplus* transformation and optimised using the same learning rate as the rest of the network. This enables adaptive fitting to the data manifold without manual tuning. To prevent geometric distortion from extreme values, we apply optional bounds (e.g., min = $1e^{-4}$, max = 5.0), which are rarely triggered in practice as c consistently converges to moderate values. We also evaluated the **sensitivity to curvature initialisation**. As shown in the table below, the model remains robust across a wide range of initial $\hat{c}$ values, consistently converging to similar values of c and achieving stable validation performance. This confirms that our adaptive strategy **provides reliable geometry without the need for extensive tuning**.

Table 5. Effect of different initializations of $\hat{c}_0$ on the final learned curvature c and C-index on BRCA.

Initialization $\hat{c}_0$	softplus($\hat{c}_0$)	Final Learned c	C-index on BRCA
0.5	0.97	1.92	0.760
1.0	1.31	2.02	0.763
2.0	2.13	1.95	0.759
5.0	5.01	2.08	0.761

1.6. Hyperbolic Embedding Visualization

To better understand the hyperbolic constraints and the resulted embeddings, we employ HoroPCA [1] to visualize the learned hyperbolic embeddings. Using the LUAD survival prediction task as an example, Fig. **??** displays features across modalities and hierarchical levels. The visualization reveals a clear hierarchical organization: patient-level embeddings are concentrated near the origin, WSI- and pathway-level embeddings lie at intermediate radii, and gene- and patch-level embeddings are positioned closer to the boundary of the Poincaré ball. Moreover, distinct decision boundaries between risk groups indicate that our method effectively aligns and distributes multimodal hierarchical features in hyperbolic space. Such structured representations provide a more faithful organization of cross-modal information and contribute to improved performance on multimodal survival prediction.

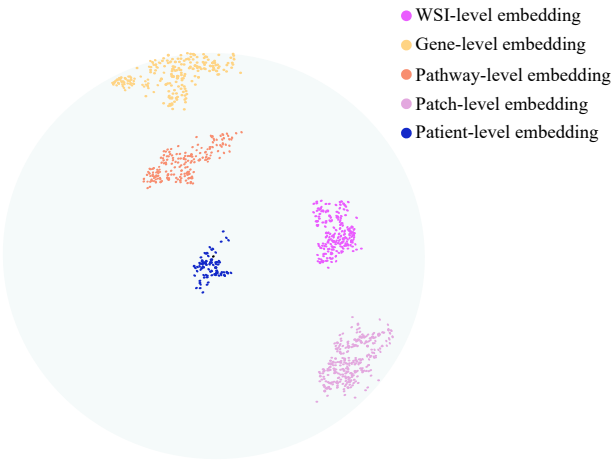

Figure 1. Distribution of key features in hyperbolic space.

1.7. Qualitative Analysis of Model Interpretability

To qualitatively demonstrate the interpretability and feature-level complementarity achieved through the fusion

