DSIB: Towards Robust Multimodal Survival Prediction via Discrete Self-Distillation Information Bottleneck

1st Ye Tian
College of Computer Science and Technology (College of Data Science)
Taiyuan University of Technology
Taiyuan, China
1036851584@qq.com

2nd Ruiqin Bai*
College of Artificial Intelligence
Taiyuan University of Technology
Taiyuan, China
bairuiqinty@163.com

***Abstract*—Multimodal survival prediction can improve prognosis by jointly modeling whole slide images (WSIs) and genomic profiles, yet practical models are brittle under multimodal heterogeneity. In particular, when training is constrained to extremely small batches (often required by WSI processing), modality-specific noise and distribution mismatch are amplified in the fused interaction space, making optimization unstable and prone to redundant correlations. We propose *Discrete Self-Distillation Information Bottleneck* (DSIB), a training-only algorithm that directly regularizes the fusion representation without modifying backbone architectures. DSIB inserts a stochastic *binary* bottleneck to explicitly control the information capacity of fused features, and injects hidden-space perturbations to expose heterogeneity-induced failure modes. Crucially, DSIB performs distributional self-distillation by minimizing the Jensen–Shannon divergence between the factorized Bernoulli posteriors of perturbed and unperturbed pathways, stabilizing learning under small-batch noise. We provide a theoretical connection showing that DSIB optimizes a tractable bound of the discrete information bottleneck objective. Experiments on three TCGA cohorts (BLCA, UCEC, LUAD) across multiple state-of-the-art multimodal survival predictors demonstrate consistent C-index gains and improved robustness under controlled noise and redundancy, with zero inference overhead.**

***Index Terms*—multimodal learning, survival prediction, information bottleneck, self-distillation, robustness**

I. Introduction

Cancer survival prediction plays a pivotal role in precision oncology by enabling personalized treatment planning and risk stratification [1]. The rapid digitization of clinical diagnostics has led to the widespread availability of heterogeneous healthcare multimedia, particularly Whole Slide Images (WSIs) that capture spatial tissue morphology and genomic profiles that reflect high-dimensional molecular alterations [2]. Learning from such complementary modalities has the potential to improve prognostic modeling compared with unimodal approaches [3], [4].

However, multimodal survival prediction models (MSPMs) remain challenging due to multimodal heterogeneity, i.e., modality-dependent noise, missingness, and distribution mismatch [5]. WSIs are prone to stain variations, background regions, and scanning artifacts, whereas genomic measurements often contain missing values and sequencing noise [6], [7]. Direct fusion can propagate these imperfections into the latent interaction space, causing one modality to dominate and suppress informative cues from the other [8], [9].

Moreover, the high computational and memory cost of WSI feature extraction, together with limited cohort size, often forces MSPMs to operate in a small-batch regime (frequently batch size 1, sometimes with gradient accumulation). In this setting, unreliable batch statistics and high-dimensional continuous fusion representations can become overly sensitive to noise and can overfit modality-specific artifacts, leading to unstable optimization and poor robustness. We refer to this phenomenon as heterogeneity-induced noise amplification, which severely undermines robust healthcare multimedia analysis.

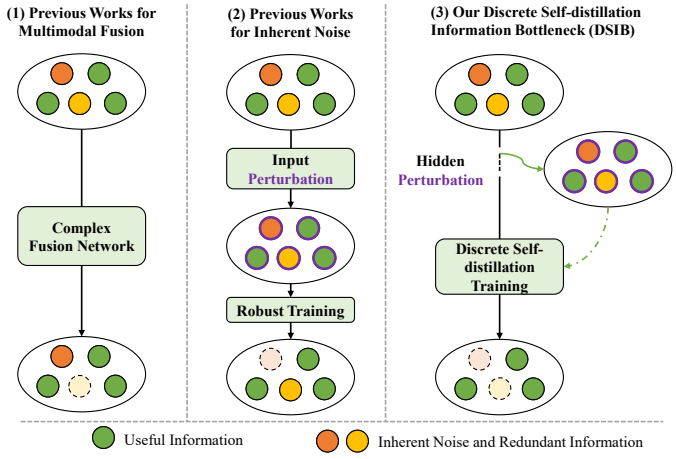

Fig. 1. Motivation and positioning of DSIB. Existing MSPMs often rely on (i) increasingly complex fusion architectures, and (ii) input-space perturbations to improve robustness. DSIB instead perturbs the fused hidden representation during training and enforces posterior-level consistency via discrete self-distillation, while keeping inference unchanged.

