OTSurv: A Novel Multiple Instance Learning Framework for Survival Prediction with Heterogeneity-aware Optimal Transport

Qin Ren[1] Yifan Wang[1] Ruogu Fang[3] Haibin Ling[1] Chenyu You[1,2]

[1]Department of Computer Science, Stony Brook University
[2] Department of Applied Mathematics & Statistics, Stony Brook University
[3] Department of Biomedical Engineering, University of Florida
{qinren,cyou}@cs.stonybrook.edu

Abstract. Survival prediction using whole slide images (WSIs) can be formulated as a multiple instance learning (MIL) problem. However, existing MIL methods often fail to explicitly capture pathological heterogeneity within WSIs, both globally – through long-tailed morphological distributions, and locally – through tile-level prediction uncertainty. Optimal transport (OT) provides a principled way of modeling such heterogeneity by incorporating marginal distribution constraints. Building on this insight, we propose **OTSurv**, a novel MIL framework from an optimal transport perspective. Specifically, OTSurv formulates survival predictions as a heterogeneity-aware OT problem with two constraints: (1) *global long-tail constraint* that models prior morphological distributions to avert both mode collapse and excessive uniformity by regulating transport mass allocation, and (2) *local uncertainty-aware constraint* that prioritizes high-confidence patches while suppressing noise by progressively raising the total transport mass. We then recast the initial OT problem, augmented by these constraints, into an unbalanced OT formulation that can be solved with an efficient, hardware-friendly matrix scaling algorithm. Empirically, OTSurv sets new state-of-the-art results across six popular benchmarks, achieving an absolute 3.6% improvement in average C-index. In addition, OTSurv achieves statistical significance in log-rank tests and offers high interpretability, making it a powerful tool for survival prediction in digital pathology. Our codes are available at https://github.com/Y-Research-SBU/OTSurv.

Keywords: Survival Prediction · Multiple Instance Learning · Optimal Transport.

1 Introduction

Survival prediction, which estimates patient-specific time-to-event probabilities (*i.e.*, overall survival), is a critical oncology task essential for optimizing therapeutic strategies [26]. In clinical practice, it mainly relies on pathologists' meticulous analysis of tissue slides. However, variations in tissue composition and

structure across different regions pose significant challenges in identifying prognostic patterns [2,6,20,35,34,19,10]. In particular, pathologists need to detect and interpret small or ambiguous regions that are crucial for survival outcomes. Capturing this heterogeneity is inherently challenging, and even slight missteps can lead to incomplete prognostic assessments that ultimately compromise patient survival.

In digital pathology, neural networks have shown great promise in WSI-based survival prediction, but their ultra-high resolution (*e.g.*, $10^5 \times 10^5$ pixels) incurs high computational and annotation costs, rendering fully supervised learning impractical [32]. Thus, weakly supervised methods, particularly Multiple Instance Learning (MIL), have become the *de facto* solution, which treats each WSI as a bag of instances with only a bag-level survival label [31,38,14,23]. Analogous to how pathologists analyze tissue heterogeneity, MIL methods are required to effectively capture this variability to ensure that crucial survival-related information is not lost. Specifically, survival-related heterogeneity manifests at two scales. At *global* level, it involves capturing sparse patterns within the long-tailed distribution of patch types [7]. At *local* level, it focuses on eliminating the predictive uncertainty of each patch [21]. However, this dual-scale heterogeneity remains under-explored in current MIL approach: some completely ignore heterogeneity (*e.g.*, Mean/Max Pooling), some focus only on single level (*global* [25,16] or *local* [27,9,14,23,31]), while others attempt to tackle both but without providing explicit guidance [33]. Additionally, computational limits force MIL models to sample tiles randomly, risking the omission of critical prognostic regions and hindering heterogeneity modeling.

In this work, we aim to address the following question: *how can MIL-based models explicitly account for both global and local pathological heterogeneity?* We note that this heterogeneity in MIL essentially manifests as feature distribution variability, with global and local heterogeneity reflected in prototype-level and instance-level feature distributions, respectively. Optimal Transport (OT) [29,12], as a mathematically interpretable framework that optimally aligns two distributions while minimizing transport cost, naturally fits this setting. By incorporating two marginal distribution constraints, OT provides a principled mechanism to explicitly model dual-scale heterogeneity.

