Inference-Time Dynamic Modality Selection for Incomplete Multimodal Classification


Siyi Du


Xinzhe Luo


Declan P. O'Regan


Chen Qin

Imperial College London

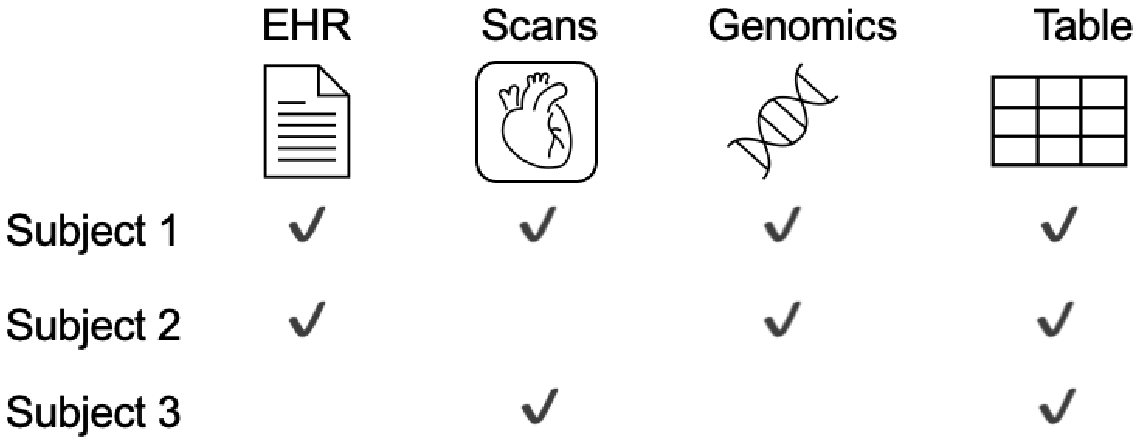
International Conference on Learning Representations (ICLR)

