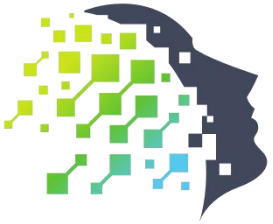


ICLR

Prototypical Information Bottlenecking And Disentangling For Multimodal Cancer Survival Prediction

Yilan Zhang[2†], Yingxue Xu[1†], Jianqi Chen[2], Fengying Xie[2*], Hao Chen[1*]

† Equal contribution *Corresponding authors

1. The Hong Kong University of Science and Technology 2. Beihang University

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

北京航空航天大學
BEIHANG UNIVERSITY

Outline

- ❑ **Research Background**
- ❑ Our Solution: PIBD
- ❑ Experiments and Conclusion

Multimodal Redundancy

Task: Survival Prediction aims to estimate the death risk of patients for prognosis, in which multimodal learning by integrating both **histological** information and **genomic** molecular profiles can benefit majority cancer types.

- However, **massive redundancy** in multimodal data prevents it from extracting **discriminative and compact** information:

Intra-modal redundancy

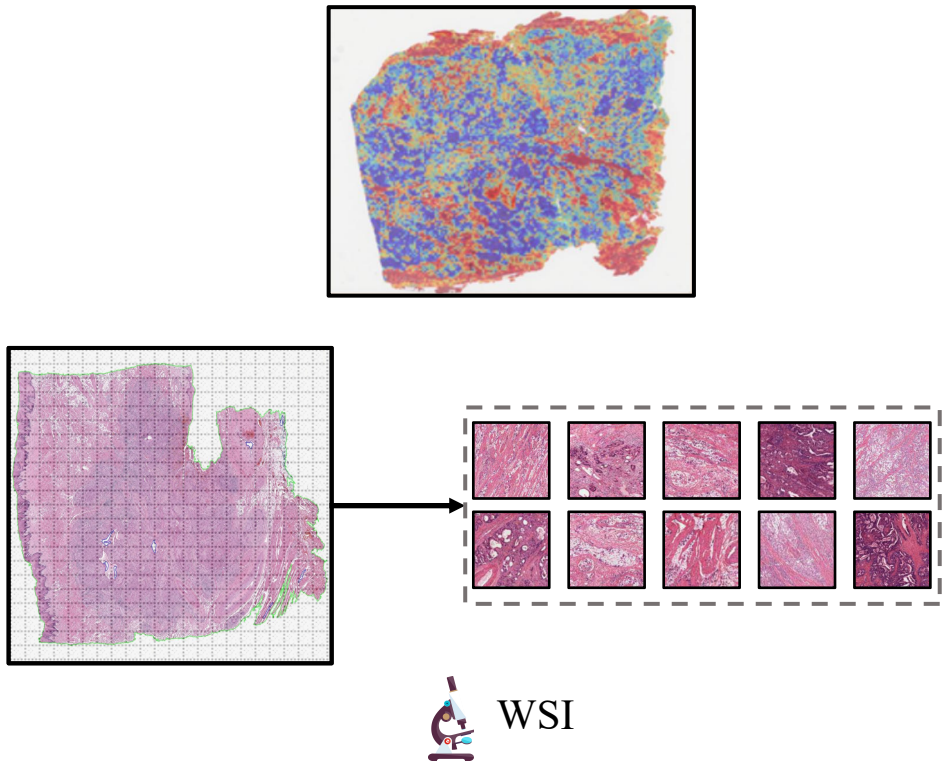
An extensive amount of intra-modal **task-unrelated information** blurs discriminability.

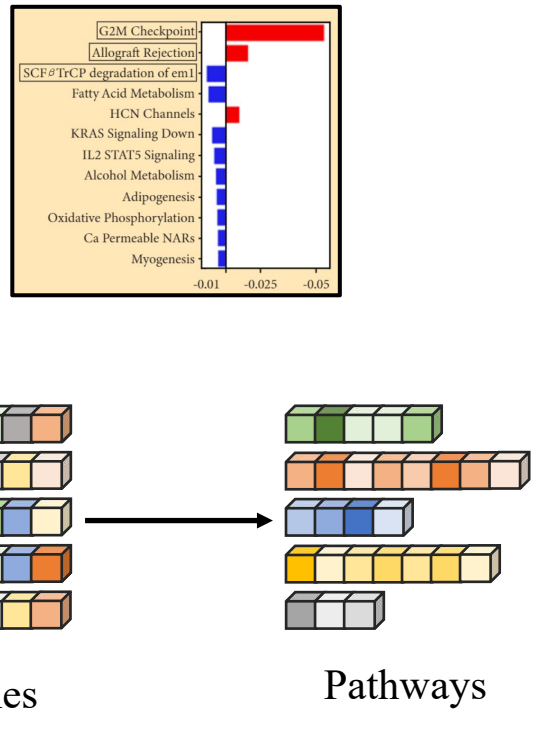
Duplicated information among modalities dominates the representation of multimodal data, which makes modality-specific information prone to being ignored.

Inter-modal redundancy

Multimodal Redundancy

Q1: Intra-modal redundancy



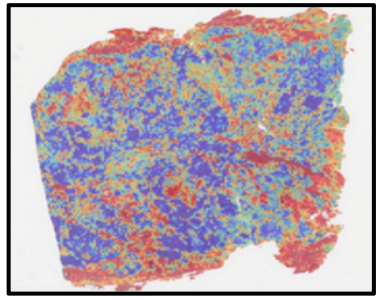


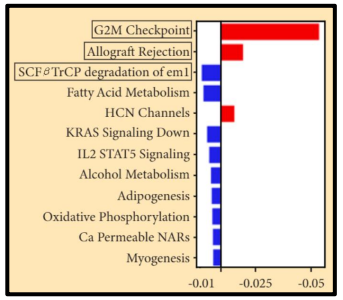
- The region of interest only occupies **a small portion** of gigapixel WSIs.
- These pathways can yield hundreds to thousands of groups, and **only a few specific pathways** exhibit a strong correlation with patient prognosis.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Jaume G, Vaidya A, Chen R, et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction[J]. arXiv preprint arXiv:2304.06819, 2023.
Jiang X, Xia Y, Meng H, et al. Identification of a nuclear mitochondrial-related multi-genes signature to predict the prognosis of bladder cancer[J]. Frontiers in Oncology, 2021, 11: 746029.

