Semantic-Anchored Evidential Fusion for Domain-Robust Whole-Slide Survival Analysis

Yucheng Xing[1,2], Ling Huang[1,3†], Pei Liu[4], Jingying Ma[1], Jiaqing Xu[1], Kai He[1], and Mengling Feng[1]

[1] National University of Singapore, Singapore
[2] National University of Singapore Guangzhou Research Translation and Innovation Institute, Guangzhou, Guangdong, China
[3] Imperial College London, London, UK
[4] Hunan University, Changsha, China

Abstract. Whole-slide images (WSIs) are widely used for computational cancer prognosis. However, current research primarily focuses on in-domain performance and fails to generalize across clinical centers. This limitation stems from their reliance on pixel-derived representations that are highly susceptible to domain-specific artifacts caused by staining protocols and scanner hardware. We hypothesize that high-level pathology semantics, such as tumor grade and micro-environmental architecture, provide a domain-invariant semantic representation that mirrors the robust diagnostic logic of human pathologists. Therefore, we propose a Semantic-Anchored Evidential Fusion Survival (SAEFS) framework, where SAEFS derives semantic anchors from WSIs via Visual Question Answering (VQA), employs a dual-stream WSI evidence extraction architecture, uses Dirichlet-based Subjective Logic to model uncertainty, and fuses semantic and visual evidence through a cautious conjunction rule to avoid overconfident fusion from correlated sources. Trained exclusively on one source domain and evaluated zero-shot across four unseen domains, SAEFS consistently outperforms state-of-the-art models both in prediction accuracy and reliability, improving the average C-index by 10.2%. Quantitative analyses further show that VQA-derived semantic features exhibit significantly lower cross-center divergence than pixel-derived features, highlighting their robustness for cross-center clinical applications. The code is available at `https://github.com/YuchengXing99/SAEFS`.

Keywords: Computational pathology · Survival prediction · Cross-center domain shift · Evidential deep learning · Uncertainty quantification · Cautious belief fusion · Semantic-Anchored evidence · Visual question answering

1 Introduction

Whole-slide images (WSIs) are gigapixel-scale digital scans of histopathology tissue sections that capture rich microscopic details, including cellular morphology,

† Corresponding author: Ling Huang <my@linghuang.co.uk>.

tissue architecture, and tumor micro-environments [24, 47]. These comprehensive pathological features enable clinicians to assess tumor progression and estimate patient prognosis, a task commonly referred to as survival analysis [3, 37]. Since accurate survival prediction can provide critical guidance for personalized treatment planning and therapeutic decision-making, there has been growing interest in developing computational approaches to automate this process from WSIs [3, 21, 62].

To handle the extreme size of WSIs, attention-based multiple instance learning (MIL) has become the dominant paradigm [19, 35, 45]. By aggregating thousands of tissue patches into slide-level representations, these models capture complex morphological patterns, such as tumor cellularity and stromal architecture, that correlate with patient prognosis. However, these models are predominantly developed and validated within single-center cohorts, leaving their cross-center generalizability largely unexamined. In practice, WSIs collected from different centers inevitably exhibit substantial domain shifts due to heterogeneous staining protocols, diverse scanner hardware, and varying laboratory pipelines [49]. As illustrated in Fig. 1(a), state-of-the-art MIL models such as TransMIL trained on The Cancer Genome Atlas (TCGA) suffer an average C-index drop of approximately 0.19 when evaluated zero-shot on the external cohort National Lung Screening Trial (NLST) [48], revealing a significant gap between benchmark and real-world performance.

Existing methods to address domain shift often rely on feature-level stain normalization [36, 40, 51], which standardizes visual appearance through color space transformations but does not account for deeper structural and scanner-level variations [49]. Domain adaptation offers a more principled alternative. Supervised approaches explicitly align source and target distributions during training [30, 46], while source-free methods [29, 59] relax this constraint by adapting a pre-trained model using only unlabeled target data at inference time. Nevertheless, both paradigms fundamentally assume that a sufficient amount of target-domain data is accessible, either during training or adaptation. However, in real-world clinical deployment, the target data is typically unseen or completely unavailable due to privacy regulations and logistical constraints, making these approaches impractical.

We argue that this failure of cross-center transfer stems from a fundamental feature of entanglement: standard encoders conflate prognostically relevant morphology with domain-specific confounders such as staining intensity and scanner color profiles. Indeed, recent work has shown that current pathology foundation models encode center-of-origin signatures more strongly than biological signals [22], confirming that visual representations remain fragile to domain-level confounders (Fig. 1(b)). In contrast, high-level pathology semantics, including descriptions of tumor grade, necrosis patterns, and the immune microenvironment, are inherently consistent across centers [42], mirroring clinical reality where a pathologist's expertise transfers because they reason via stable biological concepts rather than low-level pixel statistics. However, such semantic annotations have traditionally relied on expert pathologists to provide slide-level

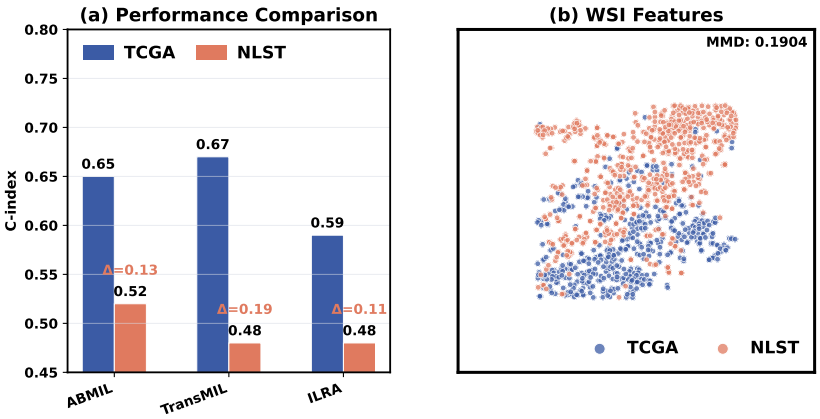


Fig. 1: Domain-shift robustness analysis for lung adenocarcinoma under TCGA to NLST transfer. (a) C-index of ABMIL [19], TransMIL [45], and ILRA [54] on the source domain (TCGA) and target domain (NLST), where Δ denotes the performance drop after transfer. (b) t-SNE visualization of cross-domain WSI features. Maximum Mean Discrepancy (MMD) values in (b) quantify the domain discrepancy in each feature space.

descriptions, making them prohibitively expensive and difficult to scale across large multi-center cohorts. These limitations motivate us to ask: *can we build a survival model that automatically extracts domain-invariant semantic anchors and leverages them to alleviate cross-center distribution shift, without requiring any target-domain data?*