Rio de Janeiro, Brazil

April 23 , 2026

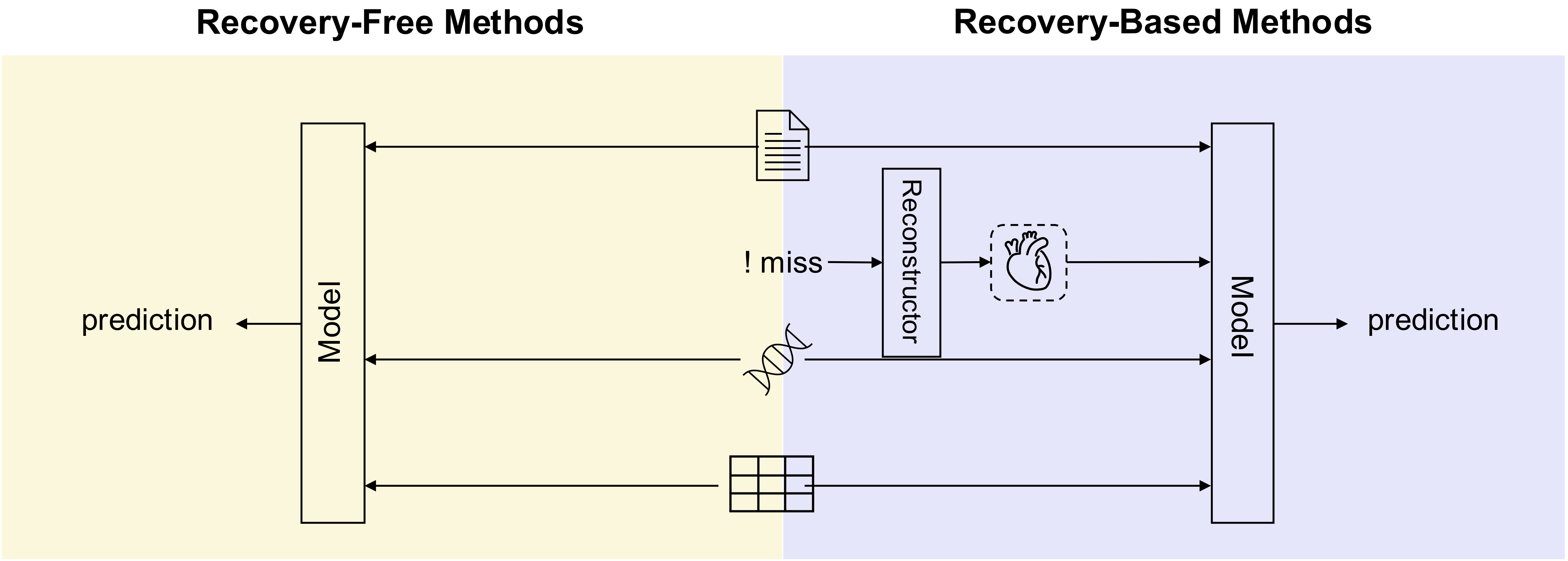
Incomplete Multimodal Learning

- Multimodal deep learning (MDL) has achieved remarkable success across various domains
- MDL models typically pre-suppose full-modality availability at inference
- **However**, in real-world applications, multimodal samples are often incomplete