*Ruiqin Bai is the corresponding author. This work was supported by the Shanxi Provincial Young Scientists Research Project under Grant No. 202303021212081.

Existing robustness strategies for MSPMs commonly focus on either enhancing fusion architectures (e.g., cross-modal attention) [10] or applying input-space perturbations [8]. While effective in some regimes, these approaches may not explicitly regularize the fused latent interaction space where cross-modal noise and redundancy interact, and they can increase training complexity or inference overhead in practice [1]. The added architectural complexity and the reliance on large-batch training usually incur substantial computational cost during training and often introduce extra overhead at inference time [11]–[13], which hinders deployment in real-world clinical settings that demand lightweight and stable systems [1].

To address these limitations, we propose **Discrete Self-Distillation Information Bottleneck (DSIB)**, a plug-and-play framework for robust multimodal survival prediction, as illustrated in Fig. 1. DSIB perturbs the *fused hidden representation* during training to explicitly regularize the latent interaction space, thereby mitigating heterogeneity-induced noise amplification. Unlike prior self-distillation schemes that match continuous features, DSIB performs *distributional self-distillation* by enforcing Jensen–Shannon (JS) consistency between the *discrete bottleneck posteriors* of perturbed and unperturbed pathways. In parallel, DSIB introduces a **discrete information bottleneck** that compresses the fusion representation into compact binary codes, suppressing redundant information propagation and improving stability under small-batch training. Importantly, DSIB is *training-only*: the auxiliary pathway and consistency objective are removed after training, resulting in **zero inference overhead**. DSIB can be seamlessly integrated into diverse MSPM backbones without structural modification. We conduct extensive experiments on three representative TCGA cancer cohorts (BLCA, UCEC, and LUAD), demonstrating that DSIB consistently improves both robustness and prognostic accuracy over state-of-the-art multimodal survival prediction models.

II. Preliminary

Multimodal Survival Prediction Model (MSPM). We study multimodal survival prediction with two typical inputs, i.e., Whole Slide Images (WSIs) and genomic data. Given N patients, the dataset is $\mathcal{D} = \{(P, G)\}_{i=1}^{N}$, where $P \in \mathbb{R}^{n_p \times d}$ denotes WSI patch features and $G \in \mathbb{R}^{n_g \times d}$ denotes genomic features. Following prior practice, genomic features are organized into six genomic sequences: 1) Tumor Suppression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, 5) Transcription, and 6) Cytokines and Growth.

The model estimates a discrete-time hazard function $h(t) = h(T = t|T > t, (P, G)) \in [0, 1]$, and the survival probability is computed as: $s(t|(P, G)) = \prod_{u=1}^{t}(1 - h(u))$.

The MSPM proceeds in a fusion–embedding pipeline:

$$H = f_{\theta_1}[\text{Fusion}(P, G)], \quad Z = f_{\theta_2}(H). \tag{1}$$

The learning objective is:

$$\mathcal{L}_{\textbf{Msp}} \equiv \mathcal{L}_{\textbf{Msp}}(f_{\theta_3}(Z), S). \tag{2}$$

The Bounds in the VIB. The variational information bottleneck (VIB) [14] constrains the hidden representation H via a neural model parameterized by θ, with the goal of improving robustness. Concretely, VIB maximizes the following trade-off:

$$I(H, Y; \theta) - \beta I(H, X; \theta), \tag{3}$$

where β is a fixed coefficient and $I(\cdot, \cdot)$ denotes mutual information. The term $I(H, Y; \theta)$ encourages H to preserve predictive information about Y, whereas $I(H, X; \theta)$ penalizes excessive dependence of H on the input X. By retaining task-relevant factors while discarding nuisance variations, H becomes less sensitive to noise; meanwhile, stronger compression also reduces redundant information inherited from X.

A standard lower bound for $I(H, Y)$ in Eq. (3) is

$$I(H, Y) \geq \int dX dY dH p(X)p(Y|X)p(H|X) \log p(Y|H), \tag{4}$$

and an upper bound for $I(H, X)$ is given by

$$I(H, X) \leq \int dH dX p(X)p(H|X) \log \frac{p(H|X)}{r(H)}, \tag{5}$$

where $r(H)$ denotes a prescribed prior distribution.

