CACSurv: Concordance-Aligned Comparative Learning with Large Language Models for Cancer Survival Prediction

Tianqi Xiang, Qixiang Zhang, Xinpeng Ding, Yi Li, Xiaomeng Li*

The Hong Kong University of Science and Technology
eexmli@ust.hk

arXiv:2608.16594v1 [cs.AI] 17 Aug 2026

Abstract

Cancer survival prediction supports treatment planning, risk stratification, and follow-up management. Existing methods primarily use structured clinical variables, whole-slide images, genomic profiles, or multimodal combinations, while patient reports remain underexplored as primary inputs. We study report-centric survival prediction using reports that organize pathological, clinical, and molecular evidence into coherent text. Large language models (LLMs) can reason over such reports, but case-wise time regression introduces two fundamental mismatches. First, a formulation mismatch arises because survival evaluation depends on correctly ordering comparable patients, whereas independent time predictions do not enforce inter-patient ranking consistency. Second, a supervision mismatch arises because a censored patient's observed time only indicates survival beyond that point and cannot serve as an exact regression target, although it still implies orderings relative to patients who died earlier. To address these mismatches, we propose **CACSurv**, a Concordance-Aligned Comparative framework for report-centric survival prediction. CACSurv reformulates survival modeling as mini-cohort comparative reasoning, where an LLM predicts relative prognostic orderings. We further introduce concordance-aligned rewards derived from comparable relations under right censoring, enabling censored outcomes to provide ranking supervision without exact event-time targets. During inference, Monte Carlo Reference Aggregation compares each evaluation patient with sampled references and aggregates the resulting positions into a cohort-level ranking. We establish **TCGA-SurvReport**, a benchmark covering six TCGA cancer cohorts. CACSurv achieves the highest C-index on all six cohorts and an average C-index of 0.722, outperforming the strongest published survival model by 6.5 percentage points and the strongest LLM time-regression baseline by 4.2 percentage points. Our code, models, and dataset will be made publicly available at https://github.com/xmed-lab/CACSurv.

Introduction

Cancer survival prediction aims to characterize a patient's prognosis from clinical evidence, providing important support for treatment planning, risk stratification, and follow-up management. Existing computational approaches primarily learn from structured clinical variables, whole-slide images (WSIs), genomic profiles, or their multimodal combinations (Shao et al. 2021; Chen et al. 2021), typically producing patient-wise risk scores or survival estimates (Fig. 1(a)). Beyond these commonly studied modalities, patient reports provide a distinct textual interface that organizes clinically interpreted evidence, including diagnosis, staging, pathological findings, and molecular characteristics (Kefeli and Tatonetti 2024). Despite their information-rich nature, patient reports remain comparatively underexplored as the primary input for cancer survival prediction. In this work, we investigate report-centric cancer survival analysis, where unified patient reports are directly used to derive prognostic predictions.

Recent advances in large language models (LLMs) have enabled effective processing of long-form text and integration of heterogeneous information (Qwen et al. 2025), making them a natural candidate for report-centric cancer survival analysis. Recent regression-aware LLM studies have further established pointwise scalar prediction as a viable paradigm for language models (Lukasik et al. 2025; Zhang et al. 2026a). Following this direction and the patient-wise formulation of conventional survival models (Fig. 1(a)), a straightforward LLM adaptation is to process each report independently and generate an absolute survival time (Fig. 1(b)). However, the ability to perform numerical regression does not by itself provide a principled formulation for survival prediction. Although natural for generative models, this case-wise time-regression formulation is not well aligned with either the evaluation or supervision of survival prediction, giving rise to two fundamental mismatches.

First, there is a **formulation mismatch** between case-wise time regression and concordance-based survival evaluation. As shown in Fig. 1(b), a time-regression LLM makes separate scalar predictions for Patients A and B. However, survival models are commonly evaluated by the concordance index (C-index), which measures whether comparable patients are ordered correctly rather than the numerical accuracy of individual predictions. Independent time regression lacks explicit constraints on such inter-patient ordering. Moreover, absolute time regression must accommodate substantially different survival time scales across cancer types, as shown in Fig. 3. Thus, the time-regression formulation focuses on absolute numerical prediction, without directly aligning its objective with the relative prognostic ordering required by survival evaluation.

Second, there is a **supervision mismatch** between exact-time regression and incomplete survival outcomes. As shown

*Corresponding author.

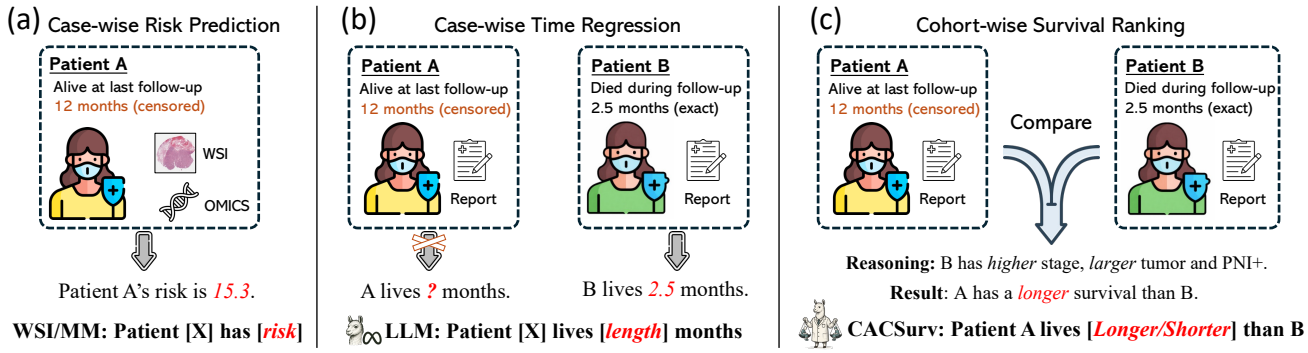

Figure 1: **Comparison of survival prediction formulations.** (a) Conventional WSI/multimodal models predict a case-wise risk score. (b) A straightforward report-based LLM performs case-wise time regression, but patients who remain alive at the last follow-up lack exact-time targets. (c) CACSurv directly ranks patients within a mini-cohort with reliable relative comparisons.

