中南大學
CENTRAL SOUTH UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

山东师范大学
SHANDONG NORMAL UNIVERSITY

Disentangled Cross-Modal Representation Learning with Enhanced Mutual Supervision

Lu Gao[1*], Wenlan Chen[1*], Daoyuan Wang[1], Fei Guo[1†], Cheng Liang[2†]

[1]School of Computer Science and Engineering, Central South University

[2]School of Information Science and Engineering, Shandong Normal University

CONTENT

01 Research Background

02 Methods

03 Experiments

04 Conclusion

CONTENT

01 Research Background

中南大學
CENTRAL SOUTH UNIVERSITY

Research Background

Audio data

Video data

Image data

After surviving the heavenly tribulation, Nezha and Ao Bing face new trials.

Text data

Information coexists across multiple views or modalities

Goal

- Extract semantically aligned representations from heterogeneous modalities such as images and text.

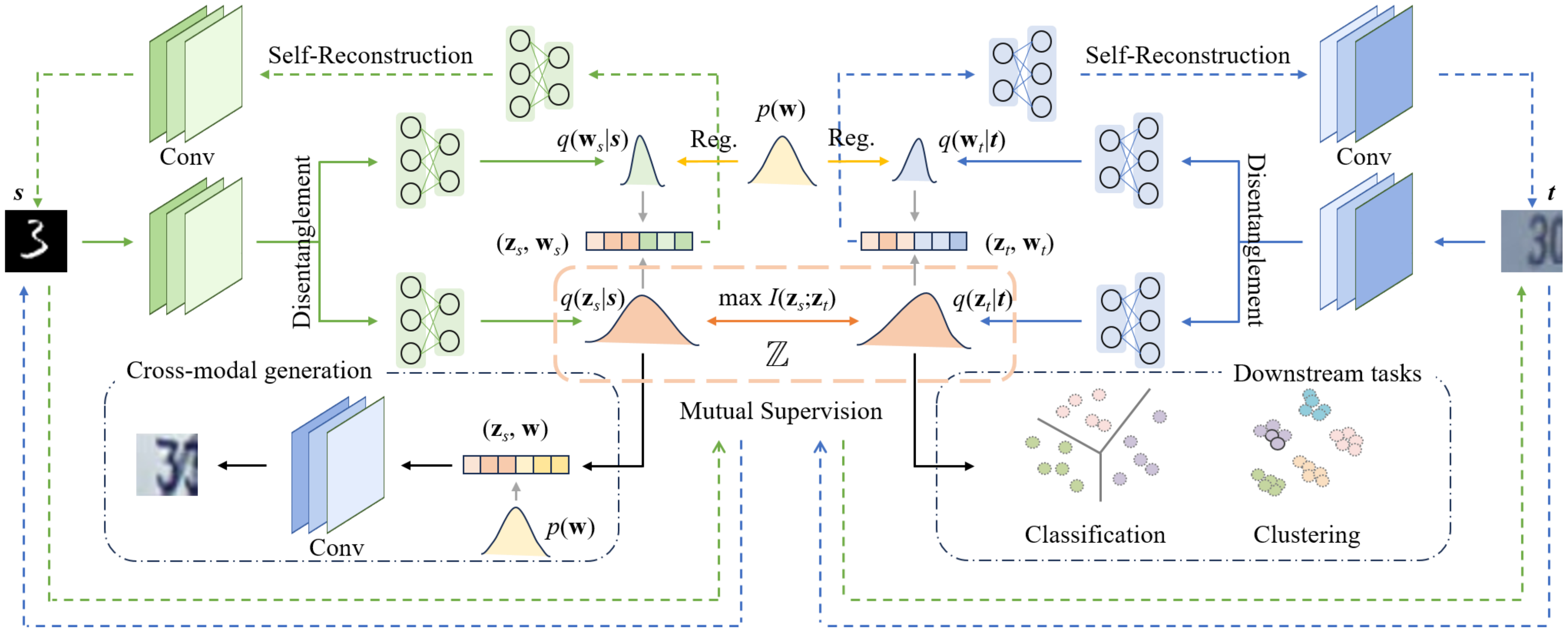
Challenges

Existing multimodal VAE-based models often struggle to effectively align heterogeneous modalities.

Many of these models also lack sufficient structural constraints to clearly separate the modality-specific and shared factors.

02 Methods

Methods



DCMEM

- Disentangle shared and modality-specific information, and enforce consistency through mutual supervision.
- Leverage the information bottleneck principle to promote semantic alignment and information complementarity across different modalities.