III. Method

A. Discrete Self-Distillation Information Bottleneck (DSIB)

In multimodal survival prediction, training is often constrained to *small batches*. WSI feature extraction is computationally expensive, and cohort sizes are limited. In this regime, common robustness strategies can be unreliable and may add extra complexity. DSIB operates on the *fusion representation* instead. It perturbs the fused hidden state and enforces *distributional consistency* through self-distillation. DSIB also introduces a **discrete** information bottleneck to suppress redundant information, which helps when batch statistics are noisy.

Discrete bottleneck on fused representations. Let H denote the fused hidden representation produced by the MSPM fusion module. DSIB inserts a stochastic **binary** bottleneck $B \in \{0, 1\}^K$ between the fusion representation and the survival head. We parameterize an independent Bernoulli posterior using a bottleneck network g_{θ_2}:

$$p = p_{\theta_2}(B{=}1|H) = \sigma(g_{\theta_2}(H)) \in (0, 1)^K, \tag{6}$$

and obtain B using a straight-through estimator or a binary-concrete relaxation during training. During inference, we only use the original pathway and thus DSIB introduces no extra cost.

Hidden-space perturbation as robustness stress test. To explicitly train invariance at the fusion level, DSIB perturbs the fused representation:

$$H_\delta = H + \delta H, \tag{7}$$

where δH can be instantiated as either (i) fixed Gaussian noise (DSIB_α) or (ii) adaptive noise (DSIB_β):

$$\delta H \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \quad (\text{DSIB}_\alpha), \tag{8}$$

$$\delta H = \phi \cdot \epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (\text{DSIB}_\beta), \tag{9}$$

where $\phi \in \mathbb{R}^d$ is a learnable scale predicted from H (e.g., $\phi = \text{Net}_\phi(H)$). The perturbed posterior is then

$$p_\delta = p_{\theta_2}(B{=}1|H_\delta) = \sigma(g_{\theta_2}(H_\delta)). \tag{10}$$

Posterior self-distillation. Rather than matching logits or predictions, DSIB distills *bottleneck posteriors*. We enforce distributional consistency between p_δ and p using Jensen–Shannon divergence (computed between the factorized Bernoulli distributions induced by p and p_δ):

$$\mathcal{L}_{\textbf{Dist}}^{\textbf{DSIB}}(p_\delta, p) \equiv \gamma_d \cdot \textbf{JS}(p_\delta || p), \tag{11}$$

where γ_d controls the strength of invariance. This distillation pathway is only used during training.

Discrete information bottleneck regularization. To explicitly control the information capacity of the binary code, we regularize the Bernoulli posterior towards a factorized Bernoulli prior $r(B) = \prod_{k=1}^{K} \text{Bernoulli}(\pi_k)$:

$$\mathcal{L}_{\textbf{Reg}}^{\textbf{DSIB}}(p) \equiv \gamma_r \cdot \text{KL}\left(\prod_{k=1}^{K} \text{Bernoulli}(p_k) \,\middle\|\, \prod_{k=1}^{K} \text{Bernoulli}(\pi_k) \right), \tag{12}$$

where π_k can be set to 0.5 or tuned as a prior.

Overall objective. DSIB augments a standard MSPM (Eqs. (1)–(2)) with additional losses:

$$\mathcal{L}^{\textbf{DSIB}} = \underbrace{\mathcal{L}_{\textbf{Msp}}^{\textbf{DSIB}}(f_{\theta_3}(B_\delta), S)}_{\text{Survival prediction}} + \underbrace{\mathcal{L}_{\textbf{Dist}}^{\textbf{DSIB}}(p_\delta, p)}_{\text{Self-Distillation}} + \underbrace{\mathcal{L}_{\textbf{Reg}}^{\textbf{DSIB}}(p)}_{\text{Regularization Term}}, \tag{13}$$

where the prediction term is

$$\mathcal{L}_{\textbf{Msp}}^{\textbf{DSIB}} \equiv \gamma_0 \cdot \mathcal{L}(f_{\theta_3}(B_\delta), S), \tag{14}$$

$\gamma_0, \gamma_d, \gamma_r$ are weighting hyperparameters, and $\mathcal{L}(\cdot)$ is a general survival prediction loss. In summary, DSIB couples (i) fusion-level perturbation, (ii) posterior consistency self-distillation, and (iii) an explicit discrete information bottleneck into a single training recipe, improving robustness while keeping inference unchanged.