by Patient A in the left example of Fig. 1(b), being alive at the last follow-up of 12 months indicates only that the patient survived for at least 12 months, while the true survival time remains unknown. Such right-censored cases are common in survival analysis and account for over 60% of our benchmark (Fig. 3). Treating the follow-up time as an exact regression target introduces incorrect supervision, while excluding censored patients discards substantial prognostic information. Importantly, these patients still provide valid ordering relations with patients who died before their last follow-up. Thus, censored outcomes are better represented as partial-order supervision than as exact-time labels.

These two mismatches share a common underlying issue: absolute-time regression treats survival as an independent numerical prediction problem, whereas both concordance-based evaluation and the supervision available under right censoring are fundamentally relational. To this end, we propose **CACSurv**, a Concordance-Aligned Comparative framework for report-centric survival analysis. To handle the formulation mismatch, **CACSurv introduces a mini-cohort comparative framework that replaces absolute-time prediction with relative prognostic ordering.** The model receives a small group of patient reports and predicts their ordering through comparative reasoning (Fig. 1(c)). This formulation makes inter-patient ranking the native prediction target and directly aligns model outputs with concordance-based evaluation. During inference, Monte Carlo Reference Aggregation repeatedly compares each evaluation patient with subsets sampled from a shared cancer-specific reference bank and aggregates the resulting relative positions into a reference-relative survival score. These scores place all evaluation patients on a common scale for cohort-level ranking. To resolve the supervision mismatch, **CACSurv designs concordance-aligned rewards derived from valid comparable relations under right censoring.** Rather than treating the last follow-up time of a censored patient as an exact survival target, these rewards evaluate whether the model correctly predicts the reliable partial-order relations implied by the observed outcomes, allowing censored patients to contribute informative supervision without mislabeling their follow-up times as observed event times or excluding them from training. Together, CACSurv aligns its prediction formulation, supervision signal, and inference procedure with survival concordance. Our main contributions are as follows:

- **Comparative survival framework.** We reformulate report-centric LLM survival analysis from independent absolute-time generation into mini-cohort prognostic ordering, directly aligning model prediction with concordance-based evaluation.
- **Concordance-aligned learning and inference.** We introduce rewards that exploit valid partial-order supervision under right censoring and Monte Carlo Reference Aggregation that converts comparative predictions to cohort-level rankings.
- **Comprehensive benchmark and validation.** We establish TCGA-SurvReport, a six-cohort benchmark that integrates pathological, clinical, and molecular evidence into unified patient reports. CACSurv achieves the highest C-index on all six cohorts, with an average of 0.722, outperforming the strongest published survival model by 6.5 percentage points and the strongest LLM time-regression baseline by 4.2 percentage points.

Related Work

Unimodal Cancer Survival Analysis

Single-modality cancer survival analysis has been extensively studied using structured clinical variables and histopathology images. Traditional approaches apply statistical or deep survival models to structured clinical features (Cox 1972; Katzman et al. 2018; Lee et al. 2018). In computational pathology, WSI-based methods commonly aggregate patch-level representations through multiple instance learning (Lu et al. 2021; Shao et al. 2021), with recent studies further improving slide representation, heterogeneous tissue modeling, and uncertainty estimation (Ding et al. 2025; Wu et al. 2025; Xing et al. 2026). Despite their architectural differences, these methods generally process each patient independently and produce a patient-specific risk score, hazard estimate, or survival distribution.

Multimodal Cancer Survival Analysis

Multimodal cancer survival analysis integrates histopathology with genomic, transcriptomic, clinical, and textual information to capture complementary prognostic evidence. Early studies model cross-modal interactions through co-attention or optimal-transport-based alignment (Chen et al. 2021; Xu and Chen 2023), while pathway-aware approaches connect histological representations with biological pathways (Jaume et al. 2024). More recent methods explore structured cross-modal alignment and latent prognostic representations (Bu et al. 2026; Zhang et al. 2026b). Pathology reports and report-derived embeddings have also been incorporated alongside WSIs, molecular profiles, and clinical variables for multimodal survival prediction (Raza et al. 2025; Song et al. 2026).

These multimodal approaches primarily focus on representation learning and information fusion across heterogeneous modalities. When pathology reports are incorporated, they are used as an additional modality alongside WSIs, molecular profiles, and clinical variables. Recent studies have also explored LLM-based prognostic assessment from clinical notes or pathology reports through LLM-assisted extraction of prognostic variables (Kiermeyer et al. 2026), survival classification at a fixed time point (Phaterpekar et al. 2026), and categorical risk or prognosis prediction (Loeffler et al. 2026; Saluja et al. 2025). These studies address related prognostic tasks but differ from censoring-aware survival ranking in their prediction targets and evaluation protocols. CACSurv uses unified patient reports as direct inputs and formulates survival prediction as comparative ranking, replacing independent patient-wise prediction and aligning its supervision and inference with survival concordance under right censoring.

Methodology

In this section, we present CACSurv, a concordance-aligned comparative learning framework for report-centric survival prediction, as illustrated in Fig. 2. We first formulate the survival prediction problem and briefly review Group Relative Policy Optimization (GRPO) (Guo et al. 2025). We then introduce comparative mini-cohort ranking during training and Monte Carlo Reference Aggregation during inference, followed by the pair and triplet concordance-aligned rewards used to optimize CACSurv.