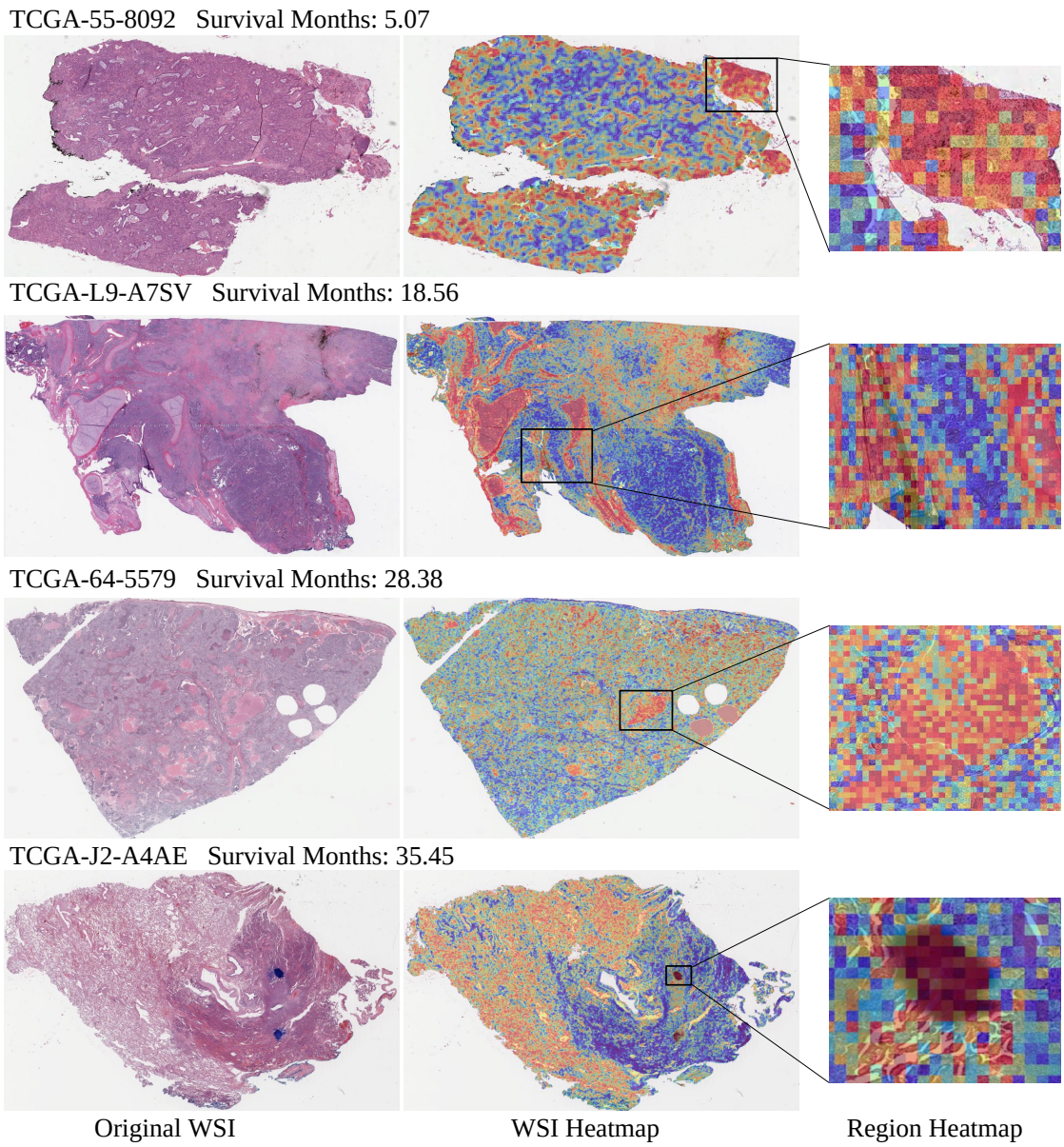

Figure 2. Pathological gene attention heatmap display on TCGA-LUAD dataset.

of histopathology and genomic modalities, we conducted visual and attribution-based analyses to examine modality-specific contributions to survival prediction. Fig. 2 illustrates the interpretability advantages of our H^2-Surv. For example, in LUAD cohorts, attention-based heatmaps highlight discriminative histopathological regions in high-risk and low-risk patients. Representative patches are extracted to illustrate morphological characteristics associated with risk. In parallel, the Integrated Gradients (IG) method [3] is applied to identify key genes with high attribution values, providing insight into molecular contributions to survival. All morphological and gene-level findings are reviewed and confirmed by expert pathologists, validating the model's ability to identify clinically meaningful biomarkers across modalities. For example, in the LUAD cohort, case TCGA-L9-A7SV shows dense cellular clustering and irregular gland formation in the heatmap, with top-ranked genes such as HS3ST6 and FAM135B exhibiting strong positive contributions, leading to a high-risk prediction consistent with its poor survival outcome.

References

[1] Ines Chami, Albert Gu, Dung P Nguyen, and Christopher Ré. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In *Proceedings of the 38th International*

Conference on Machine Learning (ICML), pages 1419–1429. PMLR, 2021. 2

[2] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE Int. Conf. Comput. Vis.*, pages 4015–4025, 2021. 2

[3] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[4] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE Int. Conf. Comput. Vis.*, pages 21241–21251, 2023. 2

[5] Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, and Hao Chen. Distilled prompt learning for incomplete multimodal survival prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5102–5111, 2025. 2

[6] Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646*, 2024. 2

[7] Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE Int. Conf. Comput. Vis.*, pages 21485–21494, 2023. 2