B. Theoretical Insight of DSIB

We provide theoretical evidence that DSIB optimizes a tractable discrete information bottleneck objective. Let X denote the multimodal input and Y the survival target. DSIB introduces a stochastic binary bottleneck $B \in \{0,1\}^K$ with posterior $p_\theta(B|X)$ induced by Eq. (6) via the fused representation $H = h(X)$, and a factorized Bernoulli prior $r(B) = \prod_{k=1}^{K} \text{Bernoulli}(\pi_k)$ as in Eq. (12).

Theorem 1. *Minimizing the regularization term $\mathcal{L}_{Reg}^{DSIB}$ is equivalent to minimizing the KL-divergence between the posterior $p_\theta(B|X)$ and the prior $r(B)$ (up to the scalar weight γ_r).*

Proof. Since both posterior and prior are factorized Bernoulli distributions,

$$\begin{aligned} p_\theta(B|X) &= \prod_{k=1}^{K} \text{Bernoulli}(B_k; p_k(X)), \\ r(B) &= \prod_{k=1}^{K} \text{Bernoulli}(B_k; \pi_k), \end{aligned} \tag{15}$$

their KL divergence decomposes across dimensions:

$$\begin{aligned} \text{KL}(p_\theta(B|X) \| r(B)) &= \sum_{k=1}^{K} \text{KL}\big(\text{Bern}(p_k) \,\|\, \text{Bern}(\pi_k)\big) \\ &= \sum_{k=1}^{K} \left[p_k \log \frac{p_k}{\pi_k} + (1-p_k) \log \frac{1-p_k}{1-\pi_k} \right], \end{aligned} \tag{16}$$

which matches Eq. (12) up to the multiplicative weight γ_r. □

Theorem 2. *Minimizing $\mathcal{L}^{DSIB}$ optimizes a tractable bound of the discrete information bottleneck objective:*

$$\max_\theta \ I_\theta(B;Y) - \beta I_\theta(B;X), \tag{17}$$

where $B \sim p_\theta(B|X)$ and $\beta > 0$.

Proof. The information bottleneck principle minimizes

$$\mathcal{L}_{IB}(\theta) = -I_\theta(B;Y) + \beta I_\theta(B;X). \tag{18}$$

(i) Variational upper bound for $I_\theta(B;X)$. By definition,

$$I_\theta(B;X) = \mathbb{E}_{p(x)}[\text{KL}(p_\theta(B|x) \,\|\, p_\theta(B))], \tag{19}$$

where $p_\theta(B)$ is the aggregated posterior. Introducing a variational prior $r(B)$ and using $\text{KL}(p_\theta(B) \| r(B)) \geq 0$ yields the standard bound

$$I_\theta(B;X) \leq \mathbb{E}_{p(x)}[\text{KL}(p_\theta(B|x) \,\|\, r(B))]. \tag{20}$$

By Theorem 1, the right-hand side corresponds to the DSIB regularizer $\mathcal{L}_{\textbf{Reg}}^{\textbf{DSIB}}$.

(ii) Variational lower bound for $I_\theta(B;Y)$. Similarly, introduce a variational decoder $q_\psi(y|b)$ and apply $\text{KL}(p_\theta(Y|B) \| q_\psi(Y|B)) \geq 0$, obtaining

$$I_\theta(B;Y) \geq H(Y) + \mathbb{E}_{p_\theta(b,y)}[\log q_\psi(y|b)]. \tag{21}$$

Dropping the constant $H(Y)$, maximizing $I_\theta(B;Y)$ is lower-bounded by maximizing $\mathbb{E}[\log q_\psi(y|b)]$, which corresponds to minimizing a negative log-likelihood surrogate. In DSIB, the survival head f_{θ_3} together with the survival loss $\mathcal{L}(\cdot)$ implements this term.

(iii) Connection to DSIB. Combining Eqs. (20)–(21) gives a tractable surrogate of $\mathcal{L}_{IB}$:

$$\mathcal{L}_{IB}(\theta) \lesssim -\mathbb{E}[\log q_\psi(Y|B)] + \beta\, \mathbb{E}_{p(x)}[\text{KL}(p_\theta(B|x) \,\|\, r(B))], \tag{22}$$

which aligns with $\mathcal{L}_{\textbf{Msp}}^{\textbf{DSIB}} + \mathcal{L}_{\textbf{Reg}}^{\textbf{DSIB}}$ up to constants and weighting. Finally, DSIB adds $\mathcal{L}_{\textbf{Dist}}^{\textbf{DSIB}}$ (Eq. (11)) to explicitly enforce posterior invariance under hidden perturbations $H \mapsto H_\delta$, strengthening robustness during training while keeping inference unchanged. □