Preliminary

Problem setup. We consider a cohort of n patients and split it into a training set $\Psi = \{p_1, \ldots, p_{n_0}\}$ and an evaluation set $\Psi^* = \{p_1^*, \ldots, p_{n_1}^*\}$, where $n_0 + n_1 = n$. Each training patient $p_i \in \Psi$ is associated with a patient record x_i and a survival label (t_i, δ_i), where t_i denotes the observed time and $\delta_i \in \{0, 1\}$ indicates whether the event is observed ($\delta_i = 1$, patient dead) or the case is right-censored ($\delta_i = 0$, patient alive). For an evaluation patient $p_j^* \in \Psi^*$, the corresponding record and label are denoted by $(x_j^*, t_j^*, \delta_j^*)$. Only patient reports are provided to CACSurv as model inputs. The survival outcomes (t_i, δ_i) of training patients are used solely to construct comparable relations and compute concordance-aligned rewards, and are not included in the model prompt. For evaluation patients, (t_j^*, δ_j^*) are used only for performance evaluation after prediction. Right censoring provides only partial survival information, so we define a patient pair as *comparable* if and only if the patient with the shorter observed time is not censored. We evaluate prognostic performance using the concordance index (C-index), which measures the agreement between predicted preferences and comparable survival outcomes.

Group Relative Policy Optimization (GRPO). We employ GRPO (Guo et al. 2025) to optimize the model with reward-based signals. Given an input, the current policy θ samples a set of G candidate outputs $\{o_i\}_{i=1}^G$, each associated with a scalar reward R_i. GRPO forms a relative advantage by standardizing each reward within the sampled group:

$$A_i = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^G\right)}{\text{std}\left(\{R_j\}_{j=1}^G\right)}. \tag{1}$$

This group-normalized advantage is shared across all tokens in o_i, leading to a clipped surrogate objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)A_i,\ \bar{r}_{i,t}(\theta)\,A_i\Big), \tag{2}$$

where $r_{i,t}(\theta)$ is the token-wise likelihood ratio, $\bar{r}_{i,t}(\theta) = \text{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)$, and $|o_i|$ is the length of the i-th output. We omit the KL term here for brevity.

CACSurv for Comparative Survival Ranking

CACSurv reformulates report-centric survival prediction as a comparison-based ranking problem. Instead of treating each patient as an atomic unit and predicting an absolute survival time or continuous risk score, CACSurv uses mini-cohorts as the basic prediction units and directly learns relative survival orderings among patients.

Mini-cohort construction. A mini-cohort is a small set of patients sampled from the full cohort, used as the basic prediction unit in CACSurv. For training, we sample an index set $I_k \subseteq \{1, \ldots, n_0\}$ for the k-th mini-cohort C_k from the training set Ψ:

$$C_k = \{p_i : i \in I_k\}, \quad m_k = |I_k|. \tag{3}$$

Each mini-cohort consists of patients from the same cancer type, with no duplicate patients within the cohort.

For evaluation, we fix a cancer-type-specific reference bank $\psi \subseteq \Psi$ containing training patients of the same cancer type as the evaluation patient. For each evaluation patient $p_j^* \in \Psi^*$, we construct K evaluation mini-cohorts by combining p_j^* with a subset of reference patients from ψ:

$$C_{j,l}^* = \{p_j^*\} \cup R_{j,l}, \quad R_{j,l} \subseteq \psi, \quad l = 1, \ldots, K. \tag{4}$$

Given any ordering y over a mini-cohort C, we write $u \succ_y v$ if u is preferred to v under y, meaning u is predicted to have a longer survival than v.

Training phase. Given a training mini-cohort C_k, let $X_k = \{x_i : p_i \in C_k\}$ denote its corresponding patient reports. CACSurv predicts a survival ordering