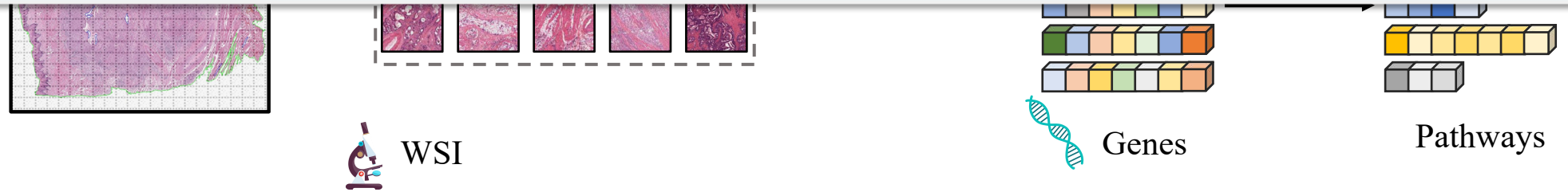
Multimodal Redundancy

Q1: Intra-modal redundancy





How can we capture the discriminative information from single modality by eliminating its redundancy?



- The region of interest only occupies **a small portion** of gigapixel WSIs.
- These pathways can yield hundreds to thousands of groups, and **only a few specific pathways** exhibit a strong correlation with patient prognosis.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Jaume G, Vaidya A, Chen R, et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction[J]. arXiv preprint arXiv:2304.06819, 2023.
Jiang X, Xia Y, Meng H, et al. Identification of a nuclear mitochondrial-related multi-genes signature to predict the prognosis of bladder cancer[J]. Frontiers in Oncology, 2021, 11: 746029.

Multimodal Redundancy

Q2: Inter-modal redundancy

❑ The redundancy stemming from this **duplicated information** can complicate the knowledge extraction.

- **Common** information often **dominates** aligning and integrating multimodal information
- Lead to the suppression of **modality-specific** information, thereby disregarding the wealth of distinctive perspectives.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.

Multimodal Redundancy

Q2: Inter-modal redundancy

❑ The redundancy stemming from this **duplicated information** can complicate the knowledge extraction.

How can we capture compact yet comprehensive knowledge from the dominant overlapping information in multimodal data?

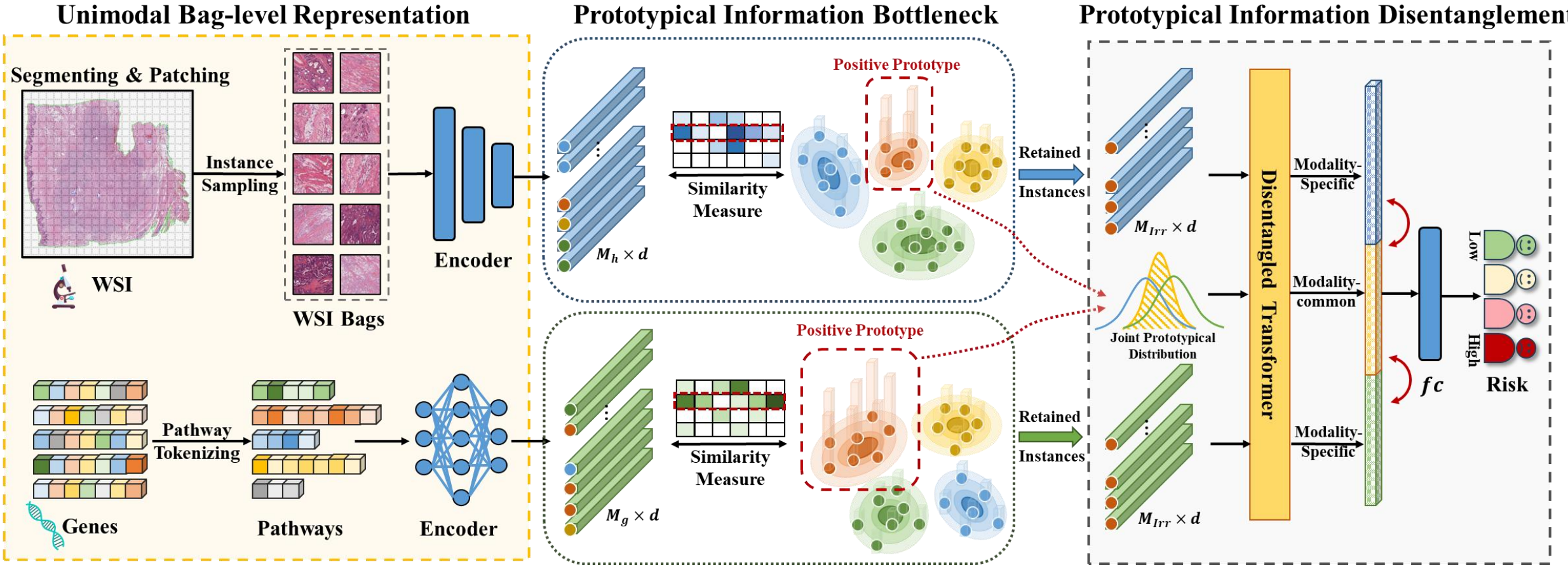
- Lead to the suppression of **modality-specific** information, thereby disregarding the wealth of distinctive perspectives.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.