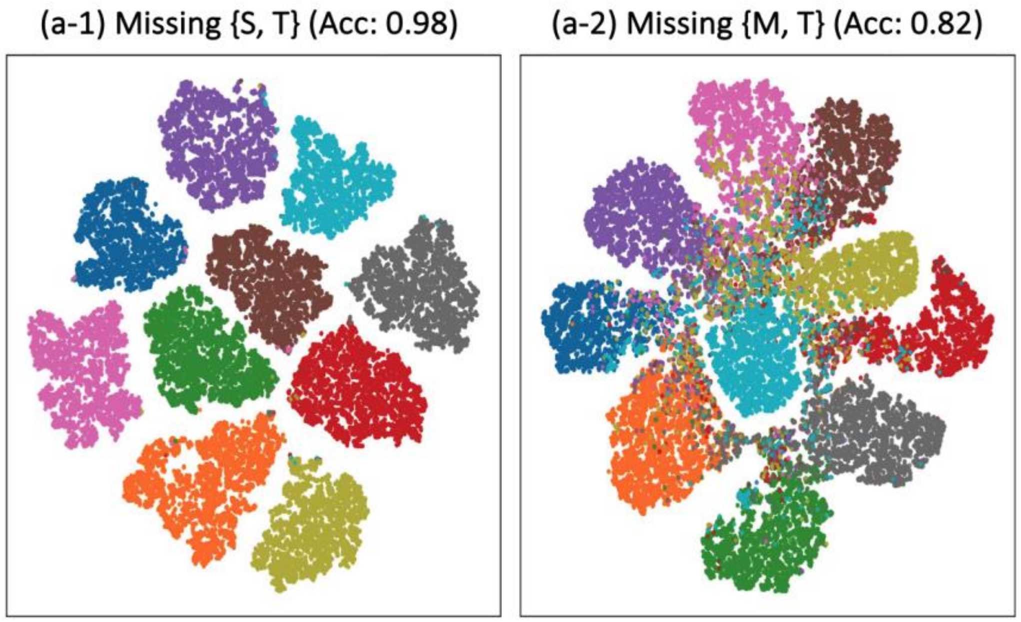


Existing Approaches for Incomplete Multimodal Learning

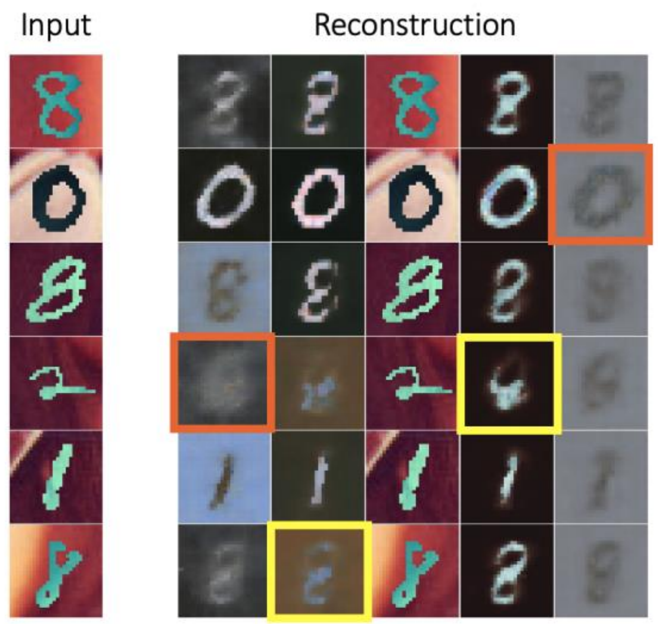
- Incomplete MDL approaches can broadly categorized into two types



Discarding-Imputation Dilemma Caused by Modality Heterogeneity



(a) t-SNE visualization of ModDrop's modality-incomplete features on the MNIST(M)-SVHN(S)-TEXT(T) (MST) training dataset



(b) Sample modality generation by MoPoE on the PolyMNIST test dataset

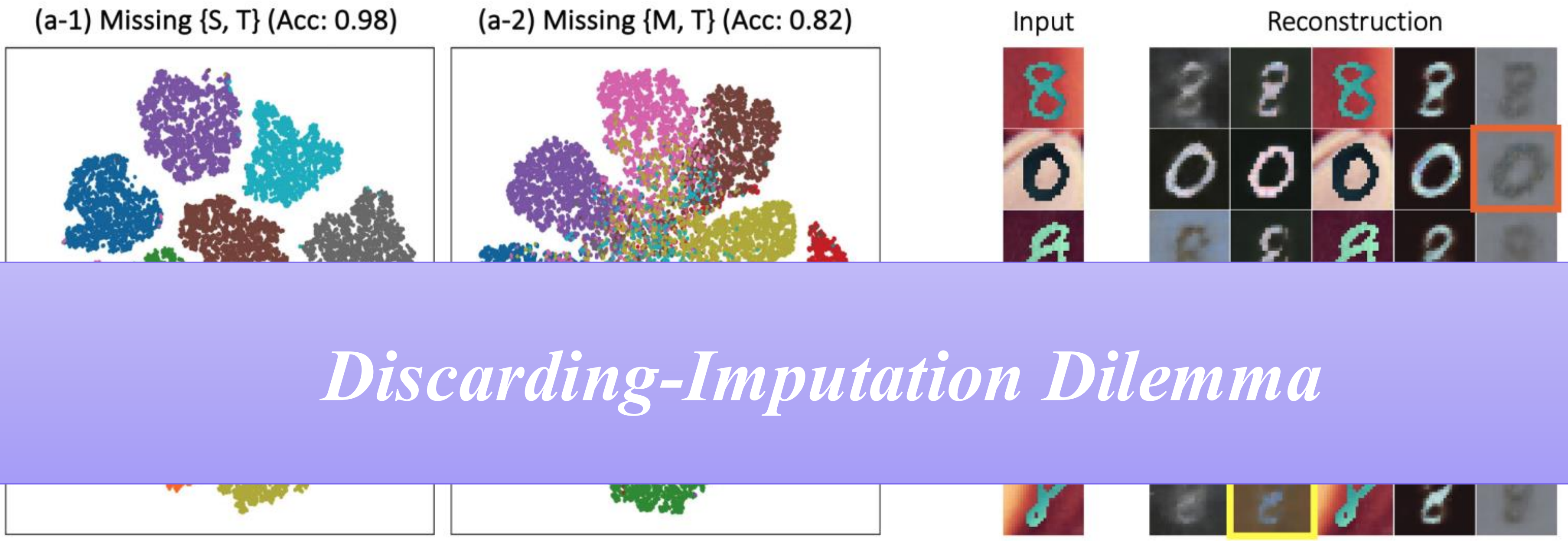
Recovery-free methods

- Learn less-distinguishable features

Recovery-based methods

- May generate unreliable reconstructions
  - Low-fidelity
  - Semantically misaligned