$$\hat{y}_k = f_\theta(X_k), \tag{5}$$

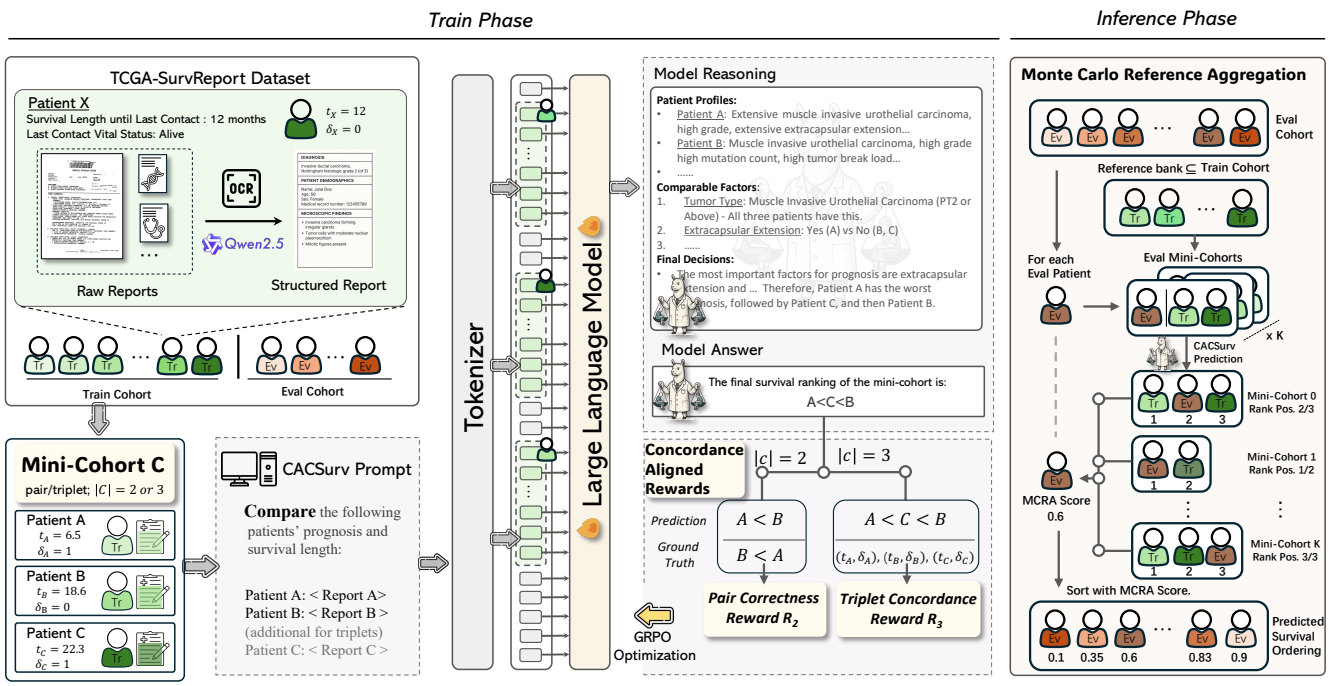

Figure 2: **Overview of CACSurv.** During training, CACSurv learns survival orderings over pair and triplet mini-cohorts using concordance-aligned rewards and GRPO. During inference, Monte Carlo Reference Aggregation constructs multiple mini-cohorts for each evaluation patient using randomly sampled reference patients, then averages its normalized rank positions to obtain the final cohort-level survival ordering.

where $\hat{y}_k$ orders the patients in C_k from shorter to longer predicted survival. We optimize f_θ with GRPO using the concordance-aligned rewards defined below. Repeated training over different mini-cohorts exposes the model to inter-patient survival relations and directly encourages concordance-consistent ranking behavior.

Inference phase (Monte Carlo Reference Aggregation). For each evaluation mini-cohort $C^*_{j,l}$, CACSurv takes the corresponding patient reports $X^*_{j,l}$ as input and predicts an ordering $\hat{y}^*_{j,l} = f_\theta(X^*_{j,l})$. We compute the normalized rank position of the evaluation patient p^*_j as:

$$s(p^*_j; C^*_{j,l}, \hat{y}^*_{j,l}) = \frac{\left|\left\{u \in C^*_{j,l} \setminus \{p^*_j\} \;:\; p^*_j \succ_{\hat{y}^*_{j,l}} u\right\}\right|}{|C^*_{j,l}| - 1}. \quad (6)$$

We aggregate its relative positions across K mini-cohorts:

$$S(p^*_j) = \frac{1}{K}\sum_{l=1}^{K} s(p^*_j; C^*_{j,l}, \hat{y}^*_{j,l}). \quad (7)$$

A larger $S(p^*_j)$ indicates longer predicted survival. The final full cohort-level ordering is obtained by sorting the evaluation patients in ascending order of $S(p^*_j)$.

Under uniform random sampling from the reference bank, each normalized rank position is a bounded Monte Carlo observation of the patient's reference-relative position. Therefore,

$$E[S(p^*_j)] = \mu_j = E_{R\sim\psi}[s(p^*_j; C_R, \hat{y}_R)], \quad (8)$$

and $S(p^*_j)$ converges to μ_j as K increases. When the predicted preferences are consistent across mini-cohorts, μ_j corresponds to the proportion of reference patients predicted to have shorter survival than p^*_j, providing a reference-relative approximation of its global position.

MCRA avoids direct comparisons among evaluation patients and assigns each patient a fixed number of comparisons independent of the evaluation cohort size, allowing it to support both small evaluation cohorts and sequential patient arrival.

Concordance-aligned Rewards

Our concordance-aligned rewards convert censored survival outcomes into reliable preference supervision. If a patient is right-censored at time t_i, the patient is known to have survival longer than any patient who died before t_i, even though the exact survival time remains unknown. Based on such comparable relations, we define a pair correctness reward for pair mini-cohorts and a triplet concordance reward for mini-cohorts of size three or beyond.

Pair correctness reward ($|C| = 2$). For a pair mini-cohort $C = \{p_i, p_j\}$, we ensure during construction that the two patients are comparable. Given a predicted ordering $\hat{y}$, we define the correctness reward as

$$R_2(C, \hat{y}) = \mathbb{1}[t_i < t_j] \cdot \mathbb{1}[p_j \succ_{\hat{y}} p_i] + \mathbb{1}[t_j < t_i] \cdot \mathbb{1}[p_i \succ_{\hat{y}} p_j]. \quad (9)$$

Since pair comparability is enforced during construction, the event indicator is omitted from Eq. 9.

Triplet concordance reward ($|C| \geq 3$). For a triplet mini-cohort C, right censoring may make only a subset of patient pairs comparable. We define the set of comparable directed relations within C as

$$\mathcal{V}(C) = \{(p_u, p_v) \in C \times C \ : \ u \neq v, \ t_v < t_u, \ \delta_v = 1\}. \tag{10}$$

The triplet concordance reward encourages ordering consistency over all comparable relations in $\mathcal{V}(C)$:

$$R_3(C, \hat{y}) = \begin{cases} \frac{1}{|\mathcal{V}(C)|} \sum_{(p_u, p_v) \in \mathcal{V}(C)} \mathbb{1}[p_u \succ_{\hat{y}} p_v], & |\mathcal{V}(C)| > 0, \\ 0, & |\mathcal{V}(C)| = 0. \end{cases} \tag{11}$$

Note that this formulation also naturally accepts mini-cohorts with more than three patients by defining $\mathcal{V}(C)$ over all within-cohort comparable relations.