Outline

- ❑ Research Background
- ❑ **Our Solution: PIBD**
- ❑ Experiments and Conclusion

PIBD: Overall Architecture



(a) **Intra-model redundancy**: A **Prototypical Information Bottleneck (PIB)** module for selecting discriminative instances within a modality.
(b) **Inter-model redundancy**: A **Prototypical Information Disentanglement (PID)** module for getting independent modality-common and modality-specific knowledge.

PIBD: Prototypical Information Bottleneck

Preliminary of Information Bottleneck

- Objective: Variable Z is maximally expressive about the target Y , while compressing the original information from the input X.

$$\text{Maximize } R_{IB} = I(Z,Y) - \beta I(Z,X)$$

$$\mathcal{L}_{IB} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}_{z\sim p(z|x_n)}[-log q_\theta(y_n|z)]}_{\text{Lower bound}} + \underbrace{\beta KL[p(z|x_n), r(z)]}_{\text{Upper bound}}$$

Bag: $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$

in which,
$q_\theta(y_n|z)$ is a variational approximation of the intractable likelihood p(y|z).
$p(z|x)$ is the posterior distribution over z. $p(z|x) \approx q_\theta(z|x) = \mathcal{N}(z|f_E^\mu(x), f_E^\Sigma(x))$, where f_E is and MLP encoder.
$r(z)$ is a prior spherical Gaussian.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In International Conference on Learning Representations, 2016.

PIBD: Prototypical Information Bottleneck

Preliminary of Information Bottleneck

- Objective: Variable Z is maximally expressive about the target Y , while compressing the original information from the input X.

How can we compress a bag X to Z by decreasing the KL divergence between p(z|x) and r(z) in the context of MIL method?

in which,
$q_\theta(y_n|z)$ is a variational approximation of the intractable likelihood p(y|z).
$p(z|x)$ is the posterior distribution over z. $p(z|x) \approx q_\theta(z|x) = \mathcal{N}(z|f_E^\mu(x), f_E^\Sigma(x))$, where f_E is and MLP encoder.
$r(z)$ is a prior spherical Gaussian.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In International Conference on Learning Representations, 2016.

PIBD: Prototypical Information Bottleneck

Apply IB for every instance z to compress Z to X.

- Directly employ the variational approximation $q_\theta(z|x)$ in VIB to learn a compact representation for **each instance** $x \in \mathbf{x}$ in the bag.

Problems:

*1) How to derive the **overall distribution of the entire bag** $p(z|\mathbf{x})$ for a bag $\mathbf{x}$ based on such a large number of individual instance distributions?*

2) All compact instance features  *A compact bag*

Information Bottleneck (IB):

✓ Provides a promising solution to **compress unnecessary** redundancy from itself while **maximizing discriminative** information about task targets.

✗ Suffer from the **high-dimensional** computational challenges posed by **massive patches** of a gigapixel WSI and thousands of pathways.

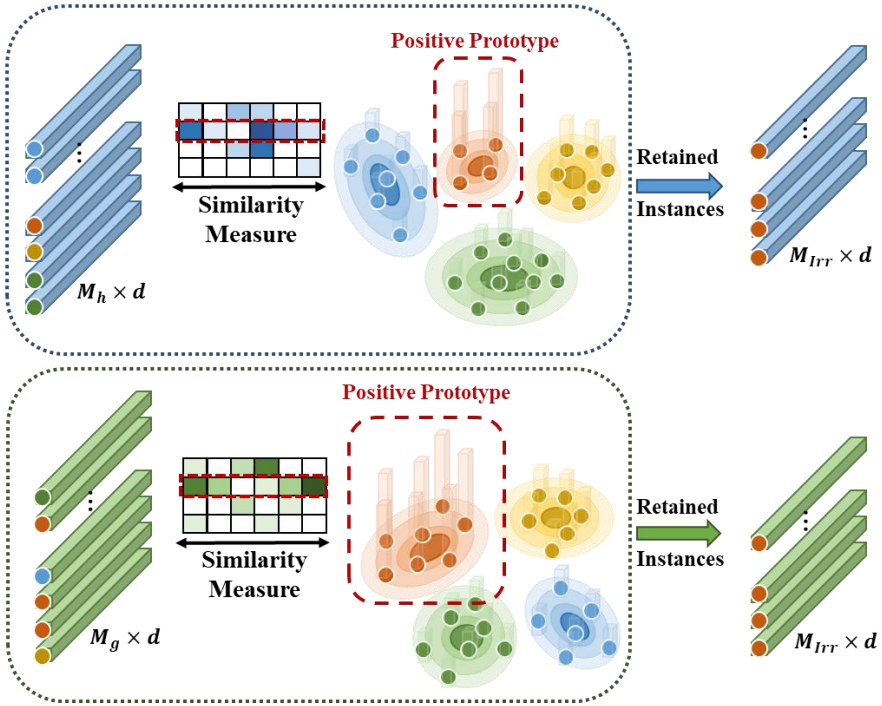
Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Li H, Zhu C, Zhang Y, et al. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7454-7463.