Building on this insight, we propose **OTSurv**, an **OT**-based MIL framework for **surv**ival prediction from an optimal transport perspective. OTSurv models survival prediction as an OT problem by aggregating a variable-sized set of instance features and mapping them onto a fixed-size set of trainable survival tokens, which is used for reference. This process explicitly accounts for both global and local heterogeneity through two tailored marginal distribution constraints. For *global heterogeneity* (long-tailed distribution), we introduce **G**lobal Long-tail **C**onstraint ($\mathbf{C_G}$), which models prior morphological distributions to prevent mode collapse and excessive uniformity by regularizing transport mass allocation. For *local heterogeneity* (difficult-to-predict patches), we propose **L**ocal Uncertainty-aware **C**onstraint ($\mathbf{C_L}$), which prioritizes high-confidence patches to suppress noise by progressively raising total transport mass. We then recast the

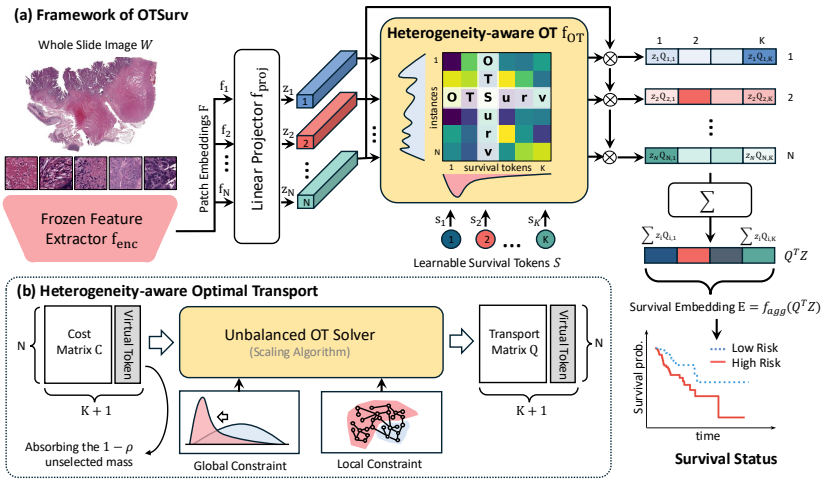

Fig. 1: **Overview of the OTSurv framework.** (a) Framework of OTSurv: A WSI W is processed into patch features Z, aligned with survival tokens S via heterogeneity-aware OT, and aggregated into a survival embedding E. (b) heterogeneity-aware Optimal Transport: By extending the cost matrix with an additional column of zeros (*i.e.*, virtual token), the unbalanced OT solver can be used for solving heterogeneity-aware OT with global and local constraints.

augmented OT problem into an unbalanced OT formulation that can be solved with an efficient, hardware-friendly matrix scaling algorithm, by incorporating a virtual survival token to absorb unselected mass.

Our key contributions are: (i) We propose **OTSurv**, a novel OT-based MIL framework for WSI survival prediction. (ii) We design two tailored marginal constraints in OT to model global and local heterogeneity in WSIs. (iii) Experiments on 6 popular benchmarks show that OTSurv outperforms previous SOTA methods in C-index, achieves statistical significance in log-rank tests, and provides strong interpretability.

2 Method

2.1 Overview

As illustrated in Fig. 1, OTSurv reframes WSI-based survival prediction as a heterogeneity-aware optimal transport problem. It is composed of four distinct, yet interlocking, modules: (i) *WSI Decomposition*: A large-scale WSI W is partitioned into N non-overlapping patches $\{x_i\}_{i=1}^N$, denoted as $x = \{x_i\}_{i=1}^N \in \mathbb{R}^{N\times c\times h\times w}$, where c is the number of color channels, and $h \times w$ is the patch resolution; (ii) *Feature Embedding*: A frozen encoder f_{enc} extracts patch-level features $F = f_{\text{enc}}(x) \in \mathbb{R}^{N\times D}$. These are then projected into a lower-dimensional latent space via a learnable linear projection, yielding instance embeddings $Z = f_{\text{proj}}(F) \in \mathbb{R}^{N\times d}$; (iii) *Feature Aggregation*: At the **core** of our approach,