Discarding-Imputation Dilemma Caused by Modality Heterogeneity



Discarding-Imputation Dilemma

(a) t-SNE visualization of ModDrop's modality-incomplete features on the MNIST(M)-SVHN(S)-TEXT(T) (MST) training dataset

(b) Sample modality generation by MoPoE on the PolyMNIST test dataset

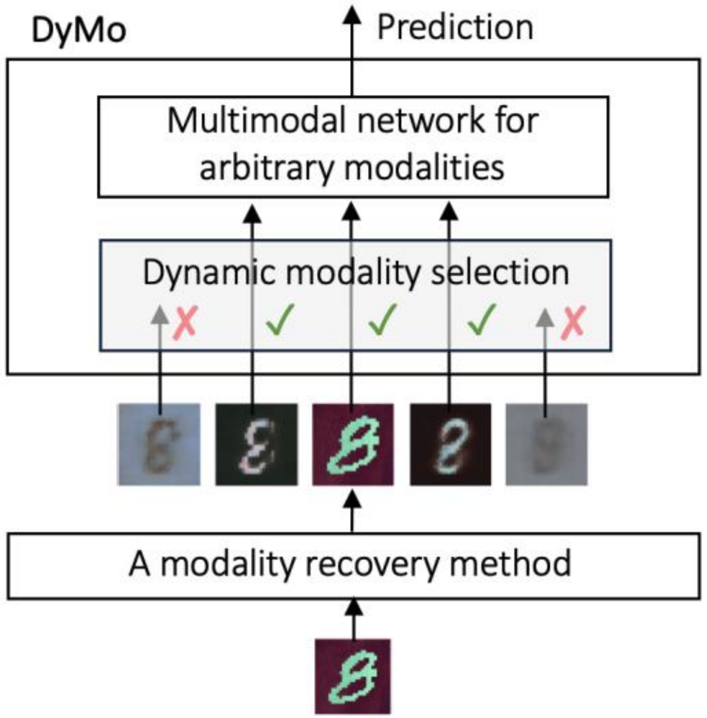
Recovery-free methods

- Learn less-distinguishable features

Recovery-based methods

- May generate unreliable reconstructions
  - Low-fidelity
  - Semantically misaligned