PIBD: Prototypical Information Bottleneck

Objective: Intra-model redundancy reduction

- Directly approximate bag-level distribution $p(z|\mathbf{x})$ with a parametric distribution $p(\hat{z})$ represented by a group of **prototypes** $\mathbf{P} = \{\mathcal{N}(\hat{z}|\mu_y, \Sigma_y)\}_{y=1}^{2N_t}$ for **different risk levels**.

$$p(z|\mathbf{x}) = p(z|\mathbf{x}, y) \approx p(\hat{z}|y)$$ "prototype distribution at risk band y"



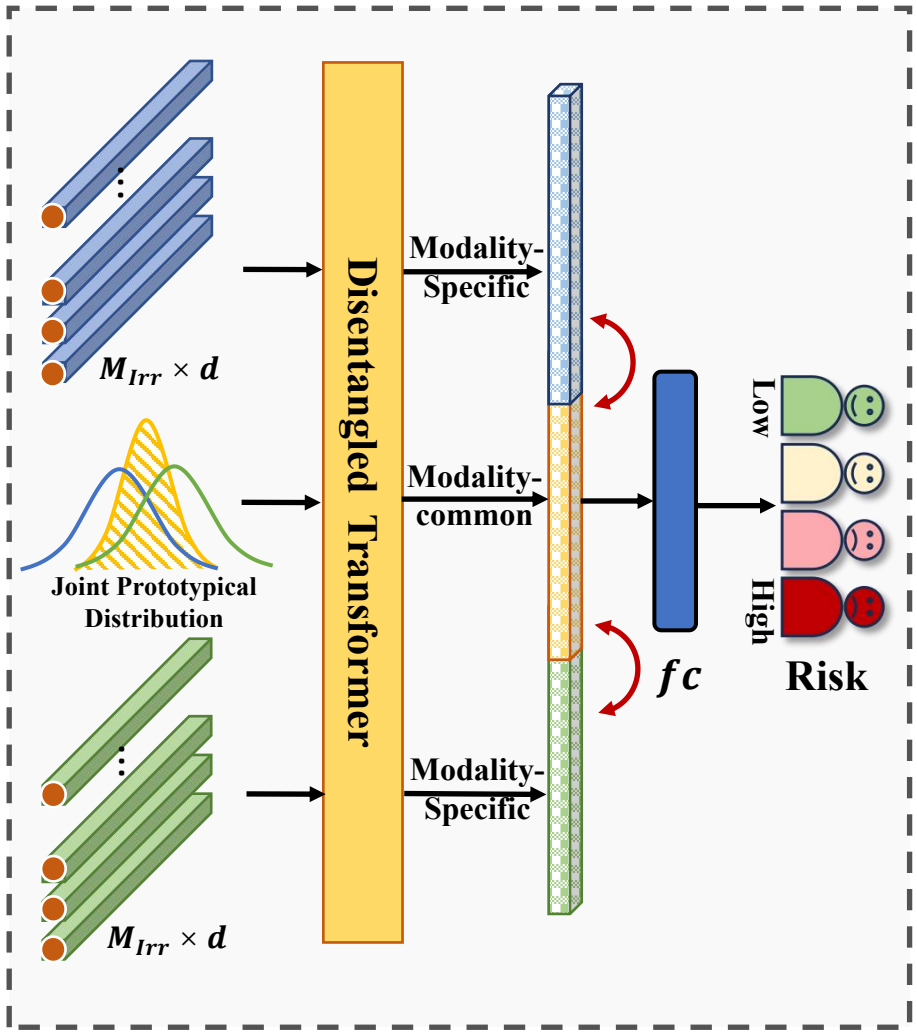
✓ To achieve this, we **maximize the similarity** between $p(\hat{z}|y)$ and distributions of latent features $z = f_E(x)$ for a bunch of instances.

$$\mathcal{L}_{pro} = \frac{1}{N_D}\sum_{i=1}^{N_D} -Sim(\hat{z}_+^{(i)}, \tilde{\mathbf{z}}_+^{(i)}) + \frac{1}{2N_t - 1}\sum_{i=1}^{2N_t-1} Sim(\hat{z}_{-,n}^{(i)}, \tilde{\mathbf{z}}_{-,n}^{(i)})$$

✓ As a result, we just need to optimize the parametric prototypes $\boldsymbol{p}(\hat{\boldsymbol{z}})$ and $\boldsymbol{f_E}$ for a bag x, instead of modeling $p(z|x)$ for each instance of the bag.

$$\mathcal{L}_{PIB} = \frac{1}{2N_t}\sum_{n=1}^{2N_t}\{\alpha\mathcal{L}_{surv}(\hat{z}^{(n)}, t^{(n)}, c^{(n)}) + \beta KL[\mathcal{N}(\hat{z}|\mu_n, \Sigma_n), r(z)]\} + \gamma\mathcal{L}_{pro}$$