the heterogeneity-aware OT module computes an optimal transport plan $Q = f_{\text{OT}}(C) \in \mathbb{R}^{N\times K}$, where the cost matrix $C \in \mathbb{R}^{N\times K}$ is computed using normalized Euclidean distance between $Z \in \mathbb{R}^{N\times d}$ and learnable survival tokens $S \in \mathbb{R}^{K\times d}$. The aggregated slide-level embedding is then obtained as $E = f_{\text{agg}}(Q^\top Z) \in \mathbb{R}^d$, where $Q^\top Z \in \mathbb{R}^{K\times d}$ and $f_{\text{agg}}(\cdot)$ is linear layer; and (iv) *Risk Prediction*: finally, the aggregated embedding E is fed into a linear predictor f_{pred} to produce the final risk score $r = f_{\text{pred}}(E) \in \mathbb{R}$. The entire model is optimized by minimizing a Cox proportional hazards loss [5].

2.2 Heterogeneity-aware OT Problem

In this section, we first overview OT theory, and then delve into an in-depth discussion of our proposed local and global OT marginal distribution constraints, which are pivotal to our framework.

General OT Formulation. To formulate heterogeneity-aware OT, we start with the OT problem, which aims to transport one distribution to another with minimal cost. Given a source distribution $\mu \in \mathbb{R}^{N\times 1}$, a target distribution $\nu \in \mathbb{R}^{K\times 1}$, and a cost matrix $C \in \mathbb{R}^{N\times K}$, our objective is to determine a transport matrix $Q \in \mathbb{R}^{N\times K}$ that minimizes:

$$\min_{Q\in\mathbb{R}_+^{N\times K}} \langle Q, C\rangle_F + F_1\left(Q\mathbf{1}_K,\, \mu\right) + F_2\left(Q^\top \mathbf{1}_N,\, \nu\right) \tag{1}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product, and Q_{ij} represents mass transported from the μ_i to ν_j. F_1 and F_2 enforce constraints on the marginal distributions of Q. When these constraints are specified as equalities (*i.e.*, $Q\mathbf{1}_K = \mu$, $Q^\top \mathbf{1}_N = \nu$), the formulation reduces to Kantorovich's classical OT problem [11]. Alternatively, when F_1 and F_2 impose inequality constraints (*e.g.*, using the KL divergence), the problem turns into unbalanced OT [15].

Global Long-tail Constraint ($\mathbf{C_G}$). WSIs exhibit a long-tailed tissue morphology, where dominant patterns prevail while rare, prognostically critical features are scarce. Without any constraint, the transport mass may collapse onto a single survival token, whereas enforcing a strict equality constraint forces a uniform mass distribution across tokens – neither case captures the true long-tailed nature of the data. To address this, we impose a KL divergence constraint on the survival token mass $Q^\top \mathbf{1}_N \in \mathbb{R}_+^{K\times 1}$ to match a desired long-tailed prior. This global constraint ($\mathbf{C_G}$) preserves tissue diversity by ensuring that every infrequent, yet critical, patterns are adequately represented for accurate survival prediction. Specifically, we apply a KL divergence penalty on $F_2(\cdot)$ while temporarily assuming a uniform distribution for $F_1(\cdot)$, where $Q\mathbf{1}_K \in \mathbb{R}_+^{N\times 1}$. This leads to the following semi-relaxed OT formulation:

$$\begin{aligned} \min_{Q\in\Pi} \quad & \langle Q, C\rangle_F + \lambda\, \mathrm{KL}\left(Q^\top \mathbf{1}_N \,\|\, \tfrac{1}{K}\mathbf{1}_K\right) \\ \text{subject to} \quad & \Pi = \left\{ Q \in \mathbb{R}_+^{N\times K} \;\middle|\; Q\mathbf{1}_K = \tfrac{1}{N}\mathbf{1}_N \right\} \end{aligned} \tag{2}$$

where λ is a weighting factor controlling the KL divergence regularization.