Contributions



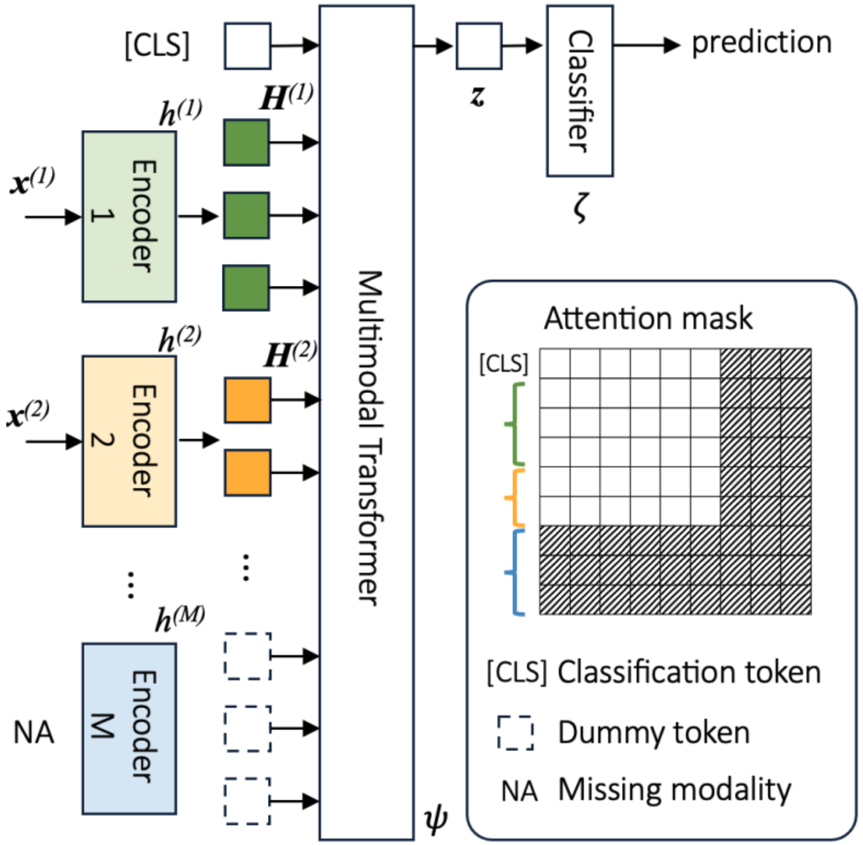
A New Perspective

- Dynamically selects and fuses recovered modalities

Our DyMo

- A dynamic selection algorithm formulated on multimodal task-relevant information gain
- A multimodal network architecture for arbitrary modality input
- A tailored training strategy to learn robust latent features

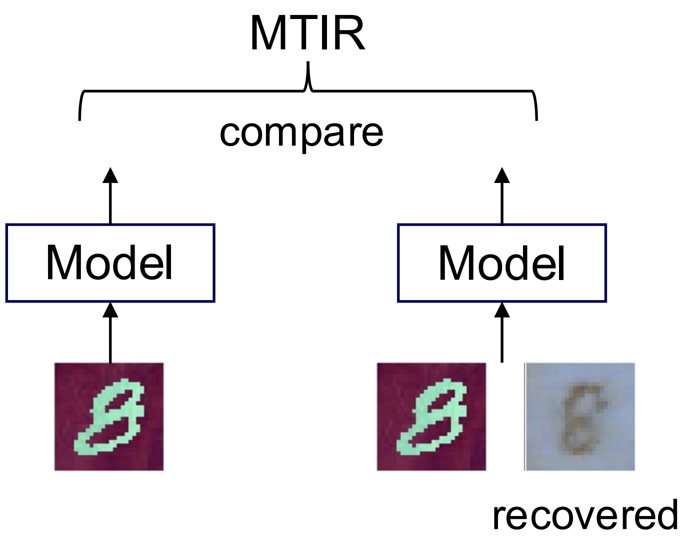
Multimodal Network Architecture for Arbitrary Modalities



Model Overview

- Modality-specific encoders
- Multimodal transformer
- Linear classifier

Multimodal Task-Relevant Information Reward (MTIR)



MTIR Reward of A Recovered Modality

- Reward > 0: Useful. Introduces additional task-relevant information.
- Reward = 0: Low-fidelity. Provide negligible benefit.
- Reward < 0: Semantically misaligned. Potentially disregard the representation.

Multimodal Task-Relevant Information Reward (MTIR)

Multimodal task-relevant information

$$I(Y;\boldsymbol{Z}) = \mathbb{E}_{p(y,\boldsymbol{z})} \log \frac{p(y,\boldsymbol{z})}{p(y)p(\boldsymbol{z})}$$

Due to unknown true data distribution, we derive a lower bound of $I(Y;Z)$, using the empirical classification cross-entropy (CE) loss

$$I(Y;\boldsymbol{Z}) \geq H(Y) - \hat{\mathcal{L}}_{\mathrm{ce}} - G\sqrt{\frac{\ln(1/\delta)}{2|\mathcal{D}|}}, \quad \text{with probability at least } 1-\delta,$$

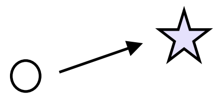
Therefore, MTIR is define as

$$\mathrm{R}(\tilde{\boldsymbol{x}}^{(u)}, \mathbb{X}^{O}) = \ell_{\mathrm{ce}}\left[f(\mathbb{X}^{O}), y\right] - \ell_{\mathrm{ce}}\left[f(\mathbb{X}^{O}, \tilde{\boldsymbol{x}}^{(u)}), y\right] = -\log p_f(y|\boldsymbol{z}) + \log p_f(y|\boldsymbol{z}^{u}),$$