PIBD: Prototypical Information Disentanglement



Objective: Inter-model redundancy reduction

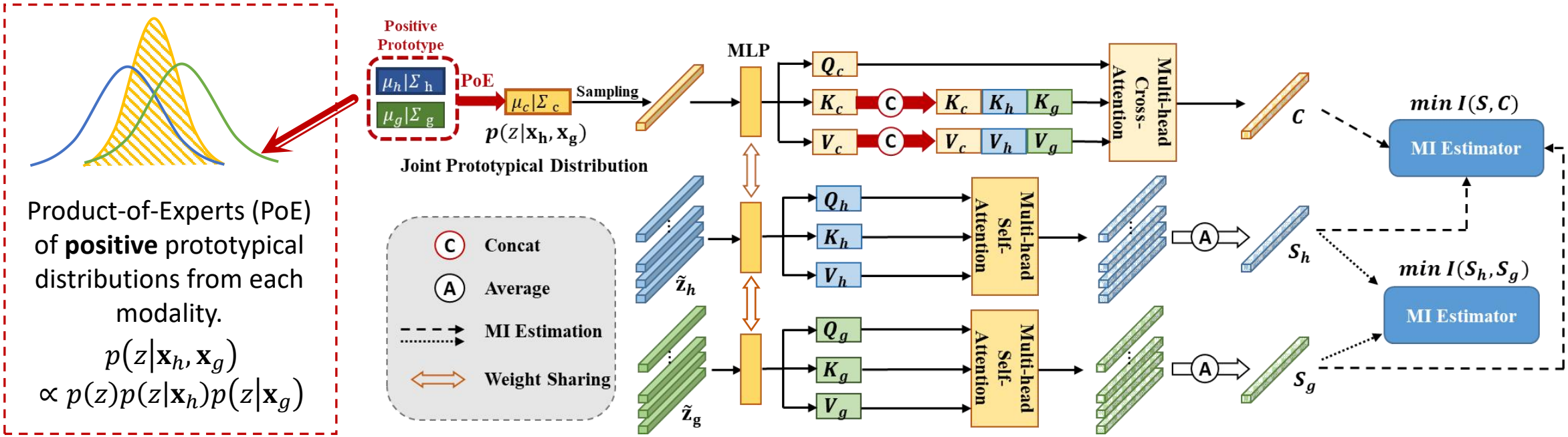
- Get independent modality-**common** and modality-**specific** knowledge from the dominant overlapping information in multimodal data.

$$minimize\ I(S, C) + I(S_h, S_g),\ where\ S = Cat(S_h, S_g)$$

In which,
S_h and S_g is the modality-specific feature of histological modality and genomic modality, respectively.
C is the modality-common feature.

PIBD: Prototypical Information Disentanglement

Disentangled Transformer



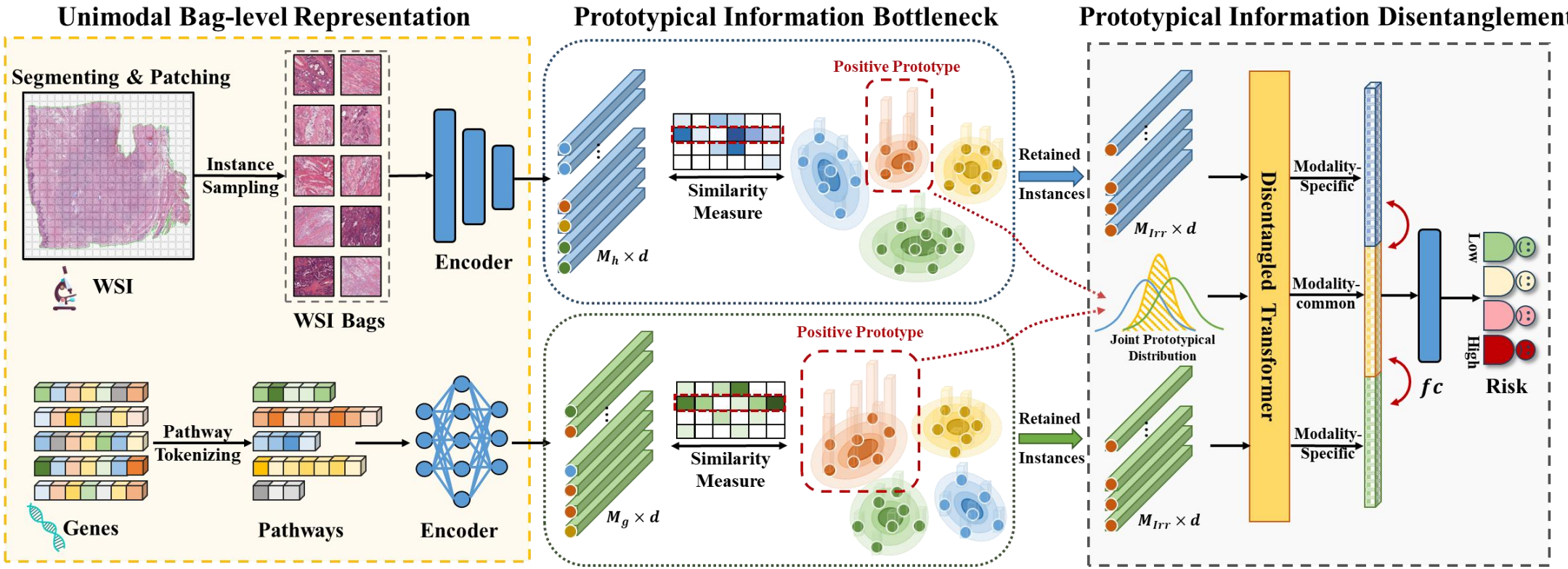
Product-of-Experts (PoE) of **positive** prototypical distributions from each modality.

$$p(z|\mathbf{x}_h, \mathbf{x}_g) \propto p(z)p(z|\mathbf{x}_h)p(z|\mathbf{x}_g)$$

- ✓ Reused the joint **prototypical distributions** $p(z|\mathbf{x}_h, \mathbf{x}_g)$ modeled by aforementioned PIB to guide the extraction of **common knowledge**.
- ✓ Enforced the model to learn knowledge **different from the joint prototypical distribution**, considered as the modality-**specific** knowledge.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Cao, Yanshuai, and David J. Fleet. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. arXiv preprint arXiv:1410.7827, 2014.