Local Uncertainty-aware Constraint ($\mathbf{C_L}$). In Eq. (2), the $F_1(\cdot)$ constraint $Q\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N$ treats all instances equally, which may lead to noisy alignments due to poor initial representations. Inspired by curriculum learning [30], a more effective approach is to start with easy samples and gradually incorporate harder ones. Instead of manually selecting confident samples via hard thresholds in the cost matrix, we reformulate the selection as a total mass constraint in Eq. (2), eliminating the need for sensitive hyperparameter tuning. This results in the following formulation:

$$\begin{aligned} \min_{Q \in \Pi} \quad & \langle Q, C \rangle_F + \lambda \, \mathrm{KL}\left(Q^\top \mathbf{1}_N \,\|\, \tfrac{\rho}{K}\mathbf{1}_K\right) \\ \text{subject to} \quad & \Pi = \left\{ Q \in \mathbb{R}_+^{N \times K} \;\middle|\; Q\mathbf{1}_K \le \tfrac{1}{N}\mathbf{1}_N,\ \mathbf{1}_N^\top Q \mathbf{1}_K = \rho \right\} \end{aligned} \tag{3}$$

where $\rho \in (0, 1]$ is the fraction of selected mass, which gradually increases during training. The resulting $Q\mathbf{1}_K$ represents the transport weights for each instance. We employ a sigmoid ramp-up function, a common technique in semi-supervised learning [22,28], to progressively increase ρ during training:

$$\rho = \rho_0 + (1 - \rho_0) \cdot e^{-5(1 - t/(T \cdot I))^2}, \tag{4}$$

where t is the current iteration, T is the ramp-up epochs and I is the number of iterations per epoch.

Heterogeneity-aware OT. We denote this OT formulation as heterogeneity-aware OT, as it incorporates dual marginal distribution constraints. By iteratively solving Eq. (3) with a sigmoid ramp-up of ρ during training, this approach seamlessly integrates instance selection, weighting, and aggregation within a unified OT framework.

2.3 Heterogeneity-aware OT Solver

Existing scaling algorithms [13] are designed for unbalanced OT, whereas our heterogeneity-aware OT, with its two tailored marginal constraints, differs from standard unbalanced OT. Despite this, by incorporating a virtual point, we can recast our formulation into an unbalanced OT problem that these efficient algorithms can solve.

As shown in Fig. 1(b), we introduce a *virtual survival token* to absorb the $1-\rho$ unselected mass, ensuring compatibility with the unbalanced OT framework. Specifically, we augment the cost matrix C with an additional column of zeros to form $\hat{C} \in \mathbb{R}_+^{N \times (K+1)}$. Consequently, Eq. (3) can be reformulated as follows:

$$\begin{aligned} \min_{\hat{Q} \in \Phi} \quad & \langle \hat{Q}, \hat{C} \rangle_F + \hat{\mathrm{KL}}\left(\hat{Q}^\top \mathbf{1}_N, \beta, \hat{\lambda}\right) \\ \text{subject to} \quad & \Phi = \left\{ \hat{Q} \in \mathbb{R}_+^{N \times (K+1)} \mid \hat{Q}\mathbf{1}_{K+1} = \frac{1}{N}\mathbf{1}_N \right\} \end{aligned} \tag{5}$$

Algorithm 1: Scaling Algorithm for heterogeneity-aware OT

Input: Cost matrix C, ϵ, λ, ρ, N, K, a large value ι
$C \leftarrow [C, \mathbf{0}_N]$, $\lambda \leftarrow [\lambda, ..., \lambda, \iota]^\top$, $\beta \leftarrow [\frac{\rho}{K}\mathbf{1}_K^\top, 1-\rho]^\top$
$\alpha \leftarrow \frac{1}{N}\mathbf{1}_N$, $b \leftarrow 1_{K+1}$, $M \leftarrow \exp(-C/\epsilon)$, $f \leftarrow \lambda \oslash (\lambda + \epsilon)$
while b *not converge* **do**
  $a \leftarrow \alpha \oslash (Mb)$; $b \leftarrow (\beta \oslash (M^\top a))^{\circ f}$
end
$Q \leftarrow \mathrm{diag}(a) M \mathrm{diag}(b)$
return $Q[:, :K]$