We answer this question affirmatively by proposing the **Semantic-Anchored Evidential Fusion Survival (SAEFS)** framework. To automatically obtain domain-invariant semantics at scale, SAEFS queries a pathology-specialized VQA model with template-structured prognostically relevant questions for each WSI, yielding a bounded, hallucination-resistant semantic space without manual annotation. Built upon these semantic anchors, SAEFS adopts a dual-stream architecture with a MIL stream for global visual context and a semantic-guided stream attending to prognosis-relevant patches, whose outputs are combined via Adaptive Evidence Mixing. The fused visual evidence and semantic evidence are then converted into belief masses and uncertainty through Dirichlet-based Subjective Logic [43], and then integrated by a cautious conjunction rule [7] that accounts for evidence dependency. When evaluated zero-shot across four unseen centers, SAEFS consistently outperforms strong unimodal and multimodal baselines, with the most significant gains observed in high distribution-shift cohorts. Our contributions are:

1. To the best of our knowledge, this is the first work to systematically study cross-center domain shift in the context of WSI-based survival analysis and propose a target-free solution that is robust to unseen target domain data.

2. We introduce a template-based VQA pipeline that automatically extracts domain-invariant semantic features from a pathology VQA model, effectively anchoring prognostic feature extraction and offering a new paradigm for robust cross-center WSI analysis.
3. We leverage cautious belief fusion to address the challenge of fusing correlated visual and semantic modalities, explicitly modeling inter-modal reliability and uncertainty to enable robust evidence integration under domain shift.

2 Related Work

WSI-based Survival Prediction. Whole-slide survival research mainly builds on the multiple instance learning (MIL) paradigm with pure visual operation, in which a slide is represented as a set of patch-level instances and an aggregation module is used to produce a slide-level risk score [31, 56]. ABMIL [19] introduced a gated attention mechanism for permutation-invariant aggregation, and CLAM [35] extended it with instance-level clustering and a data-efficient training pipeline. TransMIL [45] replaced the attention pooling with a Transformer encoder to capture inter-patch correlations, while DTFD-MIL [60] proposed a dual-tier feature distillation strategy for more discriminative bag representations. Beyond pure visual methods, multimodal approaches fuse histopathology features with genomic or transcriptomic data, e.g., MCAT [2] employs cross-modal attention between pathology and omic embeddings, SurvPath [20] models pathway-level interactions with a multimodal Transformer. More recently, the pathology foundation models pre-trained on large-scale slide collections, such as UNI [1] and CONCH [32], have provided stronger patch encoders. Despite these advances, most current methods are developed on single-center or pooled multi-center datasets and obscure the impact of domain distributional shift. Consequently, these methods suffer from substantial performance drops when applied across different data centers, as demonstrated in Sec. 4.

Domain Shift in Histopathology Domain shift in histopathology, driven by heterogeneous staining protocols and hardware-specific scanner characteristics, frequently degrades model generalization and introduces diagnostic bias. Traditional domain shift interventions primarily operate at the image level. Stain normalization techniques [36, 40, 51] attempt to map slides to a canonical color space, stain augmentation [49] seeks to desensitize models to color variations via synthetic perturbations. At the representation level, large-scale self-supervised learning (SSL) has emerged as a promising strategy for implicit domain generalization. State-of-the-art foundation models, such as CTransPath [53], Phikon [9], and UNI [1], leverage massive multi-center datasets (>100k slides) to learn robust features. However, these encoders prioritize general feature quality over explicit domain invariance; as we demonstrate in Sec. 4.4, even these high-capacity feature spaces exhibit significant cross-center divergence, suggesting that scale alone does not guarantee domain robustness. General adaptation methods [28–30, 46, 52, 59] aim to mitigate distribution shifts during deployment

by dynamically updating model statistics. However, such approaches often require access to target-domain batches, a requirement frequently unmet in clinical workflows where slides are processed asynchronically and individually.

Vision-Language Models and VQA in Pathology. The success of contrastive vision–language pre-training (CLIP [38]) has catalyzed a wave of domain-specific adaptations in pathology. Early efforts focused on fine-tuning CLIP on pathology-specific pairs harvested from medical social media (PLIP [17]) or large-scale curation of PubMed captions (CONCH [32], QUILT-1M [18]). These semantic stability attributes of those models enable robust zero-shot classification and cross-modal retrieval by aligning visual features with textual descriptions. On the generative front, architectures like LLaVA-Med [27] and PathChat [33] have extended these capabilities to biomedical VQA and open-ended slide-level dialogue. Furthermore, MI-Zero [34] demonstrates the utility of CLIP-style embeddings for WSI-level classification using textual prompts. Despite these advancements, existing Vision–Language Models (VLMs) are primarily utilized for classification, retrieval, or descriptive report generation but lack application for cancer survival analysis.

Uncertainty Estimation and Evidential Deep Learning. Quantifying predictive uncertainty is paramount for safe clinical deployment [13, 14, 16, 55, 57]. While Bayesian approaches, such as MC Dropout [10] and Deep Ensembles [25], estimate uncertainty via stochastic sampling, they incur significant computational overhead during inference. In contrast, Evidential Deep Learning (EDL) [43] provides a deterministic, single-pass alternative by placing a Dirichlet prior over the categorical distribution. Grounded in Subjective Logic [23], EDL interprets the predicted concentration parameters as evidence supporting different outcomes, offering a principled calculus for reasoning under epistemic uncertainty. Recent efforts have extended EDL to multi-view classification [12] and survival regression [21]. However, conventional fusion operators such as Dempster's combination rule [6, 44] assume source independence, which is frequently violated in domain shift where modalities share correlated information, causing overconfident predictions.

3 Method

The proposed SAEFS architecture is illustrated in Fig. 2, composing semantic anchor generation, dual-stream semantic-anchored visual evidence extraction, semantic evidence extraction, and cautious belief fusion for survival predictions.

3.1 Semantic Anchor Generation

Semantic descriptions are more domain-robust in that they encode high-level pathology concepts that are more stable than raw visual features under distribution shift. We first exploit the VQA capability to automatically generate domain-robust semantic anchors.