PIBD: Overall Architecture



$$\mathcal{L}_{PIB} = \frac{1}{2N_t}\sum_{n=1}^{2N_t}\left\{\alpha\mathcal{L}_{surv}\left(\hat{z}^{(n)}, t^{(n)}, c^{(n)}\right) + \beta KL[\mathcal{N}(\hat{z}|\mu_n, \Sigma_n), r(z)]\right\} + \gamma\mathcal{L}_{pro}$$

$$\mathcal{L}_{PID} = I(S, C) + I(S_h, S_g)$$

Overall Loss of PIBD (PIB + PID): $\mathcal{L} = \mathcal{L}_{surv} + \mathcal{L}_{PIB}^{h} + \mathcal{L}_{PIB}^{g} + \lambda\mathcal{L}_{PID}$

Outline

- ❑ Research Background
- ❑ Our Solution: PIBD
- ❑ **Experiments and Conclusion**

EXP1: Comparison

❑ Comparison Experiments

- Achieves superior performance in 4 out of 5 benchmarks.
- Outperforms the second-best method by 1.6% in overall C-index.
- Compared with other IB-based methods, we achieve superior performance on all cancer datasets (overall: 3.5%-9.3% ↑).

Model	Modality	BRCA (N=869)	BLCA (N=359)	COADREAD (N=296)	HNSC (N=392)	STAD (N=317)	Overall
†MLP	g.	0.622 ± 0.079	0.530 ± 0.077	0.712 ± 0.114	0.520 ± 0.064	0.497 ± 0.031	0.576
†SNN	g.	0.621 ± 0.073	0.521 ± 0.070	0.711 ± 0.162	0.514 ± 0.076	0.485 ± 0.047	0.570
†SNNTrans	g.	0.679 ± 0.053	0.583 ± 0.060	0.739 ± 0.124	0.570 ± 0.035	0.547 ± 0.041	0.622
†ABMIL	h.	0.672 ± 0.051	0.624 ± 0.059	0.730 ± 0.151	0.624 ± 0.042	0.636 ± 0.043	0.657
†AMISL	h.	0.681 ± 0.036	0.627 ± 0.032	0.710 ± 0.091	0.607 ± 0.048	0.553 ± 0.012	0.636
†TransMIL	h.	0.663 ± 0.053	0.617 ± 0.045	0.747 ± 0.151	0.619 ± 0.062	0.660 ± 0.072	0.661
†CLAM-SB	h.	0.675 ± 0.074	0.643 ± 0.044	0.717 ± 0.172	0.630 ± 0.048	0.616 ± 0.078	0.656
†CLAM-MB	h.	0.696 ± 0.098	0.623 ± 0.045	0.721 ± 0.159	0.620 ± 0.034	0.648 ± 0.050	0.662
‡SNNTrans+CLAM-MB	g.+h.	0.699 ± 0.064	0.625 ± 0.060	0.716 ± 0.160	0.638 ± 0.066	0.629 ± 0.065	0.661
‡Porpoise(Cat)	g.+h.	0.668 ± 0.070	0.617 ± 0.056	0.738 ± 0.151	0.614 ± 0.058	0.660 ± 0.106	0.660
‡Porpoise(KP)	g.+h.	0.691 ± 0.038	0.619 ± 0.055	0.721 ± 0.157	0.630 ± 0.040	0.661 ± 0.085	0.664
‡MCAT(Cat)	g.+h.	0.685 ± 0.109	0.640 ± 0.076	0.724 ± 0.137	0.564 ± 0.840	0.625 ± 0.118	0.647
‡MCAT(KP)	g.+h.	0.727 ± 0.027	0.644 ± 0.062	0.709 ± 0.162	0.618 ± 0.093	0.643 ± 0.075	0.668
‡MOTCat	g.+h.	0.727 ± 0.027	0.659 ± 0.069	0.742 ± 0.124	**0.656 ± 0.041**	0.621 ± 0.065	0.681
‡SurvPath	g.+h.	0.724 ± 0.094	0.660 ± 0.054	0.758 ± 0.143	0.606 ± 0.080	0.667 ± 0.035	0.683
*CLAM-SB-FT	h.	0.606 ± 0.110	0.633 ± 0.065	0.725 ± 0.150	0.620 ± 0.084	0.654 ± 0.051	0.648
*MIB	g.+h.	0.602 ± 0.112	0.573 ± 0.036	0.711 ± 0.182	0.555 ± 0.055	0.588 ± 0.057	0.606
*DeepIMV	g.+h.	0.659 ± 0.089	0.638 ± 0.054	0.749 ± 0.145	0.604 ± 0.061	0.597 ± 0.047	0.649
*L-MIB	g.+h.	0.687 ± 0.071	0.662 ± 0.093	0.720 ± 0.167	0.615 ± 0.085	0.634 ± 0.060	0.664
*,‡**PIBD**	g.+h.	**0.736 ± 0.072**	**0.667 ± 0.061**	**0.768 ± 0.124**	0.640 ± 0.039	**0.684 ± 0.035**	**0.699**

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.