where

$$\hat{C} = [C, \mathbf{0}_N], \quad \beta = \begin{bmatrix} \frac{\rho}{K}\mathbf{1}_K \\ 1-\rho \end{bmatrix}, \quad \hat{\lambda} = \begin{bmatrix} \lambda \mathbf{1}_K \\ +\infty \end{bmatrix},$$
$$\hat{\mathrm{KL}}\left(\hat{Q}^\top \mathbf{1}_N, \beta, \hat{\lambda}\right) = \sum_{i=1}^{K+1} \hat{\lambda}_i \left[\hat{Q}^\top \mathbf{1}_N\right]_i \log \frac{[\hat{Q}^\top \mathbf{1}_N]_i}{\beta_i}. \tag{6}$$

Here, the weighted KL divergence enforces that the virtual survival token absorbs exactly $1-\rho$ of the unselected mass. [37] theoretically shows that the optimal transport plan $Q^\star$ from Eq. (3) corresponds to the first K columns of the optimal transport plan $\hat{Q}^\star$ from Eq. (5). The pseudocode is provided in Alg. 1, where $\oslash$ denotes element-wise division and $\circ$ denotes Hadamard power, *i.e.*, element-wise exponentiation. Notably, all operations in Alg. 1 are differentiable, making Heterogeneity-aware OT suitable for end-to-end training.

3 Experiments

3.1 Implementation Details

Datasets. We evaluate our OTSurv on six public TCGA datasets: BLCA (n=359), BRCA (n=868), LUAD (n=412), STAD (n=318), CRC (n=296), and KIRC (n=340), following the data split in [25]. WSIs are cropped into 256×256 non-overlapping patches at 20× magnification (averaging 12,789 patches per slide). Patch features extracted by UNI [4] (1024-D) are projected to 256-D via the projector. Survival prediction uses disease-specific survival [17] as the label and is evaluated via 5-fold site-stratified cross-validation [8] with the concordance index (C-index).

Setup. We initialize the transport mass ratio at $\rho_0 = 0.1$ and progressively increase it to 1.0 in 10 epochs using a sigmoid schedule. The number of survival tokens is set to $K = 16$. Entropy regularization is set to $\lambda = 0.1$. OTSurv is trained for 50 epochs using AdamW, with an initial learning rate of 1×10^{-4} following cosine decay and a weight decay of 1×10^{-5}. The Cox loss is optimized with a batch size of 16. All experiments are on an RTX 3090 GPU (24G).

Table 1: Comparison of C-index Performance Across 6 TCGA Datasets for Different Survival Prediction Methods.

Method \ Dataset	BRCA	BLCA	LUAD	STAD	CRC	KIRC	Mean
Mean-Pooling	0.512±0.177	0.589±0.046	0.532±0.060	0.503±0.060	0.597±0.146	0.730±0.062	0.577
CLAM [18]	0.629±0.180	0.608±0.104	0.569±0.045	0.541±0.077	0.625±0.132	0.674±0.124	0.608
HIPT [3]	0.555±0.094	0.609±0.127	0.576±0.081	0.539±0.074	0.599±0.066	0.689±0.094	0.595
ABMIL [9,14]	0.567±0.092	0.551±0.055	0.564±0.044	0.561±0.043	0.657±0.117	0.675±0.121	0.596
TransMIL [23]	0.599±0.058	0.590±0.106	0.546±0.117	0.500±0.061	0.543±0.127	0.677±0.133	0.576
AttnMISL [33]	0.585±0.073	0.523±0.080	0.624±0.139	0.544±0.056	**0.725**±0.110	0.658±0.115	0.610
IB-MIL [16]	0.511±0.085	0.536±0.075	0.594±0.130	0.541±0.079	0.576±0.072	0.666±0.130	0.571
ILRA [31]	0.611±0.135	0.569±0.082	0.515±0.063	0.554±0.060	0.648±0.123	0.649±0.101	0.591
PANTHER [24]	**0.650**±0.139	0.590±0.084	0.575±0.046	0.504±0.082	0.641±0.133	0.702±0.124	0.610
OTSurv	0.621±0.071	**0.637**±0.065	**0.638**±0.077	**0.565**±0.057	0.667±0.111	**0.750**±0.149	**0.646**