Therefore, we define the unified concordance-aligned reward as

$$R(C, \hat{y}) = \mathbb{1}[|C| = 2] \cdot R_2(C, \hat{y}) + \mathbb{1}[|C| \geq 3] \cdot R_3(C, \hat{y}), \tag{12}$$

which is used as the reward signal for GRPO optimization during training.

Experiments

Dataset Construction and Cohort Statistics

We construct TCGA-SurvReport, a report-centric survival dataset covering six TCGA cancer cohorts: BLCA, BRCA, COADREAD, HNSC, LUAD, and STAD. For each patient, we collect OCR-derived pathology text from TCGA-Reports (Kefeli and Tatonetti 2024), clinical records, and molecular profiles, which provide complementary pathological, demographic, clinical, and molecular information. The three sources are aligned using TCGA patient identifiers.

The released OCR text follows heterogeneous report formats and contains duplicated content, administrative statements, and detailed specimen-processing descriptions. To address these issues, we use Qwen2.5-72B-Instruct with a fixed extractive prompt to remove irrelevant content and organize the retained information into a standardized patient report. The model is prohibited from adding diagnoses, medical facts, prognostic interpretations, or clinical implications absent from the source records. We manually reviewed a sampled subset of processed reports to assess their factual consistency with the corresponding source records.

Observed time and vital status are extracted separately as survival labels and are not included in the unified patient reports. We retain patients with both a valid unified report and an available survival record. Fig. 3 summarizes the cohort statistics across the six cancer types.

Baselines

We compare CACSurv with representative WSI-based and multimodal survival models, including TransMIL (Shao et al. 2021), TITAN (Ding et al. 2025), MCAT (Chen et al. 2021), MOTCat (Xu and Chen 2023), SurvPath (Jaume et al. 2024), PS3 (Raza et al. 2025), PAMoE (Wu et al. 2025), CIMA (Bu et al. 2026), SlotSPE (Zhang et al. 2026b), and DPSurv (Xing et al. 2026). These baselines cover WSI-only,

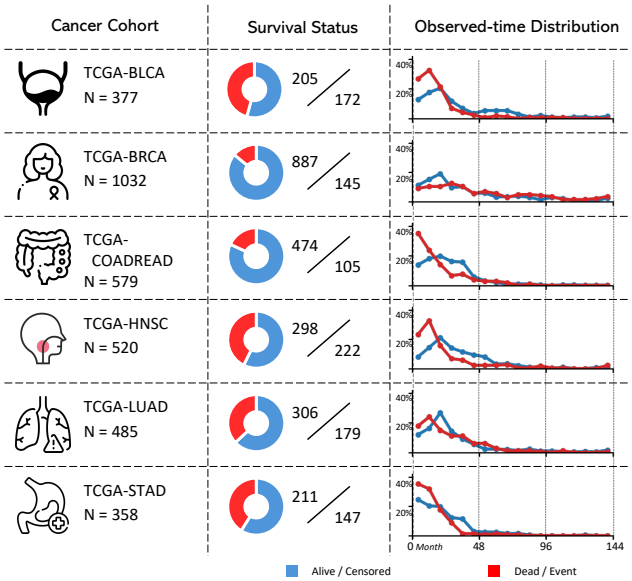

Figure 3: **Cohort statistics of TCGA-SurvReport.** For each cancer cohort, we report the number of patients, the proportions of alive/censored and dead/event cases, and their observed-time distributions. Observed time denotes the last follow-up time for censored patients and the event time for deceased patients. Each curve shows the percentage of patients within the corresponding status group in each time interval.

histology–molecular, and histology–report–molecular survival prediction settings. For report-centric comparison, we include Tabular-CoxPH (Cox 1972), Tabular-DeepHit (Lee et al. 2018), PubMedBERT-DeepSurv (Katzman et al. 2018), and BioMistral-CoxPH (Song et al. 2026). We further construct LLM-based time-regression baselines using zero-shot inference with Qwen2.5-7B/72B-Instruct (Qwen et al. 2025) and supervised fine-tuning with Qwen2.5-7B-Instruct.

Implementation and Evaluation Protocol

All methods are evaluated using the same patient cohorts and identical patient-level five-fold cross-validation splits. Models are trained only on the training patients of each fold, and performance is reported as the mean and standard deviation of the C-index across the five folds. All baseline results are obtained by rerunning the corresponding methods under this unified evaluation protocol.

For WSI-based methods, we use CONCHv1.5 (Lu et al. 2024) to extract patch features. PS3 uses PLIP (Huang et al. 2023) as it requires aligned image and text encoders. Molecular inputs are processed following the original implementation of each method. Tabular-CoxPH and Tabular-DeepHit use structured prognostic variables extracted from the reports, while PubMedBERT-DeepSurv and BioMistral-CoxPH use report embeddings produced by PubMedBERT (Gu et al. 2021) and BioMistral (Labrak et al. 2024), respectively.