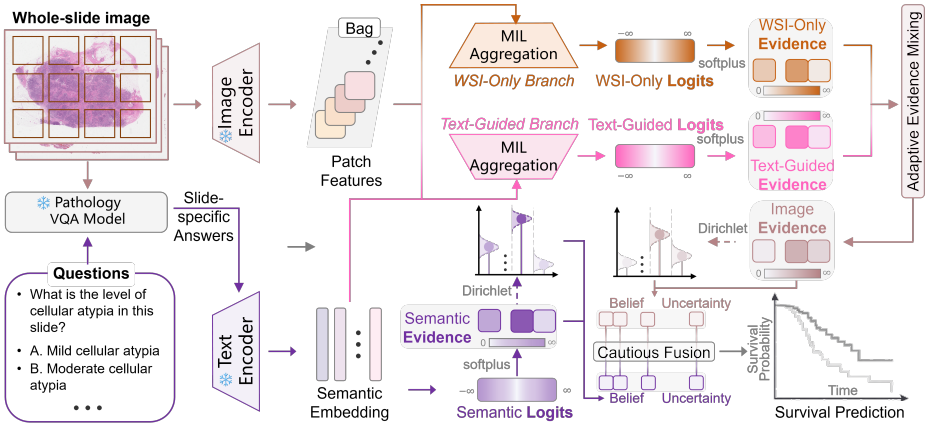


Fig. 2: Overview of the proposed framework. Template-based VQA answers serve as semantic anchors to extract text-guided visual evidence from the WSI. A parallel WSI-only branch provides complementary visual evidence. The two visual evidence streams are adaptively mixed, then fused with pure text evidence through Dirichlet-based cautious belief fusion that preserves uncertainty under modality disagreement.

Template Question Generation. To define a structured semantic space, we construct a fixed set of Q clinically grounded template questions in closed-form multiple-choice format, probing key pathology concepts including tumor morphology, grade, necrosis extent, stromal composition, and inflammatory infiltration. For example: *"What is the level of cellular atypia in this slide? A. Mild cellular atypia. B. Moderate cellular atypia. C. Severe cellular atypia. D. Cellular atypia not assessable in this slide."* These questions are generated using GPT-5.2 to capture clinically relevant prognostic factors in pathology. The closed-form design constrains the VQA model to select from predefined options rather than generating free-form text, yielding a bounded, hallucination-resistant semantic space with minimal linguistic variation. The complete set of 12 template questions, including all answer options and additional details, is provided in the Supplementary Material.

Semantic Anchor Extraction via VQA. Given the predefined question set, we query a pathology-specialized VQA model for WSI_i to obtain the corresponding answers. The answers to all Q questions are aggregated to form a structured slide-level semantic description. We denote the resulting semantic anchors for the i-th slide as $\mathcal{T}_i = [\mathcal{T}_i^1, \ldots, \mathcal{T}_i^Q]$.

3.2 Dual-Stream WSI Evidence Extraction

To extract both prognostically relevant and complementary global survival information, we construct a text-guided branch that aggregates patch features under

semantic guidance and a WSI-only branch that preserves holistic visual context, providing complementary global information.

Given the WSIs of a patient, we represent the processed patch-level features as a bag $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$, where $\mathbf{x}_n \in \mathbb{R}^D$ denotes the feature vector of the n-th patch. These features are extracted using the image encoder of a vision-language model (VLM). In parallel, the semantic anchors for the i-th slide, $\mathcal{T}_i$, are encoded by the VLM's text encoder to obtain the semantic representations $\mathbf{P}_i = [\mathbf{p}_i^1, \ldots, \mathbf{p}_i^Q]^\top \in \mathbb{R}^{Q \times D}$, where $\mathbf{p}_i^q \in \mathbb{R}^D$ denotes the feature vector of the q-th semantic anchor. Based on this instance representation, we derive two complementary branches: **(a) Text-guided WSI Evidence branch:** selectively aggregates patch features under the guidance of semantic anchors, encouraging the model to focus on prognostically relevant regions and clinically meaningful patterns beyond raw visual statistics. **(b) WSI-only branch:** encodes holistic visual context without semantic conditioning, preserving global and information-rich representations across the entire slide.

Text-Guided WSI Evidence Extraction. The embedding of each semantic anchor $\mathbf{p}^q$ serves as a query to retrieve prognostically relevant patches from the patch feature bag $\mathbf{X}$ via similarity matching and weighted averaging:

$$\mathbf{f}_{tg}^q = \sum_{i=1}^{N} \mathbf{x}_i \cdot \frac{\exp\big(\alpha \cdot \mathrm{sim}(\mathbf{p}^q, \mathbf{x}_i)\big)}{\sum_{n=1}^{N} \exp\big(\alpha \cdot \mathrm{sim}(\mathbf{p}^q, \mathbf{x}_n)\big)}, \tag{1}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity and $\alpha > 0$ is a fixed temperature hyperparameter. Applying Eq. (1) to all Q semantic anchors yields a set of text-guided features $\{\mathbf{f}_{tg}^q\}_{q=1}^Q$. We aggregate the text-guided features via mean pooling and map the resulting representation to evidential outputs through a linear projection followed by a Softplus activation:

$$\mathbf{e}_{tg} = \mathrm{Softplus}\left(\mathbf{W}_{tg}\left(\frac{1}{Q}\sum_{q=1}^{Q} \mathbf{f}_{tg}^q\right) + \mathbf{b}_{tg}\right) \in \mathbb{R}_{\geq 0}^K, \tag{2}$$

where $\mathbf{W}^{tg} \in \mathbb{R}^{K \times D}$ and $\mathbf{b}^{tg} \in \mathbb{R}^K$ are learnable parameters, and $\mathbf{e}^{tg}$ denotes a non-negative vector corresponding to K discrete survival intervals. Following the evidential learning framework, these non-negative outputs are interpreted as evidence that parameterizes the Dirichlet concentration parameters.

WSI-only Evidence Extraction. To obtain a global representation of a WSI, we apply an attention-based pooling mechanism over patch features using

$$a_i^{\mathrm{wo}} = \frac{\exp\big((\mathbf{w}^{\mathrm{wo}})^\top \tanh(\mathbf{V}^{\mathrm{wo}}\mathbf{x}_i)\big)}{\sum_{j=1}^{N} \exp\big((\mathbf{w}^{\mathrm{wo}})^\top \tanh(\mathbf{V}^{\mathrm{wo}}\mathbf{x}_j)\big)}, \qquad \mathbf{f}^{\mathrm{wo}} = \sum_{i=1}^{N} a_i^{\mathrm{wo}}\, \mathbf{x}_i, \tag{3}$$

where $\mathbf{V}^{\mathrm{wo}} \in \mathbb{R}^{D \times D}$ and $\mathbf{w}^{\mathrm{wo}} \in \mathbb{R}^D$ are learnable parameters. The aggregated representation $\mathbf{f}^{\mathrm{wo}}$ is then mapped to evidential outputs via a linear projection followed by a Softplus activation:

$$\mathbf{e}^{\mathrm{wo}} = \mathrm{Softplus}(\mathbf{W}^{\mathrm{wo}}\mathbf{f}^{\mathrm{wo}} + \mathbf{b}^{\mathrm{wo}}) \in \mathbb{R}_{\geq 0}^K, \tag{4}$$

where $\mathbf{W}^{\text{wo}} \in \mathbb{R}^{K\times D}$ and $\mathbf{b}^{\text{wo}} \in \mathbb{R}^{K}$ are learnable parameters.