IV. Experiments

A. Experimental Settings

To assess the effectiveness of our approach, we conducted experiments on The Cancer Genome Atlas (TCGA) repository[1]. TCGA is a publicly accessible cancer resource, from which we selected three high-volume cohorts that provide paired diagnostic whole-slide images (WSIs) and matched genomic profiles, together with annotated survival time and censoring indicators. Specifically, our evaluation includes 373 cases of Bladder Urothelial Carcinoma (BLCA), 480 cases of Uterine Corpus Endometrial Carcinoma (UCEC), and 453 cases of Lung Adenocarcinoma (LUAD). For each cohort, we performed 5-fold cross-validation and report the cross-validated concordance index (C-index) as the primary metric.

All models were optimized using Adam with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. Due to variable bag cardinalities across patients, we set the mini-batch size to 1 and used gradient accumulation for stable optimization. In particular, we accumulated gradients over 32 iterations, yielding an effective batch size of 32.

B. Overall Performance

Observation #1: DSIB consistently enhances performance across various MSPMs. We integrated DSIB into five representative multimodal survival frameworks, including the state-of-the-art MoME [21], evaluating them on three TCGA datasets (BLCA, UCEC, and LUAD). As summarized in Table I, both the constant (DSIB_α) and adaptive (DSIB_β) variants yield consistent C-index gains over their respective baselines, validating DSIB as a robust plug-and-play enhancement for the latent space. Notably, MoME combined with DSIB_β achieves the highest performance across all cohorts, peaking at a C-index of 0.704, 0.724, and 0.710, respectively. These improvements suggest that by introducing hidden-layer perturbations and self-distillation, DSIB effectively filters redundant features and suppresses inherent cross-modal noise without requiring structural modifications to the backbone models.

Observation #2: Adaptive noise modeling (DSIB_β) provides superior gains than constant noise (DSIB_α). The "Avg. improv." metrics in Table I highlight the clear advantage of the adaptive strategy; DSIB_α provides an average boost of approximately 1.19% to 1.22%, whereas DSIB_β significantly doubles this impact with average improvements ranging from 2.40% to 2.57%. This disparity demonstrates that learning the perturbation scale ϕ is more effective than using a fixed variance. By tailoring noise intensity to the heterogeneity of individual samples, DSIB_β avoids the risk of over-regularizing informative features while imposing stricter constraints on noisy representations, thereby fostering better generalization.

Observation #3: Latent-space refinement is essential for effective multimodal fusion. While multimodal methods generally surpass unimodal baselines (e.g., SNN or AttnMIL), confirming the benefits of integrating pathology and genomic data, they are still limited by cross-modal redundancy. The integration of DSIB enables these models to achieve superior performance; for instance, MoME+DSIB_β attains a mean C-index exceeding 0.712 across datasets. These results indicate that explicitly addressing noise and redundancy within the latent fusion space is critical for fully leveraging the complementary information from multi-source biomedical data in survival prediction.

TABLE I

Performance of MSPMs with and without DSIB on BLCA, UCEC, and LUAD. DSIB_α and DSIB_β denote constant and adaptive noise, respectively. "Improv." and "Avg. Improv." indicate relative and average improvements over the baseline.

Models	Modality		C-index		
	G	P	BLCA	UCEC	LUAD
SNN ([15])	✓		0.618	0.679	0.611
SNNTrans [15]	✓		0.659	0.656	0.638
AttnMIL ([16])		✓	0.599	0.658	0.620
CLAM-SB ([17])		✓	0.559	0.644	0.594
CLAM-MB ([17])		✓	0.565	0.609	0.582
TransMIL ([18])		✓	0.575	0.655	0.642
DTFD-MIL ([19])		✓	0.546	0.656	0.585
MCAT ([10])	✓	✓	0.672	0.649	0.659
MCAT + DSIB_α			0.680	0.657	0.667
Improv.			+1.19%	+1.23%	+1.21%
MCAT + DSIB_β			0.687	0.664	0.675
Improv.			+2.23%	+2.31%	+2.43%
Porpoise ([3])	✓	✓	0.636	0.692	0.647
Porpoise + DSIB_α			0.643	0.699	0.654
Improv.			+1.10%	+1.01%	+1.08%
Porpoise + DSIB_β			0.652	0.707	0.663
Improv.			+2.52%	+2.17%	+2.47%
MOTCAT ([12])	✓	✓	0.682	0.671	0.673
MOTCAT + DSIB_α			0.690	0.679	0.682
Improv.			+1.17%	+1.19%	+1.34%
MOTCAT + DSIB_β			0.698	0.687	0.691
Improv.			+2.35%	+2.38%	+2.67%
CMTA ([20])	✓	✓	0.672	0.691	0.676
CMTA + DSIB_α			0.681	0.700	0.685
Improv.			+1.34%	+1.30%	+1.33%
CMTA + DSIB_β			0.688	0.709	0.693
Improv.			+2.38%	+2.60%	+2.51%
MoME ([21])	✓	✓	0.686	0.706	0.691
MoME + DSIB_α			0.694	0.714	0.699
Improv.			+1.17%	+1.13%	+1.16%
MoME + DSIB_β			**0.704**	**0.724**	**0.710**
Improv.			+2.62%	+2.55%	+2.75%
Avg. improv. (DSIB_α)	-	-	**+1.19%**	**+1.17%**	**+1.22%**
Avg. improv. (DSIB_β)	-	-	**+2.42%**	**+2.40%**	**+2.57%**