To improve the robustness of MTIR, we further investigate representation shifts in the latent space

$$p(y=k|\boldsymbol{z}) = \frac{\exp(-d_\phi(\boldsymbol{z}, \boldsymbol{c}_k))}{\sum_{k'=1}^{K} \exp(-d_\phi(\boldsymbol{z}, \boldsymbol{c}_{k'}))}, \quad \boldsymbol{c}_k = \frac{1}{\sum_{i=1}^{N} \mathbb{I}[y_i = k]} \sum_{i:y_i=k} \boldsymbol{z}_i,$$

$$\mathrm{R}(\tilde{\boldsymbol{x}}^{(u)}, \mathbb{X}^{O}) = -\log \frac{\exp(-d_\phi(\boldsymbol{z}, \boldsymbol{c}_{\hat{y}}))}{\sum_{k'=1}^{K} \exp(-d_\phi(\boldsymbol{z}, \boldsymbol{c}_{k'}))} + \log \frac{\exp(-d_\phi(\boldsymbol{z}^{u}, \boldsymbol{c}_{\hat{y}^{u}}))}{\sum_{k'=1}^{K} \exp(-d_\phi(\boldsymbol{z}^{u}, \boldsymbol{c}_{k'}))}.$$



Move closer to the prototype → MTIR ↑

Intra-Class Similarity Calibration

We refine the reward by accounting for how representative a sample is within its predicted class cluster

$$\mathrm{R}^*(\tilde{\boldsymbol{x}}^{(u)}, \mathbb{X}^O) = -\log \frac{\exp(-d_\phi(\boldsymbol{z}, \boldsymbol{c}_{\hat{y}}))}{\sum_{k'=1}^{K} \exp(-d_\phi(\boldsymbol{z}, \boldsymbol{c}_{k'}))} + \alpha \times \log \frac{\exp(-d_\phi(\boldsymbol{z}^u, \boldsymbol{c}_{\hat{y}^u}))}{\sum_{k'=1}^{K} \exp(-d_\phi(\boldsymbol{z}^u, \boldsymbol{c}_{k'}))}.$$

Intra-class similarity (ICS) score calculation

$$\mathrm{ICS}(y = k, \boldsymbol{z}) = \mathbb{P}(d_\phi > d_\phi(\boldsymbol{z}, \boldsymbol{c}_k) | d_\phi > 0) = 2(1 - \Phi(d_\phi(\boldsymbol{z}, \boldsymbol{c}_k))),$$

Our asymmetric calibration term

$$\alpha = \begin{cases} 1 & \text{if } \mathrm{ICS}(y = \hat{y}^u, \boldsymbol{z}^u) > \mathrm{ICS}(y = \hat{y}, \boldsymbol{z}), \\ \frac{\mathrm{ICS}(y = \hat{y}^u, \boldsymbol{z}^u)}{\mathrm{ICS}(y = \hat{y}, \boldsymbol{z})} & \text{otherwise}, \end{cases}$$

Tailored Training Strategy

- Incomplete simulation training

$$\mathcal{L}_{\text{class}} = -\frac{1}{A}\frac{1}{B}\sum_{\mathcal{S}\sim\mathcal{U}_A}\sum_{i=1}^{B}\log p_f\left(y_i \mid \{\boldsymbol{x}^{(m)}\}_{m\in\mathcal{S}}\right),$$