Adaptive Evidence Mixing: We combine the evidential outputs of the two branches using a mixing coefficient $\lambda \in [0, 1]$:

$$\mathbf{e}_{\text{path}} = (1-\lambda)\,\mathbf{e}^{\text{wo}} + \lambda\,\mathbf{e}^{\text{tg}}, \tag{5}$$

where λ is a hyperparameter that controls the relative contribution of the WSI-only and text-guided branches, $\mathbf{e}_{\text{path}} \in \mathbb{R}^{K}_{\geq 0}$ denotes the mixed visual evidence.

3.3 Text Evidence Extraction

Beyond guiding visual feature aggregation, the semantic representations themselves can serve as an additional modality for survival prediction. We thus map the semantic embeddings directly into evidential predictions to capture domain-stable prognostic signals.

To complement the visual evidence, we derive a purely semantic evidence stream from the anchor embeddings. The Q semantic anchor embeddings $\{\mathbf{p}_q\}_{q=1}^{Q}$ are aggregated via mean pooling and mapped to a non-negative evidence vector using Softplus activation:

$$\mathbf{e}^{\text{text}} = \text{Softplus}\left(\mathbf{W}^{\text{text}}\left(\frac{1}{Q}\sum_{q=1}^{Q}\mathbf{p}_q\right) + \mathbf{b}^{\text{text}}\right) \in \mathbb{R}^{K}_{\geq 0}, \tag{6}$$

where $\mathbf{W}^{\text{text}} \in \mathbb{R}^{K\times D}$ and $\mathbf{b}^{\text{text}} \in \mathbb{R}^{K}$ are learnable parameters.

3.4 Cautious Belief Fusion

The textual and visual evidence streams are inherently dependent and combining them directly naively may amplify shared signals and lead to overconfident predictions. We therefore introduce cautious belief fusion, which preserves uncertainty and ensures more reliable predictions.

Mapping evidence into belief. We now have two evidence sources:$\mathbf{e}_{\text{path}}$(Eq. (5) from WSI source and $\mathbf{e}^{\text{text}}$ (Eq. (6)) from text source. Each evidence vector $\mathbf{e}$ defines a Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha} = \mathbf{e} + \mathbf{1} \in \mathbb{R}^{K}_{\geq 0}$. Following Subjective Logic [23], we convert the Dirichlet parameters into a *belief–uncertainty pair* $(\mathbf{b}, u)$:

$$b_k = \frac{e_k}{S}, \qquad u = \frac{K}{S}, \qquad S = \sum_{k=1}^{K}\alpha_k, \tag{7}$$

where S is the Dirichlet strength and K is the number of discrete survival intervals. The belief mass b_k represents the evidence-supported probability for the k-th survival interval, while u captures epistemic uncertainty due to lack of evidence. By construction, $\sum_k b_k + u = 1$. Applying Eq. (7) to $\mathbf{e}_{\text{path}}$ and $\mathbf{e}^{\text{text}}$ yields two opinions: $(\mathbf{b}_{\text{path}}, u_{\text{path}})$ and $(\mathbf{b}_{\text{text}}, u_{\text{text}})$.

Cautious conjunction rule. We fuse the two opinions using the *cautious conjunction rule* [7], which operates through the *commonality function* $q_k = b_k + u$:

$$q_k^{\text{fused}} = \min\big(q_k^{\text{path}},\ q_k^{\text{text}}\big), \qquad u^{\text{fused}} = \min\big(u_{\text{path}},\ u_{\text{text}}\big). \tag{8}$$

The fused belief masses are recovered and normalized:

$$\tilde{b}_k^{\text{fused}} = \max\big(q_k^{\text{fused}} - u^{\text{fused}},\ 0\big), \quad b_k^{\text{fused}} = \frac{\tilde{b}_k^{\text{fused}}}{Z}, \quad u^{\text{fused}} \leftarrow \frac{u^{\text{fused}}}{Z}, \tag{9}$$

where $Z = \sum_j \tilde{b}_j^{\text{fused}} + u^{\text{fused}}$. The key property of cautious fusion is uncertainty preservation: if either modality assigns high uncertainty, the fused opinion retains that uncertainty rather than allowing the other modality's potentially unreliable confidence to dominate.

3.5 Training Objective

Let $y_i \in \{0, \ldots, K-1\}$ denote the discretized survival interval and $c_i \in \{0, 1\}$ the censorship indicator ($c_i = 1$ if censored) for sample i. From the fused belief–uncertainty pair $(\mathbf{b}^{\text{fused}}, u^{\text{fused}})$ obtained in Eq. (9), we recover the fused Dirichlet parameters:

$$S^{\text{fused}} = \frac{K}{u^{\text{fused}} + \epsilon}, \qquad \alpha_k^{\text{fused}} = b_k^{\text{fused}} \cdot S^{\text{fused}} + 1, \tag{10}$$

where ϵ is a small constant for numerical stability. The expected categorical probability for the k-th interval is $p_k = \alpha_k^{\text{fused}} / S^{\text{fused}}$.

Survival loss. For uncensored samples ($c_i = 0$), we maximize the hazard probability at the event interval; for censored samples ($c_i = 1$), we maximize the survival probability beyond the last observed interval:

$$\mathcal{L}_{\text{surv}} = -\sum_{i=1}^{N} \Big[(1 - c_i) \log p_{y_i}^{(i)} + c_i \log \sum_{k > y_i} p_k^{(i)} \Big]. \tag{11}$$

The first term encourages the model to assign a high probability to the interval where the event occurs. The second term ensures that for censored patients, who are known to survive beyond interval y_i, the model assigns a high cumulative probability to the subsequent intervals.

KL regularizer. To prevent the model from accumulating spurious evidence for incorrect intervals, we add a KL-divergence regularizer that penalizes deviation from a non-informative uniform Dirichlet [43]:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{N} \text{KL}\big[\text{Dir}(\tilde{\boldsymbol{\alpha}}^{(i)}) \parallel \text{Dir}(\mathbf{1})\big], \tag{12}$$

where $\tilde{\alpha}_k^{(i)} = \alpha_k^{(i)} - e_k^{(i)} \cdot \mathbf{1}[k = y_i] + 1$ removes the evidence for the ground-truth interval before penalization, so that only the evidence assigned to *incorrect* intervals is regularized toward zero. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{surv}} + \lambda_t \cdot \mathcal{L}_{\text{KL}}, \tag{13}$$

Table 1: Cross-center survival prediction results (TCGA → CPTAC/NLST) under unimodal settings. C: C-index↑, S: IBS↓, I: INBLL↓. Results are reported as mean ± std over three runs. Best results are in **bold** and our results are color-coded.