Methods

Paired Setting

- Mutual Supervision $\boldsymbol{s} \leftrightarrow \boldsymbol{z} \leftrightarrow \boldsymbol{t}$
- Generative Process $p_{\theta,\psi_z}(\boldsymbol{t},\boldsymbol{s},\boldsymbol{z}_s,\boldsymbol{w}_s) = p(\boldsymbol{t})\, p_{\psi_z}(\boldsymbol{z}_s \mid \boldsymbol{t})\, p(\boldsymbol{w}_s)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s)$
- Inference Process $q_{\phi,\varphi}(\boldsymbol{z}_s,\boldsymbol{w}_s,\boldsymbol{t} \mid \boldsymbol{s}) = q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_{\phi_w}(\boldsymbol{w}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)$

ELBO Variational Framework

$$\log p_{\theta,\psi_z}(\boldsymbol{t},\boldsymbol{s}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}_s,\boldsymbol{w}_s|\boldsymbol{s})}\left[\frac{q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)}{q_{\phi_z,\varphi}(\boldsymbol{t} \mid \boldsymbol{s})} \log \frac{p_{\psi_z}(\boldsymbol{z}_s \mid \boldsymbol{t})\, p(\boldsymbol{w}_s)\, p_\theta(\boldsymbol{s} \mid \boldsymbol{z}_s,\boldsymbol{w}_s)}{q_{\phi_z}(\boldsymbol{z}_s \mid \boldsymbol{s})\, q_{\phi_w}(\boldsymbol{w}_s \mid \boldsymbol{s})\, q_\varphi(\boldsymbol{t} \mid \boldsymbol{z}_s)}\right] + \log q_{\phi_z,\varphi}(\boldsymbol{t}|\boldsymbol{s}) + \log p(\boldsymbol{t}).$$

Shared Representation

Prior

Decoder

Cross-modal Reconstruction

Encoder

Variable Independence

Structured Representation Learning

$$\max I(\boldsymbol{z}_s;\boldsymbol{t};\boldsymbol{s}) - I(\boldsymbol{z}_s;\boldsymbol{w}_s)$$

Shared Information

Redundant Information

Paired Loss

$$\mathcal{L}_{Bi}(\boldsymbol{s},\boldsymbol{t})$$

Shared Representations Alignment

$$\max I(\boldsymbol{z}_s;\boldsymbol{z}_t)$$

Methods

Partially Missing Setting

- Mutual Supervision $\boldsymbol{s} \leftrightarrow \boldsymbol{z} \leftrightarrow t$
- Generative Process $p_{\theta,\psi_z}\left(\boldsymbol{s}, \boldsymbol{z}_s, \boldsymbol{w}_s\right) = p_{u^t}\left(\boldsymbol{z}_s\right) p\left(\boldsymbol{w}_s\right) p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right)$
- Inference Process $q_{\boldsymbol{\phi}}(\boldsymbol{z}_s, \boldsymbol{w}_s | \boldsymbol{s}) = q_{\phi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{s}\right) q_{\phi_w}\left(\boldsymbol{w}_s \mid \boldsymbol{s}\right)$

ELBO Variational Framework

Decoder Prior Anchor Prior

$$\log p_{\theta,\psi_z}\left(\boldsymbol{s}\right) \geq \mathbb{E}_{q_{\phi}(\boldsymbol{z}_s, \boldsymbol{w}_s | \boldsymbol{s})} \log \frac{p_\theta\left(\boldsymbol{s} \mid \boldsymbol{z}_s, \boldsymbol{w}_s\right) p\left(\boldsymbol{w}_s\right) p_{u^t}\left(\boldsymbol{z}_s\right)}{q_{\phi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{s}\right) q_{\phi_w}\left(\boldsymbol{w}_s \mid \boldsymbol{s}\right)},$$

Encoder Variable Independence

VampPrior Anchor Prior

- Integral Form

$$p_{u^t}\left(\boldsymbol{z}_s\right) = \int p\left(\boldsymbol{t}\right) p_{\psi_z}\left(\boldsymbol{z}_s \mid \boldsymbol{t}\right) dt$$

- Monte Carlo Estimation

$$p_{u^t}\left(\boldsymbol{z}_s\right) = \frac{1}{B} \sum_{i=1}^{B} p_{\psi_z}\left(\boldsymbol{z}_s | \boldsymbol{u}_i^t\right)$$

Missing Loss

$$\mathcal{L}_s\left(\boldsymbol{s}\right)$$

Paired Loss

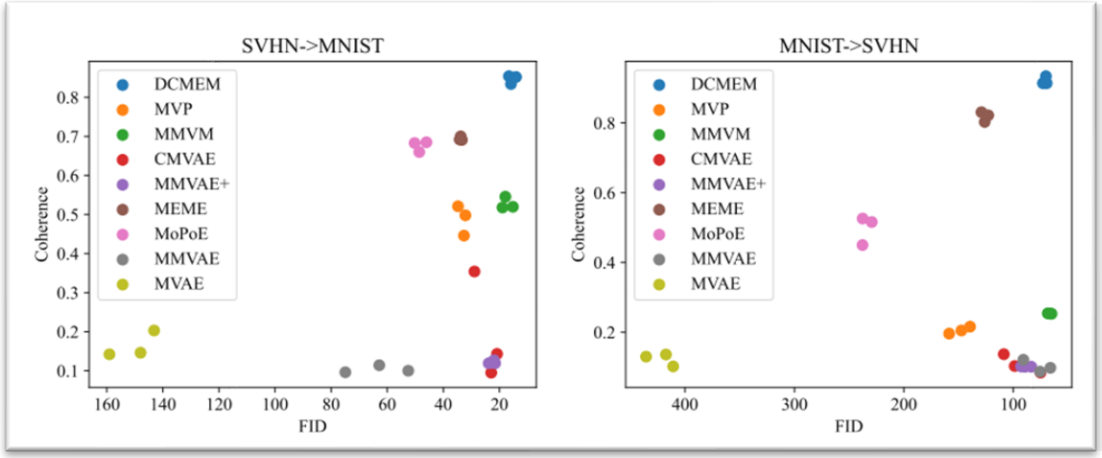
$$\mathcal{L}_{Bi}\left(\boldsymbol{s}, \boldsymbol{t}\right)$$