- Auxiliary missing-agnostic contrastive loss

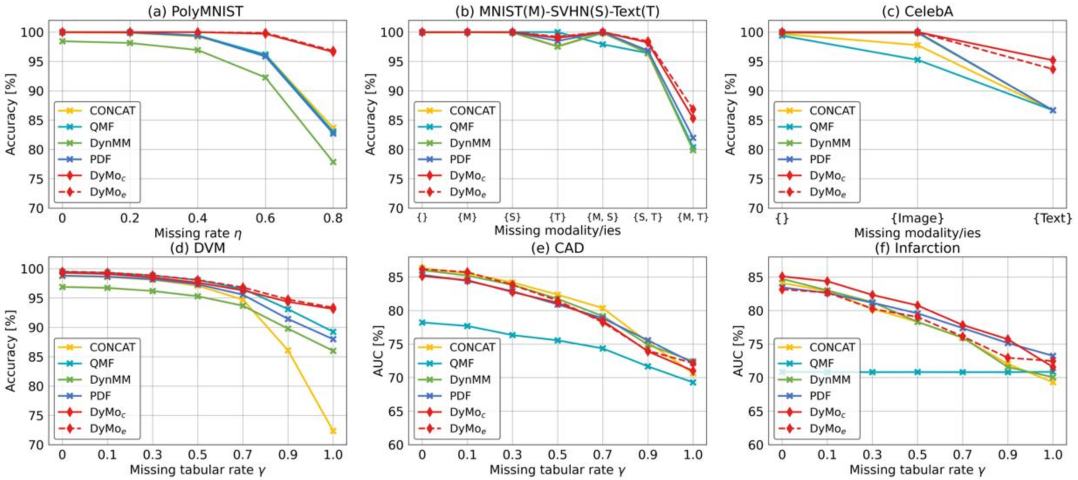
$$\mathcal{L}_{\text{aux}} = -\frac{1}{A}\frac{1}{B}\sum_{\mathcal{S}\sim\mathcal{U}_A}\sum_{i=1}^{B}\log\frac{\exp\left(-d_\phi(\boldsymbol{z}_i, \boldsymbol{c}_{y_i})/t\right)}{\sum_{k'=1}^{K}\exp\left(-d_\phi(\boldsymbol{z}_i, \boldsymbol{c}_{k'})/t\right)},$$

- Overall loss

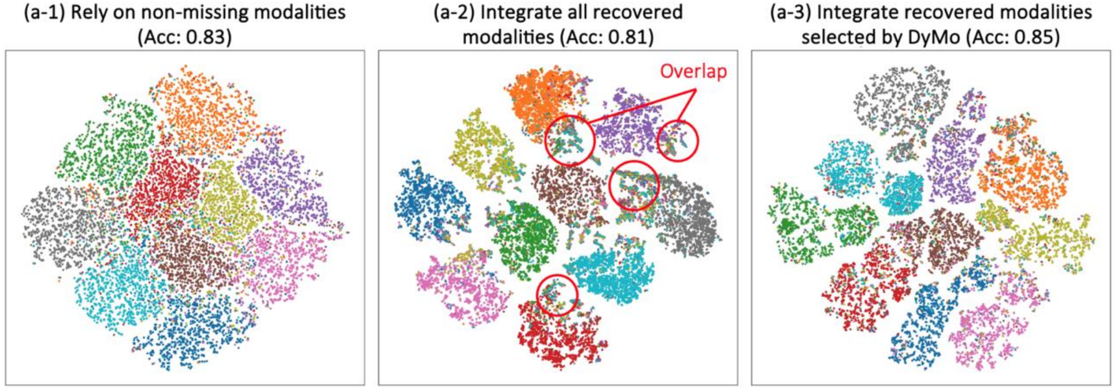
$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{aux}}.$$