Method	CPTAC-LUAD			CPTAC-UCEC			CPTAC-KIRC			NLST-LUAD			Avg		
	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$
ABMIL	0.464±0.01	0.702±0.01	1.962±0.06	0.534±0.01	0.685±0.00	2.984±0.03	0.600±0.01	0.814±0.01	3.412±0.20	0.535±0.01	0.706±0.00	2.112±0.01	0.533	0.727	2.617
TransMIL	0.500±0.02	0.946±0.00	5.081±0.05	0.636±0.04	0.833±0.03	5.307±0.23	0.622±0.01	0.868±0.00	5.431±0.15	0.516±0.03	0.908±0.00	4.728±0.02	0.568	0.889	5.136
DSMIL	0.469±0.00	0.613±0.00	1.575±0.02	0.607±0.02	0.662±0.02	2.199±0.11	0.637±0.00	0.772±0.00	2.596±0.07	0.534±0.02	0.525±0.02	1.308±0.04	0.562	0.643	1.920
ILRA	0.519±0.05	0.974±0.03	6.553±0.13	0.638±0.06	0.877±0.00	6.899±0.45	0.616±0.01	0.904±0.00	7.071±0.01	0.596±0.02	0.865±0.02	5.745±0.07	0.593	0.905	6.567
BayesMIL	0.402±0.00	0.619±0.02	1.578±0.05	**0.702±0.02**	0.658±0.00	2.322±0.00	0.564±0.00	0.787±0.00	2.816±0.05	0.503±0.00	0.581±0.01	1.484±0.02	0.543	0.661	2.050
UMSA	0.435±0.03	0.709±0.02	1.979±0.09	0.540±0.03	0.700±0.05	3.005±0.20	0.605±0.01	0.816±0.01	3.390±0.13	0.530±0.02	0.688±0.03	2.020±0.15	0.528	0.728	2.599
ACMIL	0.405±0.10	0.679±0.00	1.895±0.02	0.644±0.00	0.721±0.00	2.723±0.06	0.642±0.00	0.790±0.00	2.791±0.01	0.559±0.00	0.793±0.02	2.577±0.10	0.563	0.746	2.497
OTSurv	0.529±0.06	0.570±0.12	1.432±0.32	0.612±0.00	0.702±0.00	2.035±0.04	0.652±0.01	0.724±0.03	2.110±0.07	0.644±0.01	0.523±0.00	1.307±0.01	0.609	0.630	1.721
Ours	**0.663±0.02**	**0.376±0.01**	**0.910±0.03**	0.682±0.02	**0.270±0.00**	**0.703±0.00**	**0.677±0.03**	**0.272±0.01**	**0.700±0.02**	**0.662±0.02**	**0.278±0.01**	**0.709±0.02**	**0.671**	**0.299**	**0.755**

where $\lambda_t = \min(1,\ t/T_{\text{anneal}})$ is an annealing coefficient that linearly increases from 0 to 1 over the first T_{anneal} training steps, allowing the model to freely accumulate evidence in early training before the regularizer takes full effect.

4 Experiments

4.1 Experimental Setup

Datasets. We use The Cancer Genome Atlas (TCGA) database [50] as the source domain for training and test the trained model on four external cohorts from independent centers to evaluate cross-domain generalization under the zero-shot setting. TCGA covers three cancer types: Lung Adenocarcinoma (LUAD), Uterine Corpus Endometrial Carcinoma (UCEC), and Kidney Renal Clear Cell Carcinoma (KIRC). The four external cohorts include CPTAC [8] (CPTAC-LUAD, CPTAC-UCEC, and CPTAC-KIRC), and NLST [48] (NLST-LUAD). These cohorts exhibit substantial distribution shifts in staining protocols, scanner hardware, and tissue preparation pipelines, posing a challenging benchmark for domain robustness. No data from the target cohorts are used during training.
Evaluation Metrics. Following [15], we adopt three complementary metrics for survival prediction assessment: concordance index (C-index) measures the discriminative ability of the model by computing the fraction of concordant pairs among all comparable patient pairs; Integrated Brier Score (IBS) evaluates the calibration and sharpness of the predicted survival probabilities over time; and Integrated Negative Binomial Log-Likelihood (INBLL) assesses the overall probabilistic fit of the predicted survival distributions.
Baselines. We compare SAEFS against two groups of state-of-the-art methods. **(1) Unimodal methods:** ABMIL [19], TransMIL [45], DSMIL [26], ILRA [54], BayesMIL [5], UMSA [21], ACMIL [61], and OTSurv [41]. **(2) Multimodal methods** that fuse WSI features with VQA-derived textual representations: MCAT [2], MOTCAT [58], SurvPath [20], and PS3 [39]. All baselines are trained on the same TCGA splits with the CONCH [32] patch encoder for fair comparison and evaluated zero-shot on external cohorts. Methods that use target-domain data for training, e.g., domain adaptation, are excluded for fairness. See the Supplementary Material for implementation details.

Table 2: Cross-domain survival prediction results for multimodal methods (TCGA → CPTAC/NLST). C: C-index↑, S: IBS↓, I: INBLL↓. Results are reported as mean ± std over three runs. Best results are in **bold** and our results are color-coded.

Method	CPTAC-LUAD			CPTAC-UCEC			CPTAC-KIRC			NLST-LUAD			Avg		
	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$	$C\uparrow$	$S\downarrow$	$I\downarrow$
MCAT	0.440±0.05	0.286±0.00	0.873±0.03	0.602±0.07	**0.214**±0.01	0.824±0.05	0.591±0.03	0.217±0.01	0.793±0.03	0.460±0.05	0.266±0.04	0.822±0.11	0.523	0.246	0.828
MOTCAT	0.436±0.03	**0.274**±0.02	**0.844**±0.09	0.635±0.06	0.217±0.00	0.791±0.06	0.555±0.04	**0.211**±0.01	0.733±0.04	0.510±0.01	**0.245**±0.02	**0.708**±0.05	0.534	**0.237**	0.769
SurvPath	0.426±0.07	0.289±0.01	0.894±0.04	0.572±0.02	0.220±0.01	0.842±0.04	0.627±0.04	0.213±0.01	0.756±0.04	0.431±0.01	0.279±0.02	0.850±0.07	0.514	0.250	0.836
PS3	0.455±0.04	0.287±0.01	1.023±0.06	0.666±0.02	0.218±0.02	1.058±0.09	0.596±0.04	0.229±0.01	1.020±0.06	0.529±0.04	0.254±0.01	0.872±0.04	0.562	0.247	0.993
Ours	**0.663**±0.02	0.376±0.01	0.910±0.03	**0.682**±0.02	0.270±0.00	**0.703**±0.00	**0.677**±0.03	0.272±0.01	**0.700**±0.02	**0.662**±0.02	0.278±0.01	0.709±0.02	**0.671**	0.299	**0.755**