3.2 Comparison with State-of-the-Art Methods

In Table 1, across six TCGA cancer datasets, OTSurv consistently outperforms Mean-Pooling, CLAM [18], HIPT [3], ABMIL [9,14], TransMIL [23], AttnMISL [33], IB-MIL [16], ILRA [31], and PANTHER [24]. Moreover, we perform log-rank tests [1] on high- and low-risk cohorts, defined by splitting the predicted risks at the median (50%). Fig. 2(a) presents Kaplan-Meier survival curves along with p-values for six cancer types, all of which are below the 0.05 significance threshold, thereby demonstrating effective risk stratification.

3.3 Ablation Studies

Design choices. Our ablation study, as in Table 2, shows that replacing the KL-based global constraint with an equality constraint ($\mathbf{C_G}$ ✗) and replacing the progressive partial local constraint with either a fixed partial constraint ($\mathbf{C_L}$ ✗ with fixed $\rho = 0.8$) or a fixed full constraint ($\mathbf{C_L}$ ✗ with fixed $\rho = 1.0$) both degrade performance – highlighting the necessity of both $\mathbf{C_G}$ and $\mathbf{C_L}$. We also evaluated several other model design choices: (i) Replacing our mass-based instance selection with a cost matrix-based hard thresholding method leads to performance degradation. (ii) Replacing the sigmoid ramp-up of ρ with a linear ramp-up remains effective; (iii) Substituting Cox loss with NLL survival loss [36] yields inferior performance, likely due to information loss when converting regression to classification; (iv) Omitting the linear projector and training on raw patch embeddings significantly reduces performance, underscoring the importance of dimension reduction.

OT Hyperparameters. We evaluate OT hyperparameters – including the initial mass ratio ρ_0, KL weight λ, ramp-up epochs T, the number of survival tokens K and the number of sampled patches per WSI N. Our results show that moderate variations have little impact, whereas extreme settings degrade performance, supporting our design choices.

Table 2: Ablations on model design choices and OT hyperparameters.