CACSurv uses Qwen2.5-7B-Instruct as the backbone. For

Method	BLCA	BRCA	COADREAD	HNSC	LUAD	STAD	AVG@6
WSI/Multimodal Survival Models							
TransMIL	$0.571_{\pm 0.083}$	$0.640_{\pm 0.056}$	$0.642_{\pm 0.084}$	$0.568_{\pm 0.068}$	$0.548_{\pm 0.027}$	$0.599_{\pm 0.052}$	$0.595_{\pm 0.039}$
TITAN	$0.625_{\pm 0.065}$	$0.666_{\pm 0.033}$	$0.649_{\pm 0.060}$	$0.586_{\pm 0.033}$	$0.607_{\pm 0.048}$	$0.548_{\pm 0.074}$	$0.613_{\pm 0.043}$
PAMoE	$0.646_{\pm 0.054}$	$0.704_{\pm 0.038}$	$0.702_{\pm 0.050}$	$0.620_{\pm 0.032}$	$0.646_{\pm 0.035}$	$0.612_{\pm 0.070}$	$0.655_{\pm 0.036}$
DPSurv	$0.611_{\pm 0.064}$	$0.600_{\pm 0.055}$	$0.585_{\pm 0.089}$	$0.531_{\pm 0.074}$	$0.570_{\pm 0.020}$	$0.536_{\pm 0.041}$	$0.572_{\pm 0.033}$
MCAT	$0.620_{\pm 0.012}$	$0.564_{\pm 0.026}$	$0.612_{\pm 0.052}$	$0.518_{\pm 0.024}$	$0.571_{\pm 0.031}$	$0.528_{\pm 0.075}$	$0.569_{\pm 0.042}$
MOTCat	$0.623_{\pm 0.012}$	$0.570_{\pm 0.041}$	$0.617_{\pm 0.030}$	$0.548_{\pm 0.018}$	$0.561_{\pm 0.022}$	$0.531_{\pm 0.075}$	$0.575_{\pm 0.050}$
SurvPath	$0.581_{\pm 0.038}$	$0.623_{\pm 0.086}$	$0.657_{\pm 0.095}$	$0.539_{\pm 0.032}$	$0.571_{\pm 0.035}$	$0.562_{\pm 0.075}$	$0.589_{\pm 0.043}$
PS3	$0.638_{\pm 0.075}$	$0.632_{\pm 0.056}$	$0.613_{\pm 0.083}$	$0.622_{\pm 0.023}$	$0.658_{\pm 0.033}$	$0.556_{\pm 0.049}$	$0.620_{\pm 0.061}$
CIMA	$0.605_{\pm 0.082}$	$0.607_{\pm 0.010}$	$0.643_{\pm 0.054}$	$0.576_{\pm 0.027}$	$0.609_{\pm 0.034}$	$0.542_{\pm 0.092}$	$0.597_{\pm 0.034}$
SlotSPE	$0.658_{\pm 0.007}$	$0.709_{\pm 0.009}$	$0.710_{\pm 0.006}$	$0.607_{\pm 0.016}$	$0.656_{\pm 0.016}$	$0.601_{\pm 0.009}$	$0.657_{\pm 0.045}$
Report-centric Survival Models							
Tabular-CoxPH	$0.623_{\pm 0.046}$	$0.619_{\pm 0.071}$	$0.683_{\pm 0.045}$	$0.591_{\pm 0.035}$	$0.654_{\pm 0.048}$	$0.598_{\pm 0.085}$	$0.628_{\pm 0.032}$
Tabular-DeepHit	$0.618_{\pm 0.034}$	$0.651_{\pm 0.046}$	$0.696_{\pm 0.041}$	$0.589_{\pm 0.034}$	$0.671_{\pm 0.034}$	$0.621_{\pm 0.078}$	$0.641_{\pm 0.036}$
PubMedBERT-DeepSurv	$0.493_{\pm 0.092}$	$0.609_{\pm 0.062}$	$0.596_{\pm 0.051}$	$0.564_{\pm 0.072}$	$0.595_{\pm 0.025}$	$0.507_{\pm 0.029}$	$0.561_{\pm 0.045}$
BioMistral-CoxPH	$0.489_{\pm 0.046}$	$0.582_{\pm 0.052}$	$0.615_{\pm 0.037}$	$0.515_{\pm 0.047}$	$0.528_{\pm 0.061}$	$0.539_{\pm 0.066}$	$0.545_{\pm 0.046}$
Qwen2.5-7B Time-ZS	$0.504_{\pm 0.009}$	$0.693_{\pm 0.036}$	$0.669_{\pm 0.024}$	$0.518_{\pm 0.010}$	$0.573_{\pm 0.035}$	$0.553_{\pm 0.029}$	$0.585_{\pm 0.072}$
Qwen2.5-72B Time-ZS	$0.646_{\pm 0.053}$	$0.711_{\pm 0.051}$	$0.745_{\pm 0.043}$	$0.646_{\pm 0.025}$	$0.663_{\pm 0.060}$	$0.670_{\pm 0.050}$	$0.680_{\pm 0.047}$
Qwen2.5-7B Time-SFT	$0.629_{\pm 0.049}$	$0.607_{\pm 0.093}$	$0.649_{\pm 0.064}$	$0.620_{\pm 0.048}$	$0.615_{\pm 0.020}$	$0.634_{\pm 0.038}$	$0.626_{\pm 0.013}$
CACSurv-7B (Pair-only)	$\underline{0.661_{\pm 0.060}}$	$\underline{0.745_{\pm 0.074}}$	$\underline{0.766_{\pm 0.057}}$	$\underline{0.667_{\pm 0.040}}$	$\underline{0.687_{\pm 0.046}}$	$\underline{0.683_{\pm 0.054}}$	$\underline{0.702_{\pm 0.040}}$
CACSurv-7B	$\mathbf{0.685_{\pm 0.061}}$	$\mathbf{0.763_{\pm 0.063}}$	$\mathbf{0.806_{\pm 0.058}}$	$\mathbf{0.689_{\pm 0.046}}$	$\mathbf{0.693_{\pm 0.064}}$	$\mathbf{0.698_{\pm 0.047}}$	$\mathbf{0.722_{\pm 0.046}}$

Table 1: Main results from 5-fold cross-validation on TCGA-SurvReport (C-index, mean±std). Time-ZS and Time-SFT denote direct survival-time prediction using zero-shot inference and supervised fine-tuning, respectively. Pair-only uses only pair mini-cohorts during both training and inference. Best results are shown in **bold**, and second-best results are underlined.

each fold, we sample 1,000 pair and 1,000 triplet mini-cohorts from each cancer type and optimize the model for one epoch. Each pair is constructed to be comparable, and each triplet contains at least one comparable patient pair. During training, observed time and vital status are used only to enforce these sampling constraints and compute the concordance-aligned rewards; only patient reports are provided to the model. During inference, the training patients form the reference bank, and each test patient is compared against $K = 40$ randomly sampled reference mini-cohorts. Test survival outcomes, including observed time and vital status, are used only after prediction to calculate the final C-index.