4.2 Comparison with State-of-the-Art Methods

Comparison with Unimodal Methods. Table 1 summarizes the zero-shot cross-center performance of our method against eight state-of-the-art WSI-only survival prediction methods. Our method achieves the highest average C-index of 0.671, exceeding the strongest baseline by +0.062. The improvement is particularly notable on the high-shift CPTAC-LUAD cohort, where SAEFS attains a C-index of 0.663 compared to 0.529 for the best competing method. **Notably, the performance gains become more pronounced as the domain shift increases, indicating that SAEFS is particularly effective under severe cross-center distribution shifts.** In terms of prediction uncertainty, our method also achieves the lowest average IBS (0.299) and INBLL (0.755), indicating more accurate and reliable survival estimation under domain shift. These results suggest that combining semantic anchoring with evidential fusion improves cross-domain robustness compared with purely visual approaches.

Comparison with Multimodal Methods. Table 2 compares SAEFS with several multimodal survival prediction methods that combine WSI features with semantic representations. SAEFS achieves the highest average C-index of 0.671, exceeding the strongest competing method by +0.109. Notably, while multimodal baselines yield slightly lower IBS scores, we attribute this to inter-modal correlation: naive concatenation-based fusion implicitly double-counts shared prognostic signals between visual and semantic modalities, producing overconfident predictions that appear sharp but are poorly calibrated. In contrast, SAEFS explicitly models per-modal reliability through cautious belief fusion, avoiding such overconfidence and achieving the best average INBLL of 0.755 alongside substantially stronger discriminative power. The calibration analysis in Sec. 4.5 corroborates this interpretation, showing that SAEFS produces well-calibrated survival estimates while baseline models exhibit systematic deviation between predicted and observed probabilities.

Model complexity. Although SAEFS adds a semantic branch, its trainable fusion module remains lightweight, containing only 270K parameters, compared with 659K for ABMIL, 2.41M for TransMIL, and 2.74M for MCAT. Semantic anchors are generated once per slide using the VQA model and cached offline, introducing no additional computational cost during fusion-model training or inference.

Table 3: Ablation study (average over four external datasets). TCGA→CPTAC/NLST zero-shot evaluation. A1 removes the text-guided visual stream (WSI-only, $\lambda = 0$). A2 removes the global WSI branch and uses only text-guided visual features (TG-WSI only, $\lambda = 1$). A3 uses only semantic features (Text-only). A4 replaces evidential fusion with simple feature concatenation. A5 replaces the cautious fusion rule with the Dempster–Shafer rule. Δ is computed w.r.t. the full model.

		C-index ↑		**IBS ↓**		**INBLL ↓**	
Action	Configuration	Value	Δ	Value	Δ	Value	Δ
A1	WSI-only ($\lambda = 0$)	0.5980	-0.0860	0.2933	-0.0098	0.7452	-0.0196
A2	TG-WSI only ($\lambda = 1$)	0.6408	-0.0432	0.2964	-0.0067	0.7490	-0.0158
A3	Text-only	0.5327	-0.1513	0.3769	+0.0738	0.9417	+0.1769
A4	Concatenation fusion	0.5310	-0.1530	0.2119	-0.0912	0.7551	-0.0098
A5	Dempster fusion	0.5758	-0.1082	0.3113	+0.0082	0.7824	+0.0176
	SAEFS (Full model)	0.6840	—	0.3031	—	0.7648	—

4.3 Ablation Study

To assess the effectiveness of each component of SAEFS, we conduct a systematic ablation study averaged over the four external datasets (Table 3). The ablation variants progressively modify the visual streams and the fusion mechanism to analyze their individual impact on cross-center survival prediction.

Effect of visual and semantic modalities. We first examine the role of the dual visual branches. Removing the text-guided visual stream (A1: WSI-only) leads to a notable C-index drop of -0.0860, indicating that semantic guidance provides important cues for identifying prognostically relevant regions under domain shift. Using only the text-guided visual stream without the global WSI branch (A2: TG-WSI only) also degrades performance (-0.0432), suggesting that global visual context remains complementary to the semantically guided features. Finally, the text-only variant (A3) yields the largest degradation (-0.1513), showing that VQA-derived semantics alone, while relatively domain-stable, are insufficient for fine-grained survival discrimination without visual grounding.

Effect of cautious belief fusion. We next analyze the fusion mechanism for visual and semantic evidence. Replacing the evidential fusion with a simple concatenation-based predictor (A4) significantly degrades performance (-0.1530), highlighting the importance of uncertainty-aware fusion when combining heterogeneous modalities under domain shift. Furthermore, substituting the cautious conjunction rule with the conventional Dempster's combination rule (A5) also reduces performance (-0.1082) and slightly worsens IBS and INBLL. This suggests that the independence assumption underlying Dempster's combination rule may be violated in our setting, where visual and semantic evidence can be strongly correlated. The cautious fusion rule, which is designed to handle dependent evidence more conservatively, therefore yields more reliable predictions. Overall, the full SAEFS model achieves the best balance between discrimination and cal-

Table 4: Control experiments on CPTAC-UCEC under zero-shot transfer (TCGA→CPTAC-UCEC). C: C-index↑; S: IBS↓; I: INBLL↓. The best result for each metric is shown in bold.

Method	C ↑	S ↓	I ↓
OTSurv (VQA concat)	0.596	0.600	1.873
SAEFS (random anchors)	0.504	0.532	1.424
SAEFS (SlideChat [4])	0.678	0.535	1.432
SAEFS (Ours)	**0.682**	**0.270**	**0.703**

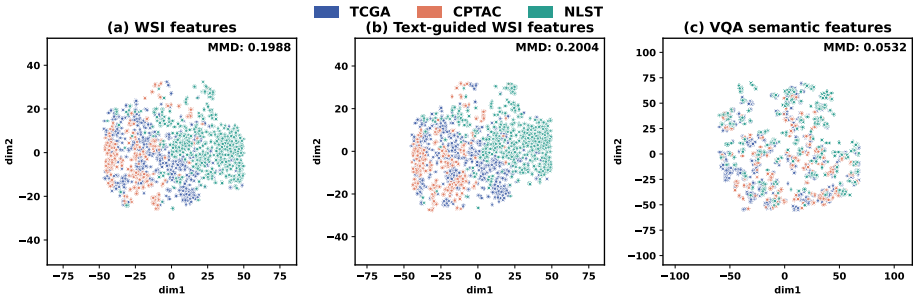


Fig. 3: Cross-center feature distributions under domain shift. t-SNE visualization of three feature representations extracted from TCGA, CPTAC, and NLST: (a) raw WSI features, (b) text-guided WSI features, and (c) VQA-derived semantic features. Colors indicate different centers. Semantic features exhibit substantially better cross-center alignment with a much lower Maximum Mean Discrepancy (MMD).

ibration, confirming the importance of both the dual-stream semantic guidance and the cautious evidential fusion design.