Method		Dataset						Mean
		BRCA	BLCA	LUAD	STAD	CRC	KIRC	
OTSurv		0.621±0.071	0.637±0.065	0.638±0.077	0.565±0.057	0.667±0.111	0.750±0.149	**0.646**
(1) Ablation on model design choices								
$\mathbf{C_G}$ ✓	$\mathbf{C_L}$ ✓ (Sigmoid ρ)	0.621±0.071	0.637±0.065	0.638±0.077	0.565±0.057	0.667±0.111	0.750±0.149	0.646
	$\mathbf{C_L}$ ✗ (Fix $\rho = 0.8$)	0.646±0.106	0.628±0.065	0.631±0.080	0.547±0.051	0.631±0.183	0.747±0.134	0.638
	$\mathbf{C_L}$ ✗ (Fix $\rho = 1.0$)	0.556±0.185	0.628±0.064	0.637±0.071	0.551±0.046	0.616±0.127	0.746±0.135	0.622
$\mathbf{C_G}$ ✗	$\mathbf{C_L}$ ✓ (Sigmoid ρ)	0.617±0.131	0.604±0.104	0.637±0.097	0.553±0.064	0.611±0.148	0.724±0.124	0.624
	$\mathbf{C_L}$ ✗ (Fix $\rho = 0.8$)	0.566±0.199	0.592±0.104	0.638±0.062	0.551±0.033	0.624±0.148	0.737±0.179	0.618
	$\mathbf{C_L}$ ✗ (Fix $\rho = 1.0$)	0.635±0.102	0.590±0.119	0.646±0.074	0.501±0.119	0.530±0.160	0.745±0.145	0.608
Patch Select.	Mass → Cost	0.589±0.136	0.612±0.079	0.637±0.048	0.515±0.093	0.559±0.148	0.722±0.171	0.606
Ramp-up	Sig. → Lin.	0.675±0.113	0.632±0.055	0.639±0.072	0.547±0.046	0.632±0.117	0.742±0.134	0.645
Loss	Cox → NLL	0.624±0.117	0.609±0.112	0.651±0.084	0.565±0.040	0.606±0.163	0.762±0.125	0.636
Projector	w/o Proj.	0.601±0.127	0.587±0.062	0.604±0.059	0.518±0.098	0.544±0.097	0.724±0.145	0.596
(2) OT Hyperparameters								
Initial Mass Ratio ρ_0	0.1 → 0.2	0.638±0.118	0.579±0.092	0.656±0.109	0.579±0.054	0.606±0.113	0.750±0.156	0.635
	0.1 → 0.05	0.639±0.119	0.570±0.077	0.666±0.111	0.561±0.077	0.570±0.094	0.761±0.147	0.628
KL Weight λ	0.1 → 0.2	0.615±0.159	0.621±0.056	0.583±0.048	0.552±0.102	0.685±0.192	0.755±0.113	0.635
Ramp-up Epoches T	10 → 20	0.626±0.105	0.612±0.112	0.645±0.081	0.560±0.040	0.606±0.159	0.773±0.125	0.637
	10 → 30	0.662±0.122	0.640±0.032	0.556±0.062	0.527±0.048	0.612±0.085	0.776±0.120	0.629
Survival Token K	16 → 8	0.659±0.085	0.630±0.066	0.636±0.068	0.548±0.053	0.619±0.121	0.767±0.115	0.643
	16 → 32	0.644±0.087	0.628±0.069	0.636±0.071	0.570±0.053	0.621±0.123	0.761±0.136	0.643
Patch N	All → 3000	0.646±0.106	0.628±0.065	0.631±0.080	0.547±0.051	0.631±0.183	0.747±0.134	0.638

3.4 Model Interpretation

In Fig. 2(b), we compare the heatmap of OTSurv with tissue-type map. The heatmap is computed as Attention $= Q \cdot |W_{\text{agg}}^{\top}|$, where $Q \in \mathbb{R}^{N \times K}$ is the transport matrix and $W_{\text{agg}} \in \mathbb{R}^{1 \times K}$ represents the aggregation weight of each survival token in f_{agg}. The tissue-type map is generated using a linear classifier trained on the CRC-100K dataset (F1 = 0.93, AUC = 0.89) with patch features extracted by UNI [4]. Besides, we average patch-level attentions across 9 tissue types in all patches of the TCGA-CRC dataset. As shown in the bottom row of Fig. 2(b), patches positioned further to the right have higher attention values (*i.e.*, greater impact on survival prediction), aligning with clinical findings. This demonstrates that OTSurv has the capability to capture biologically relevant patterns and support interpretable risk stratification.

4 Conclusion

We introduced OTSurv, a novel optimal transport-based MIL method for survival prediction that explicitly models dual-scale pathological heterogeneity – mirroring how pathologists assess diverse tissue patterns. Our approach formulates survival prediction as a heterogeneity-aware OT problem by incorporating a global long-tail constraint to model morphological distributions and a local uncertainty-aware constraint to select high-confidence patches. We then reformulate the problem as an unbalanced OT task, solved efficiently via a matrix scaling algorithm. Experiments show that OTSurv achieves state-of-the-art performance and captures biologically meaningful features.