Comparisons with State-of-the-Art Methods

Table 1 presents the survival prediction results across the six cancer cohorts. CACSurv achieves the highest C-index on every cohort and the best average C-index of 0.722. Its simplified pair-only variant consistently ranks second, with an average C-index of 0.702.

Among existing published survival models, SlotSPE achieves the strongest average performance of 0.657. CACSurv exceeds SlotSPE by 6.5 percentage points and consistently outperforms the evaluated WSI-based, multimodal, and conventional report-centric survival models. These results highlight the strong prognostic value of unified patient reports and the effectiveness of comparative survival modeling.

We further construct LLM-based time-regression baselines that directly predict absolute survival time from each report. CACSurv with a 7B backbone outperforms the strongest of these baselines, Qwen2.5-72B Time-ZS, by 4.2 percentage points, indicating that comparative prediction is more effective for this task than direct absolute-time regression.

Setting	Eval.Pair	Eval.Mix
Train.Pair	0.702	0.703
Train.Mix	0.715	**0.722**

Table 2: Ablation on pair and triplet mini-cohort instantiations during training and inference. Mix means Pair+Triplet.

Optimization	BLCA	COADREAD	HNSC	AVG@6
Rank-SFT	0.656	0.740	0.653	0.682
Ours	**0.685**	**0.806**	**0.689**	**0.722**

Table 3: Ablation of ranking-based SFT and our concordance -aligned reward optimization under the same comparative formulation.

Ablation on Pair and Mixed Mini-Cohorts

We evaluate pair and triplet mini-cohort compositions during both training and inference in Table 2. Train.Pair trains with 2,000 pair mini-cohorts, whereas Train.Mix trains with 1,000 pairs and 1,000 triplets, keeping the total number of training

mini-cohorts, backbone, data splits, and training epochs unchanged. Pair-and-triplet training consistently outperforms pair-only training, improving the average C-index from 0.702 to 0.715 under pair inference and from 0.703 to 0.722 under pair-and-triplet inference. The best performance is achieved when mixed mini-cohorts are used in both stages, indicating that triplet relations provide complementary comparative supervision and can be further exploited during inference.

Ablation on Reward-based Optimization

We compare Rank-SFT and concordance-aligned reward optimization under identical settings, varying only the optimization method. Rank-SFT directly fine-tunes the model to generate the target ranking, whereas CACSurv optimizes the concordance-aligned rewards with GRPO. Table 3 reports three representative cohorts and the six-cohort average for compactness, while CACSurv consistently outperforms Rank-SFT across all six cohorts. Overall, concordance-aligned reward optimization improves the average C-index from 0.682 to 0.722, demonstrating its effectiveness beyond directly supervising the target ranking.

Setting	$K = 10$	$K = 20$	$K = 40$	$K = 80$
Train.Pair, Eval.Pair	0.692	0.700	0.702	0.704
Train.Mix, Eval.Pair	0.701	0.707	0.715	0.718
Train.Mix, Eval.Mix	**0.706**	**0.715**	**0.722**	**0.723**

Table 4: Sensitivity to the number of reference comparisons K during inference. Mix means Pair+Triplet.

Sensitivity to Number of Reference Comparisons

Table 4 reports the sensitivity to the number of reference comparisons K. We keep the trained models and all other inference settings unchanged for fairness. Performance improves consistently as K increases and largely converges at $K = 40$. Under CACSurv's default setting (Train.Mix and Eval.Mix), further increasing K from 40 to 80 only improves the C-index from 0.722 to 0.723. We therefore adopt $K = 40$ as an acceptable accuracy-complexity tradeoff.

CACSurv Rollout Example

Fig. 4 presents a CACSurv rollout example for a pair mini-cohort. The model first identifies patient-specific prognostic factors, then compares the two patients based on the extracted evidence, and finally produces the survival ranking. This example shows how CACSurv converts report evidence into an explicit comparative prediction.

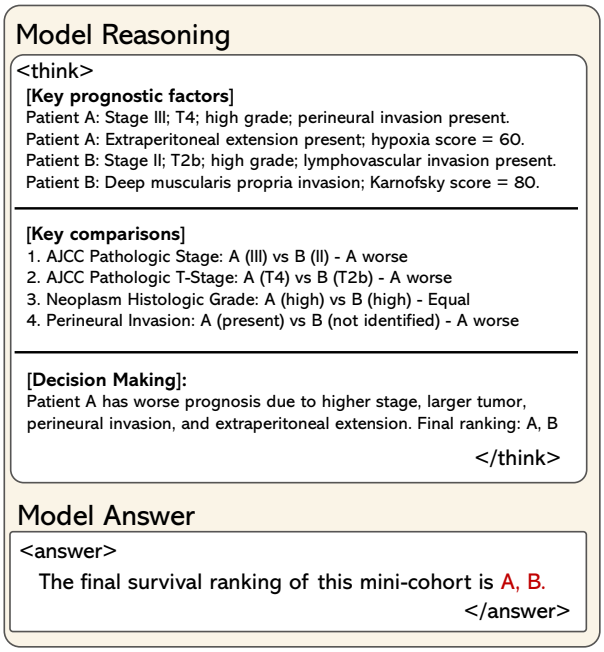

Figure 4: **Example CACSurv rollout on a pair mini-cohort.** In the `<think>` section, the model identifies patient-specific prognostic factors, compares discriminative evidence across patients, and derives a ranking decision. The predicted survival ordering is then returned in the `<answer>` section.