Attribution of performance gains. We conduct four control experiments to examine whether the improvement arises from the proposed fusion strategy or simply from using a strong VQA model (Table 4). Directly concatenating the same VQA features with OTSurv does not improve over unimodal OTSurv, suggesting that the additional VQA information alone is insufficient. Replacing the semantic anchors with random text leads to a substantial performance drop, highlighting the importance of meaningful semantic guidance. Moreover, replacing ALPaCA with SlideChat yields comparable C-index performance, showing that SAEFS is not restricted to a particular VQA backbone. Combined with the text-only ablation (A3), these results support that the cross-center improvement mainly comes from the uncertainty-aware dual-stream fusion rather than the choice of the VQA model.

Sensitivity to the number of questions. We further vary the number of template questions and observe a consistent improvement as coverage increases (Q=4: 0.633, Q=8: 0.660, Q=12: 0.682), indicating that a broader set of prognostic factors provides more complete semantic coverage.

4.4 Analysis of Domain Robustness

Cross-Center Feature Divergence. To empirically validate the effectiveness of semantic information for cross-center generalization, we quantify the distributional divergence between the source domain (TCGA) and external cohorts (CPTAC and NLST) using Maximum Mean Discrepancy (MMD) [11]. We compare three feature representations: (a) raw WSI features extracted from the patch encoder, (b) text-guided WSI features produced by the semantic cross-attention (Eq. 1), and (c) VQA-derived semantic features from the text encoder.

As illustrated in Fig. 3, the raw WSI features (Fig. 3a) exhibit clear distributional separation between TCGA and the external cohorts, with an MMD of 0.1988. The text-guided WSI features (Fig. 3b) show a comparable level of divergence (MMD = 0.2004), indicating that although semantic guidance alters the visual representation, the underlying domain shift in pixel-level features remains substantial. In contrast, the VQA semantic features (Fig. 3c) display strong overlap across centers, yielding a dramatically lower MMD of 0.0532—approximately a 73% reduction compared with the raw visual features. This result suggests that high-level pathology semantics features derived from template-structured VQA are substantially more domain-invariant than pixel-derived representations, validating the improved cross-center robustness of SAEFS.

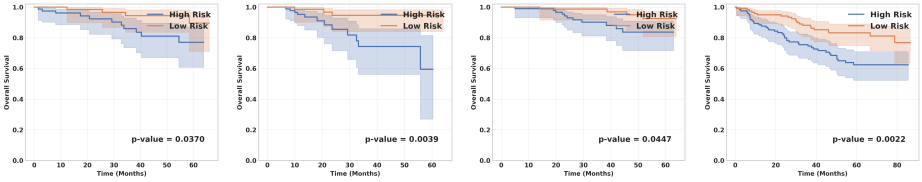
4.5 Risk Stratification and Calibration Analysis

Kaplan–Meier Risk Stratification. To assess whether SAEFS produces clinically meaningful risk predictions, we perform Kaplan-Meier survival analysis on each external cohort. Patients are stratified into high-risk and low-risk groups based on the median predicted risk score. As shown in Fig. 4, SAEFS achieves statistically significant stratification (log-rank test) between risk groups across all four external cohorts: CPTAC-KIRC ($p = 0.0370$), CPTAC-LUAD ($p = 0.0039$), CPTAC-UCEC ($p = 0.0047$), and NLST-LUAD ($p = 0.0022$). These results indicate that SAEFS produces consistent and clinically meaningful risk stratification even under substantial cross-center distribution shift.

Calibration Analysis. Beyond discrimination, reliable clinical deployment requires well-calibrated survival probabilities. Fig. 5 shows calibration curves comparing SAEFS with representative baselines across the four external cohorts. SAEFS consistently produces curves closer to the ideal diagonal, indicating improved calibration prediction under domain shift. Notably, although several multimodal baselines achieve slightly lower IBS and INBLL scores, their calibration curves deviate more noticeably from the diagonal line, suggesting less reliable probability estimates. The above observations indicate that SAEFS provides a better balance between predictive accuracy and probability calibration, attributable to the evidential modeling and cautious belief fusion mechanism.

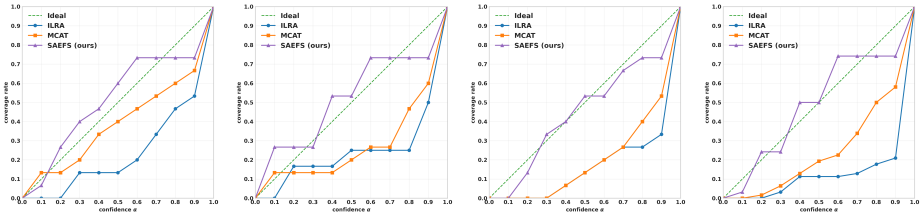
5 Conclusion

In this work, we address the domain shift challenge in WSI survival analysis. We show that high-level pathological semantics are more domain-invariant than



(a) CPTAC-KIRC **(b)** CPTAC-LUAD **(c)** CPTAC-UCEC **(d)** NLST-LUAD

Fig. 4: Kaplan–Meier risk stratification curves on external datasets (TCGA → CPTAC/NLST). High- and low-risk groups are separated by the median predicted risk.



(a) CPTAC-KIRC **(b)** CPTAC-LUAD **(c)** CPTAC-UCEC **(d)** NLST-LUAD

Fig. 5: Calibration plots comparing predicted and observed survival probabilities on external cohorts (TCGA → CPTAC/NLST). Curves closer to the diagonal indicate better calibration of survival risk predictions.

features captured by traditional MIL encoders. To operationalize this insight, we propose **SAEFS**, a survival framework that extracts semantics via template-structured VQA to anchor cross-center survival prediction in a semantic space, while employing cautious belief fusion to combine both visual and textual evidence. Zero-shot evaluations across four unseen multi-center cohorts show that SAEFS significantly outperforms state-of-the-art methods in prognostic accuracy and uncertainty calibration, yielding a more reliable and generalizable survival analysis framework. Future work will explore open-ended VLMs to derive finer semantic descriptions of rare morphological subtypes and investigate uncertainty-aware VLMs that provide semantic-level confidence for downstream diagnosis. By shifting from pixel-centric to clinical semantic-centric learning, our work advances a scalable paradigm for robust cross-center diagnostic systems.