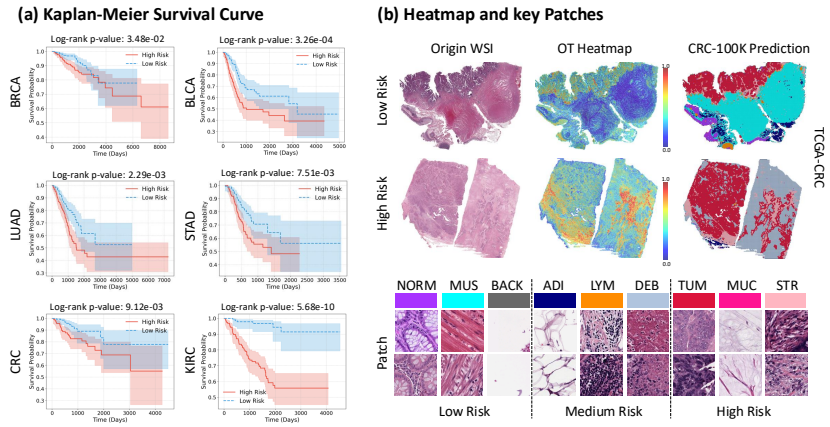

Fig. 2: (a) Kaplan–Meier survival curves, (b) OTSurv interpretability analysis.

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

References

1. Bland, J.M., Altman, D.G.: The logrank test. BMJ **328**(7447), 1073 (2004)
2. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine (2019)
3. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: CVPR (2022)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
5. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) (1972)
6. Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine (2012)
7. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: CVPR (2016)
8. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature Communications (2021)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML (2018)

10. Jin, C., Luo, L., Lin, H., Hou, J., Chen, H.: Hmil: Hierarchical multi-instance learning for fine-grained whole slide image classification. IEEE Transactions on Medical Imaging (2024)
11. Kantorovich, L.: On the transfer of masses (in russian). In: Doklady Akademii Nauk. vol. 37, p. 227 (1942)
12. Khamis, A., Tsuchida, R., Tarek, M., Rolland, V., Petersson, L.: Scalable optimal transport methods in machine learning: A contemporary survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
13. Knight, P.A.: The sinkhorn–knopp algorithm: convergence and applications. SIAM Journal on Matrix Analysis and Applications (2008)
14. Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., Yang, L.: Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: CVPR (2023)
15. Liero, M., Mielke, A., Savaré, G.: Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. Inventiones Mathematicae (2018)
16. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: CVPR (2023)
17. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.: An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell (2018)
18. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering (2021)
19. Ma, J., Xu, Y., Zhou, F., Wang, Y., Jin, C., Guo, Z., Wu, J., Tang, O.K., Zhou, H., Wang, X., et al.: Pathbench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology. arXiv preprint arXiv:2505.20202 (2025)
20. McGranahan, N., Swanton, C.: Clonal heterogeneity and tumor evolution: past, present, and the future. Cell (2017)
21. Ren, Q., Zhao, Y., He, B., Wu, B., Mai, S., Xu, F., Huang, Y., He, Y., Huang, J., Yao, J.: Iib-mil: Integrated instance-level and bag-level multiple instances learning with label disambiguation for pathological image analysis. In: MICCAI (2023)
22. Samuli, L., Timo, A.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
23. 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. In: NeurIPS (2021)
24. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: CVPR (2024)
25. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: ICML (2024)
26. Song, A.H., Williams, M., Williamson, D.F., Chow, S.S., Jaume, G., Gao, G., Zhang, A., Chen, B., Baras, A.S., Serafin, R., et al.: Analysis of 3d pathology samples using weakly supervised ai. Cell (2024)
27. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: ICCV (2023)

28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS (2017)
29. Villani, C., et al.: Optimal transport: old and new, vol. 338. Springer (2009)
30. Wang, X., Chen, Y., Zhu, W.: A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
31. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: ICLR (2023)
32. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
33. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis (2020)
34. You, C., Dai, W., Liu, F., Min, Y., Dvornek, N.C., Li, X., Clifton, D.A., Staib, L., Duncan, J.S.: Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
35. You, C., Dai, W., Min, Y., Liu, F., Clifton, D., Zhou, S.K., Staib, L., Duncan, J.: Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. In: NeurIPS (2023)
36. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
37. Zhang, C., Ren, H., He, X.: P^2 ot: Progressive partial optimal transport for deep imbalanced clustering. arXiv preprint arXiv:2401.09266 (2024)
38. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: CVPR (2022)