Conclusion

We present CACSurv, a concordance-aligned comparative framework for report-centric cancer survival prediction. CACSurv addresses the formulation and supervision mismatches of conventional LLM-based time regression by directly predicting prognostic orderings within mini-cohorts and optimizing concordance-aligned rewards derived from valid comparable relations under right censoring. Monte Carlo Reference Aggregation further converts these local comparative predictions into stable cohort-level survival rankings. We also establish TCGA-SurvReport, a unified report-centric benchmark spanning six TCGA cancer cohorts. Under identical patient-level five-fold cross-validation splits, CACSurv achieves the best performance among all evaluated methods on every cohort, with an average C-index of 0.722, exceeding the strongest published survival model by 6.5 percentage points and the strongest LLM time-regression baseline by 4.2 percentage points. Further analyses show that pair and triplet mini-cohorts provide complementary benefits during training and inference, and concordance-aligned reward optimization consistently outperforms ranking-based SFT. These findings demonstrate that aligning the prediction target, learning signal, and inference procedure with survival concordance provides an effective approach to LLM-based survival analysis under right censoring.

References

Bu, Y.; Niu, Q.; Li, Z.; Xu, Y.; Wang, J.; and Yu, G. 2026. Cancer survival prediction by cyclic generation and multi-grained alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 19781–19789.

Chen, R. J.; Lu, M. Y.; Weng, W.-H.; Chen, T. Y.; Williamson, D. F.; Manz, T.; Shady, M.; and Mahmood, F. 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4025.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the royal statistical society: Series B (methodological)*, 34(2): 187–202.

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. 2025. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 1–13.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081): 633–638.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T. J.; and Zou, J. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9): 2307–2316.

Jaume, G.; Vaidya, A.; Chen, R. J.; Williamson, D. F.; Liang, P. P.; and Mahmood, F. 2024. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11579–11590.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1): 24.

Kefeli, J.; and Tatonetti, N. 2024. TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. *Patterns*, 5(3).

Kiermeyer, N.; Lenfers, T.; Dada, A.; Friedrich, J.; Khattab, S.; Knop, E.; Egger, J.; Pauly, M.; Jung, A.; Montavon, G.; et al. 2026. Large language models enable prognostic stratification of cancer patients using real-world clinical notes. *PLOS digital health*, 5(7): e0001546.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, 5848–5864.

Lee, C.; Zame, W.; Yoon, J.; and Van Der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Loeffler, C. M.; Reitsam, N. G.; Wolf, F.; Stueker, E. H.; Muti, H. S.; Wiest, I. C.; and Kather, J. N. 2026. SCRIPT: Stratified clinical risk prediction from pathology reports using large language models. *Journal of Pathology Informatics*, 22: 100673–100673.

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. 2024. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3): 863–874.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6): 555–570.

Lukasik, M.; Meng, Z.; Narasimhan, H.; Chang, Y.-W.; Menon, A. K.; Yu, F.; and Kumar, S. 2025. Better autoregressive regression with LLMs via regression-aware fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Phaterpekar, T.; Zeng, Z.; Mali, Y.; Leung, B.; Ho, C.; Ng, R.; Bates, A.; and Nunez, J. 2026. Investigating fine-tuning versus zero-shot learning for general large language models when predicting cancer survival from initial oncology consultation documents. *ESMO Real World Data and Digital Oncology*, 12: 100703–100703.

Qwen; Yang, A.; Yang, B.; Zhang, B.; et al. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Raza, M.; Azam, A.; Qaiser, T.; and Rajpoot, N. 2025. Ps3: A multimodal transformer integrating pathology reports with histology images and biological pathways for cancer survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22175–22186.

Saluja, R.; Rosenthal, J.; Windon, A.; Artzi, Y.; Pisapia, D. J.; Liechty, B. L.; and Sabuncu, M. R. 2025. Cancer type, stage and prognosis assessment from pathology reports using LLMs. *Scientific Reports*, 15(1): 27300.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34: 2136–2147.

Song, S.; Borjigin-Wang, M.; Madejski, I. R.; and Grossman, R. L. 2026. Multimodal Cancer Modeling in the Age of Foundation Model Embeddings. In Argaw, P.; Zhang, H.; Jabbour, S.; Chandak, P.; Ji, J.; Mukherjee, S.; Salaudeen, O.; Chang, T.; Healey, E.; Gröger, F.; Adibi, A.; Hegselmann, S.; Wild, B.; and Noori, A., eds., *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, 202–227. PMLR.

Wu, J.; Chen, M.; Ke, X.; Xun, T.; Jiang, X.; Zhou, H.; Shao, L.; and Kong, Y. 2025. Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5144–5153.

Xing, Y.; Huang, L.; Ma, J.; Hong, R.; Qiu, J.; Liu, P.; He, K.; Fu, H.; and Feng, M. 2026. DPsurv: Dual-Prototype Evidential Fusion for Uncertainty-Aware and Interpretable Whole Slide Image Survival Prediction. In *Forty-third International Conference on Machine Learning*.

Xu, Y.; and Chen, H. 2023. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21241–21251.

Zhang, Y.; Chen, T.; Zhou, M.; Leong, O.; Wu, Y. N.; and Lukasik, M. 2026a. REAL: Regression-Aware Reinforcement Learning for LLM-as-a-Judge. In *Forty-third International Conference on Machine Learning*.

Zhang, Y.; Nanbo, L.; Yang, C.; Schmidhuber, J.; and Gao, X. 2026b. Structural Prognostic Event Modeling for Multimodal Cancer Survival Analysis. In *The Fourteenth International Conference on Learning Representations*.