EXP1: Comparison

❑ KM Analysis

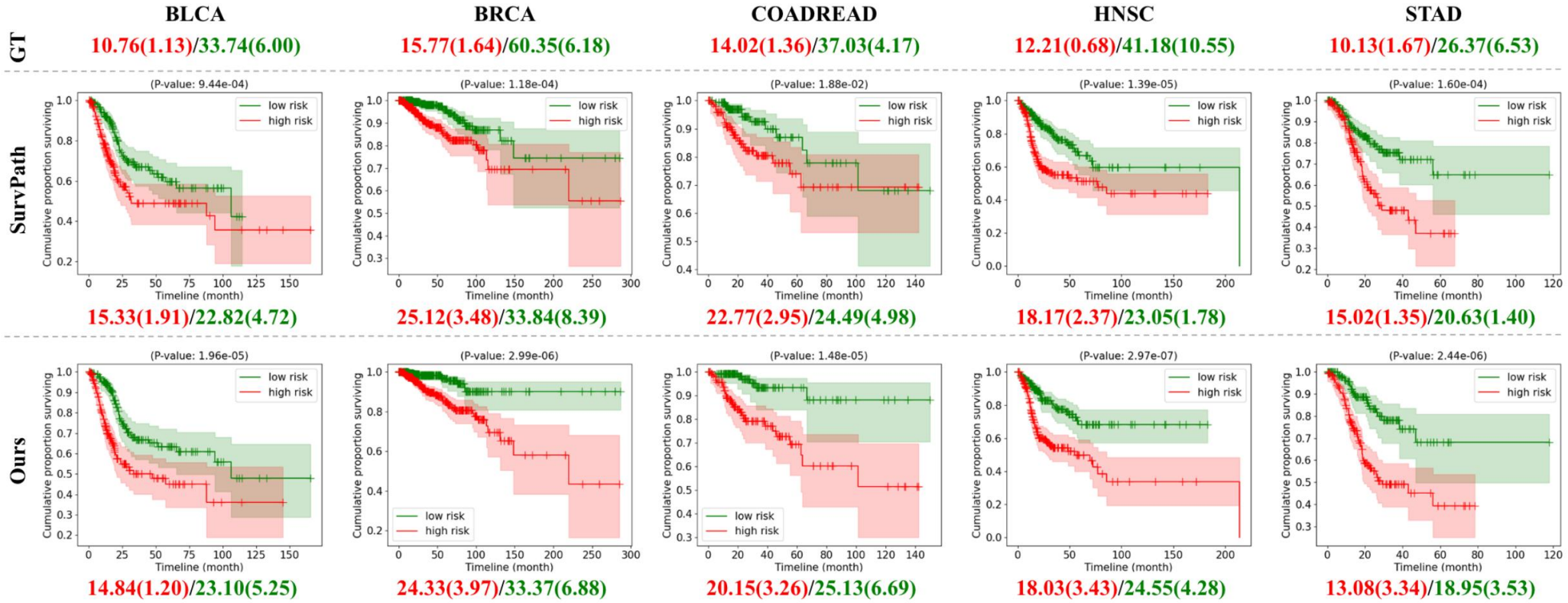


Figure 3: Kaplan-Meier curves of predicted high-risk (red) and low-risk (green) groups. A P-value $<$ 0.05 indicates statistical significance, and the shaded regions represent the confident intervals. The median survival months are reported in the format of "high-risk: mean(std)/low-risk: mean(std)"

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.

Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, Paul Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. arXiv preprint arXiv:2304.06819, 2023.

EXP2: Ablation Study

❑ Ablation study

Variants	PIB	PID	BRCA	BLCA	COADREAD	HNSC	STAD	Overall
AP			0.684 ± 0.044	0.619 ± 0.090	0.713 ± 0.161	0.567 ± 0.073	0.609 ± 0.048	0.638
PIB(AP)	✓		0.705 ± 0.108	0.593 ± 0.038	0.753 ± 0.143	0.623 ± 0.107	0.613 ± 0.071	0.657
TransMIL			0.672 ± 0.088	0.636 ± 0.059	0.750 ± 0.133	0.591 ± 0.080	0.662 ± 0.090	0.662
PIB(TransMIL)	✓		0.696 ± 0.069	0.648 ± 0.074	0.757 ± 0.176	0.615 ± 0.062	0.643 ± 0.074	0.672
PIBD	✓	✓	$\mathbf{0.736 \pm 0.072}$	$\mathbf{0.667 \pm 0.061}$	$\mathbf{0.768 \pm 0.124}$	$\mathbf{0.640 \pm 0.039}$	$\mathbf{0.684 \pm 0.035}$	**0.699**

*Average Pooling (AP)

For ablating PIB, we established two baselines:

1. one involves direct average pooling (AP) on original features
2. the other employs a non-disentangled TransMIL encoder as a strong baseline.

For ablating PID, we conduct a comparison between our PIBD and the baseline using the **non-disentangled** TransMIL with PIB.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Shao Z, Bian H, Chen Y, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification[J]. Advances in neural information processing systems, 2021, 34: 2136-2147.

EXP3: Discriminativeness of PIB

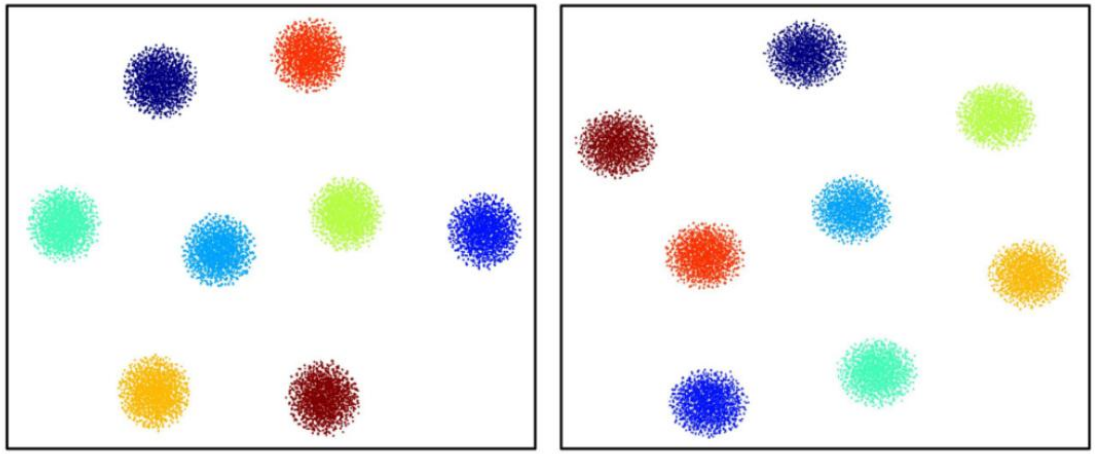


Figure 4: Visualization of prototypes.