C. Analysis

To gain a deeper understanding of DSIB's internal mechanisms and its broader applicability in survival analysis, our analysis focuses on the following four research questions (RQs):

- **RQ1:** Does DSIB effectively mitigate the inherent noise and redundancy prevalent in heterogeneous clinical data?
- **RQ2:** Is DSIB compatible with other robust training methods?

[1] https://gdc.cancer.gov

TABLE II
C-INDEX OF MCAT UNDER DIFFERENT GAUSSIAN NOISE INTENSITIES ON LUAD. "DECR." DENOTES THE RELATIVE DECREASE FROM THE ORIGINAL (ORI.) PERFORMANCE AFTER NOISE INJECTION.

Models	Intensity	Ori.	Noise	Decr.↓
MCAT [10]	0.01	0.659	0.638	3.19%
MCAT+DSIB$_\alpha$		0.667	0.652	2.25%
MCAT+DSIB$_\beta$		0.675	0.661	2.07%
MCAT	0.05	0.659	0.598	9.26%
MCAT+DSIB$_\alpha$		0.667	0.621	6.90%
MCAT+DSIB$_\beta$		0.675	0.637	5.63%

TABLE III
C-INDEX OF MCAT ON LUAD UNDER GENOMIC AND PATHOLOGICAL REDUNDANCY (RED.). "DECR." DENOTES THE RELATIVE PERFORMANCE DROP FROM THE ORIGINAL (ORI.) MODEL. RESULTS ARE REPORTED FOR G+P, P-ONLY, AND G-ONLY SETTINGS.

Models	Type	Ori.	w/ Red.	Decr.↓
MCAT [10]	G+P	0.659	0.604	8.35%
MCAT+DSIB$_\alpha$		0.666	0.649	2.55%
MCAT+DSIB$_\beta$		0.674	0.666	1.19%
MCAT	P	0.659	0.621	5.77%
MCAT+DSIB$_\alpha$		0.666	0.653	1.95%
MCAT+DSIB$_\beta$		0.674	0.670	0.59%
MCAT	G	0.659	0.636	3.49%
MCAT+DSIB$_\alpha$		0.666	0.660	0.90%
MCAT+DSIB$_\beta$		0.674	0.672	0.30%

We mainly consider three MSPMs, i.e., MCAT [10], MOTCAT [12], MoME [21] on the dataset LUAD for the following analysis.

Mitigating Inherent Noise and Redundant Information (RQ1). We investigate whether DSIB effectively mitigates heterogeneity-induced noise and filters redundancy through controlled stress tests.

(1) Robustness against Feature Noise. We simulate inherent clinical noise by injecting Gaussian perturbations into the embeddings: $\mathbf{x}_{\text{noisy}} = \mathbf{x} + \lambda \cdot \boldsymbol{\epsilon}$. As reported in Table II, DSIB exhibits superior resilience. Notably, under high-intensity noise (λ=0.05), the baseline suffers a severe performance drop of **9.26%**, whereas DSIB$_\beta$ limits the degradation to just **5.63%**.

(2) Filtering Redundant Information. We further introduce task-irrelevant redundancy by concatenating random noise dimensions to genomic profiles (50% expansion) and injecting uninformative patterns into WSI embeddings. As shown in Table III, while the baseline performance deteriorates due to overfitting nuisance signals, DSIB maintains stability across all settings. This confirms that the discrete bottleneck successfully compresses the joint representation, discarding redundancy while preserving prognostic evidence.