Experiments

- +1.61% on PolyMNIST, +1.68% on MST, and +3.88% on CelebA

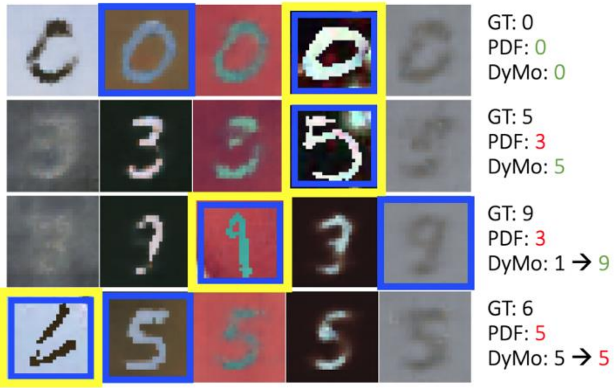


Model	PolyMNIST Acc (%) ↑			MST Acc (%) ↑			CelebA Acc (%) ↑			DVM Acc (%) ↑			CAD AUC (%) ↑			Infarction AUC (%) ↑		
	Missing Rate η			Missing Modality/ies						Missing Tabular Rate γ								
	0	0.6	0.8	{}	{S,T}	{M,T}	{}	{I}	{T}	0	0.7	1	0	0.7	1	0	0.7	1
(a) Recovery-based Methods for Missing Modality																		
MultiAE	99.77	95.36	84.39	99.96	97.00	81.60	89.98	15.51	89.71	-	-	-	-	-	-	-	-	-
MultiAE†	99.94	97.50	89.86	99.87	98.33	83.44	88.66	72.04	87.44	-	-	-	-	-	-	-	-	-
MoPoE	99.79	93.93	79.84	99.62	90.86	79.01	38.97	13.91	34.84	-	-	-	-	-	-	-	-	-
MoPoE†	99.63	96.81	87.06	99.39	96.50	82.54	68.22	56.90	65.75	-	-	-	-	-	-	-	-	-
M3Care	99.93	56.66	40.53	99.99	16.03	9.34	92.33	99.92	51.75	98.44	-	11.92	85.62	-	64.99	70.61	-	70.53
M3Care†	99.99	97.27	87.92	99.98	98.27	85.16	98.73	97.14	91.32	98.94	-	**93.43**	72.48	-	**72.48**	83.27	-	68.44
OnlineMAE	**100**	98.29	90.09	99.90	98.14	84.14	86.67	86.67	86.67	90.92	-	89.90	85.22	-	70.96	84.05	-	61.39
(b) Recovery-free Methods for Missing Modality																		
ModDrop	99.97	97.66	88.44	**100**	98.21	82.47	99.93	99.93	87.32	99.02	93.23	87.97	85.10	76.65	69.18	84.76	74.64	72.16
MTL	99.97	98.43	91.14	99.96	**98.60**	84.37	99.69	99.26	89.38	99.44	95.53	92.32	84.87	77.72	70.23	83.59	75.73	69.90
MAP	99.86	43.00	23.19	**100**	9.83	10.13	99.98	99.93	85.33	98.86	86.96	63.15	84.39	75.61	70.11	84.62	71.73	69.17
MAP†	99.99	96.74	76.20	99.99	97.84	11.36	**100**	99.99	86.06	99.37	94.83	91.15	85.26	75.84	68.76	**85.49**	75.88	70.81
MUSE	99.93	94.73	77.56	99.86	97.14	35.96	99.93	99.86	88.35	96.86	-	1.64	83.47	-	53.23	84.40	-	66.78
(c) Dynamic Recovery Method for Missing Modality																		
DyMo$_c$	**100**	99.71	96.61	**100**	98.22	85.31	**100**	**100**	**95.20**	99.30	96.31	93.14	85.14	**78.49**	71.02	85.10	**77.85**	71.58
DyMo$_e$	**100**	**99.87**	**96.81**	**100**	98.43	**86.84**	**100**	**100**	93.67	**99.50**	**96.81**	93.36	**86.17**	78.24	72.17	83.16	76.14	**72.47**



(a) t-SNE visualization on the MST test dataset with missing {M,T}

(b) Case study on PolyMNIST



IMPERIAL

Inference-Time Dynamic Modality Selection for Incomplete Multimodal Classification

Paper is available at

Code is available at

Thank you!

Contact us: {s.du23, x.luo, declan.oregan, c.qin15}@imperial.ac.uk