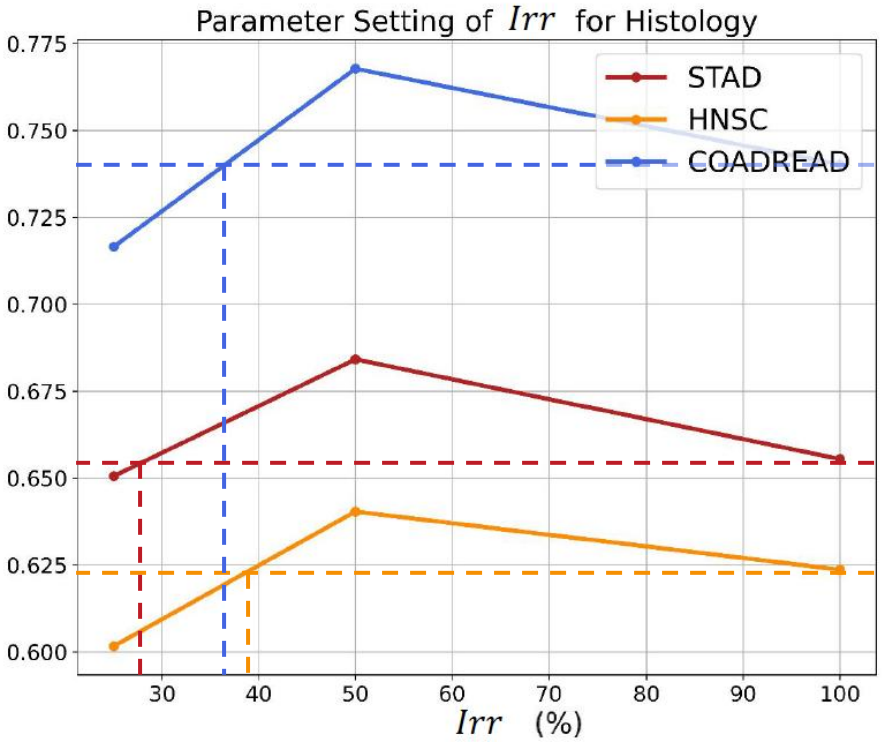
Table 3: Interventions in PIB. We conduct interventions by either removing the positive prototype or randomly deleting one of the negative prototypes.

Intervention	BLCA	COADREAD	STAD
Positive	0.401 ± 0.086	0.471 ± 0.196	0.384 ± 0.110
Negative	0.645 ± 0.067	0.731 ± 0.106	0.672 ± 0.055
w/o Intervention	0.667 ± 0.061	0.768 ± 0.124	0.684 ± 0.035

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.
Shao Z, Bian H, Chen Y, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification[J]. Advances in neural information processing systems, 2021, 34: 2136-2147.

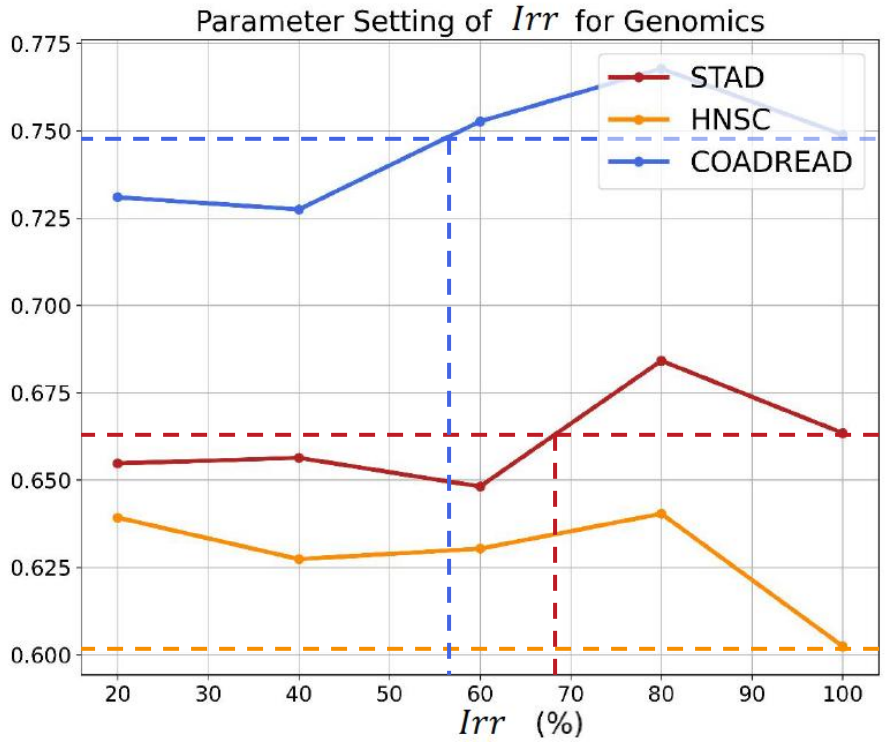
EXP4: Redundancy Removal of PIB

❑ Information retention rate



↓ 60~75%

↓ 50%



↓ 30~45%

↓ 20%

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.

Conclusion

- Inspired by information theory for mitigating redundancy, we propose a new multimodal cancer survival framework, PIBD, addressing both **“intra-modal”** and **“inter-modal”** redundancy challenges.

- We design a new IB variant, PIB, that models prototypes for selecting discriminative information to reduce intra-modal redundancy, w.r.t addressing sparsity of patch features in **MIL** via **prototypes**. This provides **a new solution to compress information of a bag via information bottleneck**.

- PID addresses inter-modal redundancy by **decoupling multimodal data into distinct components** with the guidance of joint prototypical distribution.

Zhang Y., Xu Y., et al. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. ICLR, 2024.

Thanks!