Compatibility with Robustness Methods (RQ2). We also study whether DSIB is compatible with an existing robustness-enhancing approach, LD-CVAE. As summarized in Table IV, integrating DSIB with LD-CVAE yields consistent additional gains across settings, demonstrating that DSIB can be seamlessly combined with other robustness techniques under different MSPM robustness strategies.

TABLE IV
STATE-OF-THE-ART ROBUST TRAINING METHODS LD-CVAE FOR MSP WITH AND WITHOUT DSIB ON THE LUAD.

Models	Ori.	LD-CVAE [1]	LD-CVAE+DSIB$_\alpha$	LD-CVAE+DSIB$_\beta$
MoME	0.691	0.696	0.703	0.709

TABLE V
ABLATION STUDY OF DSIB ON THE MOME.

Method	BLCA	UCEC	LUAD	Avg.
MoME [21]	0.686	0.706	0.691	0.694
+DSIB w/o $\mathcal{L}_{\text{dist}}$	0.700	0.716	0.704	0.707
+DSIB w/o $\mathcal{L}_{\text{reg}}$	0.692	0.715	0.699	0.702
+DSIB w/o Noise	0.687	0.710	0.693	0.697
Full DSIB (ours)	**0.704**	**0.724**	**0.710**	**0.713**

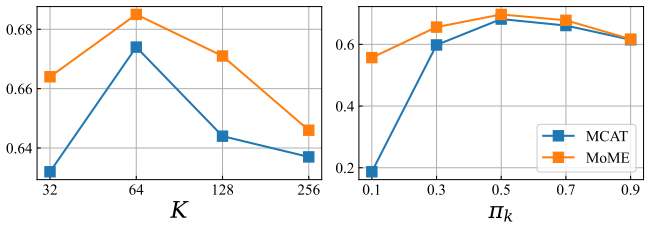

Fig. 2. C-index on the dataset UCEC with different values of parameter K and π_k.

TABLE VI
PERFORMANCE OF MOME UNDER DIFFERENT DIVERGENCES.

Models	Divergence	BLCA	UCEC	LUAD
MoME+DSIB$_\alpha$	$KL(p\|\|p_\delta)$	0.683	0.702	0.688
	$KL(p_\delta\|\|p)$	0.687	0.706	0.691
	JS	**0.695**	**0.715**	**0.700**
MoME+DSIB$_\beta$	$KL(p\|\|p_\delta)$	0.691	0.712	0.697
	$KL(p_\delta\|\|p)$	0.695	0.717	0.701
	JS	**0.703**	**0.723**	**0.709**

D. Ablation Study

(1) Effects of each Component in DSIB. We perform ablations on the MoME baseline (Table V). Removing noise injection (w/o Noise) leads to the largest drop (0.712→0.697), indicating that latent perturbation is the key factor for robust representation learning. Without the regularization term (w/o $\mathcal{L}_{\text{reg}}$), performance decreases to 0.702, showing that the VIB-style constraint is necessary for suppressing redundancy. Without self-distillation (w/o $\mathcal{L}_{\text{dist}}$), the score drops to 0.707, confirming that consistency guidance stabilizes the perturbed features. Overall, the full DSIB achieves the best result (0.712), suggesting these components are complementary.

(2) Impact of Bottleneck Dimension K and Prior π_k. To investigate the sensitivity of DSIB to its core components, we evaluate the performance variance by sweeping the bottleneck dimension $K \in \{32, 64, 128, 256\}$ and varying the prior-related parameter π_k. As illustrated in Fig. 2, the model

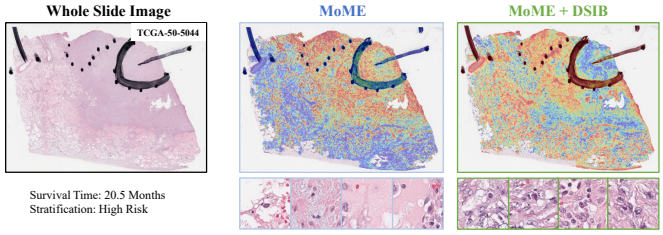

Fig. 3. Attention visualization of MoME [21] and MoME+DSIB.

exhibits robust performance across a reasonable range of hyperparameters, with $K = 32$ and $\pi_k = 0.5$ emerging as the optimal configuration for most cohorts. This suggests that a relatively compact bottleneck effectively filters out modality-specific noise and redundant correlations while preserving essential prognostic signals.