Acknowledgement

This research was supported by the National Medical Research Council (NMRC) Healthy and Meaningful Longevity–Cognition Grant (NICCOG2024-0028), the AI for Public Health Program at the Saw Swee Hock School of Public Health, National University of Singapore, and the Dame Julia Higgins Postdoc Collaborative Research Fund. This work was also supported by the NUS Guangzhou Research Translation and Innovation Institute (GRTII)-Affiliated PhD Scholarship Program.

References

1. 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 **30**(3), 850–862 (2024)
2. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
3. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer cell **40**(8), 865–878 (2022)
4. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Hu, M., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In: CVPR. pp. 5134–5143 (June 2025)
5. Cui, Y., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.W., Xue, C.J., Chan, A.B.: Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: 11th International Conference on Learning Representations (ICLR 2023) (2023)
6. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. In: Classic works of the Dempster-Shafer theory of belief functions, pp. 57–72. Springer (2008)
7. Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artificial intelligence **172**(2-3), 234–264 (2008)
8. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The cptac data portal: a resource for cancer proteomics research. Journal of proteome research **14**(6), 2707–2713 (2015)
9. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Camara, A., Mac Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. MedRxiv pp. 2023–07 (2023)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
11. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The journal of machine learning research **13**(1), 723–773 (2012)
12. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE transactions on pattern analysis and machine intelligence **45**(2), 2551–2566 (2022)
13. Huang, L., Ruan, S., Xing, Y., Feng, M.: A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. Medical Image Analysis **97**, 103223 (2024)
14. Huang, L., Xing, Y., Denoeux, T., Feng, M.: An evidential time-to-event prediction model based on gaussian random fuzzy numbers. In: International Conference on Belief Functions. pp. 49–57. Springer (2024)
15. Huang, L., Xing, Y., Lin, Q., Duan, J., Ruan, S., Feng, M.: Esurvfusion: An evidential multimodal survival fusion model based on epistemic random fuzzy sets. IEEE Transactions on Fuzzy Systems (2025)
16. Huang, L., Xing, Y., Mishra, S., Denœux, T., Feng, M.: Evidential time-to-event prediction with calibrated uncertainty quantification. International Journal of Approximate Reasoning **181**, 109403 (2025)

17. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
18. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36**, 37995–38017 (2023)
19. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
20. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: cvpr. pp. 11579–11590 (2024)
21. Jiang, S., Cai, L., Gan, Z., Wang, Y., Tang, G., Zhang, Y.: Uncertainty-aware survival analysis with dirichlet distribution for multi-scale pathology and genomics. IEEE Transactions on Medical Imaging (2025)
22. de Jong, E.D., Marcus, E., Teuwen, J.: Current pathology foundation models are unrobust to medical center differences. arXiv preprint arXiv:2501.18055 (2025)
23. Jsang, A.: Subjective Logic: A formalism for reasoning under uncertainty. Springer Publishing Company, Incorporated (2018)
24. Kumar, N., Gupta, R., Gupta, S.: Whole slide imaging (wsi) in pathology: current perspectives and future directions. Journal of digital imaging **33**(4), 1034–1040 (2020)
25. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
26. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR. pp. 14318–14328 (2021)
27. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
28. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
29. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
30. Liu, J., Li, H., Yang, C., Deutges, M., Sadafi, A., You, X., Breininger, K., Navab, N., Schüffler, P.J.: Hasd: hierarchical adaption for pathology slide-level domain-shift. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 332–342. Springer (2025)
31. Liu, P., Zeng, X., Ma, T., Xing, Y., Ren, X., Liu, Y.: Sparse task vector mixup with hypernetworks for efficient knowledge transfer in whole-slide image prognosis. In: CVPR. pp. 35238–35247 (2026)
32. 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)
33. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al.: A multimodal generative ai copilot for human pathology. Nature **634**(8033), 466–473 (2024)

34. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: CVPR. pp. 19764–19775 (2023)
35. 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 **5**(6), 555–570 (2021)
36. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. pp. 1107–1110. IEEE (2009)
37. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J., Cooper, L.A.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences **115**(13), E2970–E2979 (2018)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
39. Raza, M., Azam, A., Qaiser, T., Rajpoot, N.: 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. pp. 22175–22186 (2025)
40. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2002)
41. Ren, Q., Wang, Y., Fang, R., Ling, H., You, C.: Otsurv: A novel multiple instance learning framework for survival prediction with heterogeneity-aware optimal transport. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 439–449. Springer (2025)
42. Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al.: Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell reports **23**(1), 181–193 (2018)
43. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
44. Shafer, G.: A mathematical theory of evidence. Princeton university press (2020)
45. 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. Advances in neural information processing systems **34**, 2136–2147 (2021)
46. Shou, Y., Cao, X., Yan, P., Hui, Q., Zhao, Q., Meng, D.: Graph domain adaptation with dual-branch encoder and two-level alignment for whole slide image-based survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19925–19935 (2025)
47. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
48. Team, N.L.S.T.R.: Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine **365**(5), 395–409 (2011)
49. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., Van Der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical image analysis **58**, 101544 (2019)

50. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
51. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging **35**(8), 1962–1971 (2016)
52. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
53. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
54. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)
55. Xing, Y., Huang, L., Ma, J., Hong, R., Qiu, J., Liu, P., He, K., Fu, H., Feng, M.: Dpsurv: Dual-prototype evidential fusion for uncertainty-aware and interpretable whole-slide image survival prediction. arXiv preprint arXiv:2510.00053 (2025)
56. Xing, Y., Liu, P., Ma, J., Hong, R., Qiu, J., Liu, T., He, K., Huang, L., Feng, M.: Bridging the modality bottleneck in pathology mil through virtual molecular staining. arXiv preprint arXiv:2605.16392 (2026)
57. Xing, Y., Mo, H., Wang, Z., Huang, L., Feng, M.: Evidential fusion network for multimodal survival prediction under missing modalities. arXiv preprint arXiv:2606.20757 (2026)
58. Xu, Y., Chen, H.: 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. pp. 21241–21251 (2023)
59. Yang, S., Wang, Y., Van De Weijer, J., Herranz, L., Jui, S.: Generalized source-free domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8978–8987 (2021)
60. 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. pp. 18802–18812 (2022)
61. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European conference on computer vision. pp. 125–143. Springer (2024)
62. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: CVPR. pp. 7234–7242 (2017)