(3) Effects of JS-divergence. We compare KL- and JS-divergence for $\mathcal{L}_{\text{Dist}}^{\text{DSIB}}$ (Table VI). JS-divergence consistently achieves the best performance across all datasets and DSIB variants. Compared with the asymmetric KL objective, the symmetric and bounded JS objective encourages more balanced alignment between p_δ and p, leading to more stable optimization and improved generalization.

E. Visualization

As illustrated in Fig. 3, the visualization compares the attention maps of MoME and MoME+DSIB on a TCGA whole-slide image. The MoME model (center) shows scattered attention across various regions, possibly focusing on irrelevant areas. In contrast, MoME+DSIB (right) refines the attention, highlighting more relevant regions, suggesting that DSIB helps the model suppress noise and redundant features. The coarse black lines or dots in the original WSI are typically markings made by pathologists on physical slides or artifacts, background impurities, or non-tissue regions identified during scanning. Our method effectively avoids these distractions, enhancing robustness and interpretability in survival prediction tasks.

V. Conclusion

In this paper, we introduced the Discrete Self-Distillation Information Bottleneck (DSIB) framework to enhance robustness in multimodal survival prediction. DSIB effectively mitigates the challenges of multimodal heterogeneity by regularizing the fusion representation, suppressing noise, and filtering redundant information. Our extensive experiments on TCGA datasets demonstrate that DSIB consistently improves prognostic accuracy and robustness, with zero inference overhead, making it suitable for practical clinical deployment. DSIB provides a lightweight, plug-and-play solution that can be seamlessly integrated into existing multimodal survival prediction models, improving their stability without requiring architectural modifications.

References

[1] J. Zhou, J. Tang, Y. Zuo, P. Wan, D. Zhang, and W. Shao, "Robust multimodal survival prediction with conditional latent differentiation Variational AutoEncoder," in *CVPR*, 2025, pp. 10 384–10 393.
[2] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 757–770, 2020.
[3] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer cell*, vol. 40, no. 8, pp. 865–878, 2022.
[4] M. Qu, G. Yang, D. Di, T. Su, Y. Gao, Y. Song, and L. Fan, "Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance," *arXiv preprint arXiv:2505.11997*, 2025.
[5] Y. An, J. Chen, H. Lin, Z. Liu, S. Feng, H. Zhang, R. Lan, Z. Liu, and X. Pan, "Ca-mlif: Cross-attention and multimodal low-rank interaction fusion framework for tumor prognostic prediction," in *AAAI*, vol. 39, no. 2, 2025, pp. 1764–1772.
[6] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *AAAI*, vol. 32, no. 1, 2018.
[7] L. Fan, A. Sowmya, E. Meijering, and Y. Song, "Cancer survival prediction from whole slide images with self-supervised learning and slide consistency," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1401–1412, 2022.
[8] W. Shao, T. Wang, Z. Huang, Z. Han, J. Zhang, and K. Huang, "Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3739–3747, 2021.
[9] D. Di, C. Zou, Y. Feng, H. Zhou, R. Ji, Q. Dai, and Y. Gao, "Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5800–5815, 2022.
[10] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *ICCV*, 2021, pp. 4015–4025.
[11] G. Jaume, A. Vaidya, R. J. Chen, D. F. K. Williamson, P. P. Liang, and F. Mahmood, "Modeling dense multimodal interactions between biological pathways and histology for survival prediction," in *CVPR*, 2024, pp. 11 579–11 590.
[12] Y. Xu and H. Chen, "Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction," in *ICCV*, 2023, pp. 21 241–21 251.
[13] Y. Zhang, Y. Xu, J. Chen, F. Xie, and H. Chen, "Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction," *arXiv preprint arXiv:2401.01646*, 2024.
[14] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.
[15] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *NeurIPS*, vol. 30, 2017.
[16] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *ICML*, 2018, pp. 2127–2136.
[17] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.
[18] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *NeurIPS*, vol. 34, pp. 2136–2147, 2021.
[19] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *CVPR*, 2022, pp. 18 802–18 812.
[20] F. Zhou and H. Chen, "Cross-modal translation and alignment for survival analysis," in *ICCV*, 2023, pp. 21 485–21 494.
[21] C. Xiong, H. Chen, H. Zheng, D. Wei, Y. Zheng, J. J. Sung, and I. King, "Mome: Mixture of multimodal experts for cancer survival prediction," in *MICCAI*, 2024, pp. 318–328.