Total Loss

$$\mathcal{L}\left(\mathcal{D}\right) = \sum_{s,t \in \mathcal{D}_{s,t}} \mathcal{L}_{Bi}\left(\boldsymbol{s}, \boldsymbol{t}\right) + \sum_{s \in \mathcal{D}_s} \mathcal{L}_s\left(\boldsymbol{s}\right) + \sum_{t \in \mathcal{D}_t} \mathcal{L}_t\left(\boldsymbol{t}\right).$$

03 Experiments

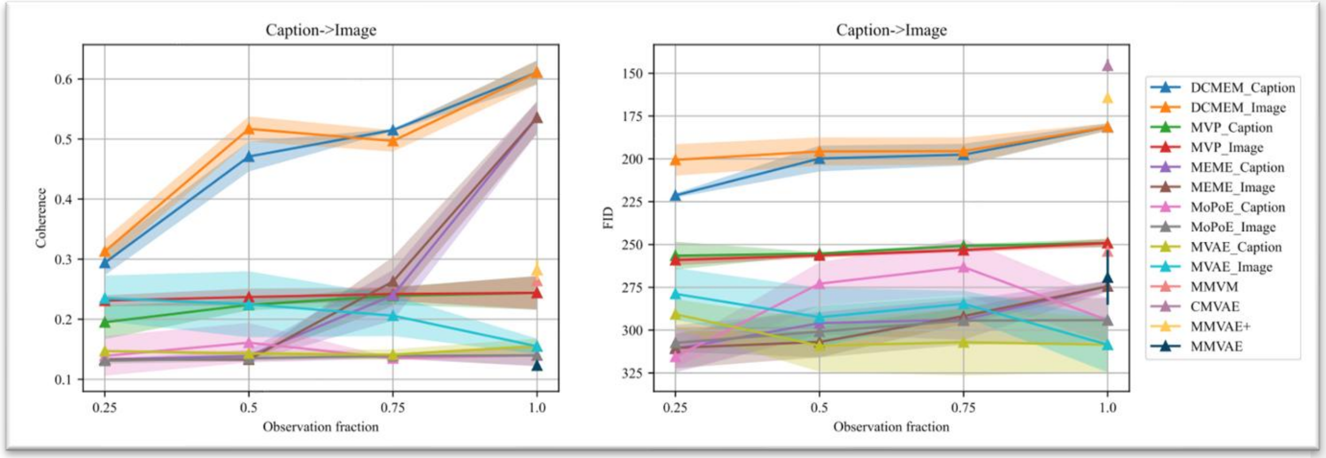
Generate performance



Paired Setting (MNIST–SVHN)

- **Task:** Cross-modal digit generation
- **Metrics:** FID (generation quality) and Coherence (generation consistency)

- DCMEM achieves higher sample quality and better cross-modal consistency.

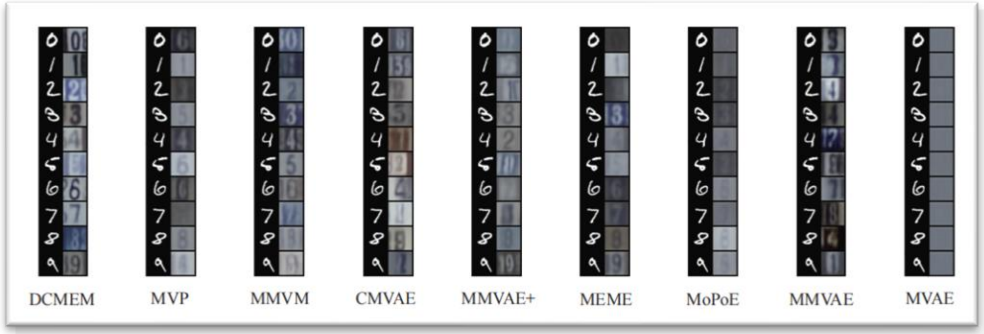


Partially Missing Setting (CUBICC)

- **Task:** Cross-modal image generation
- **Metrics:** FID and Coherence with varying proportions of observed paired data

- Achieves the best generation quality and coherence under all observation ratios.
- Maintains strong cross-modal alignment with limited paired data.

Generate performance

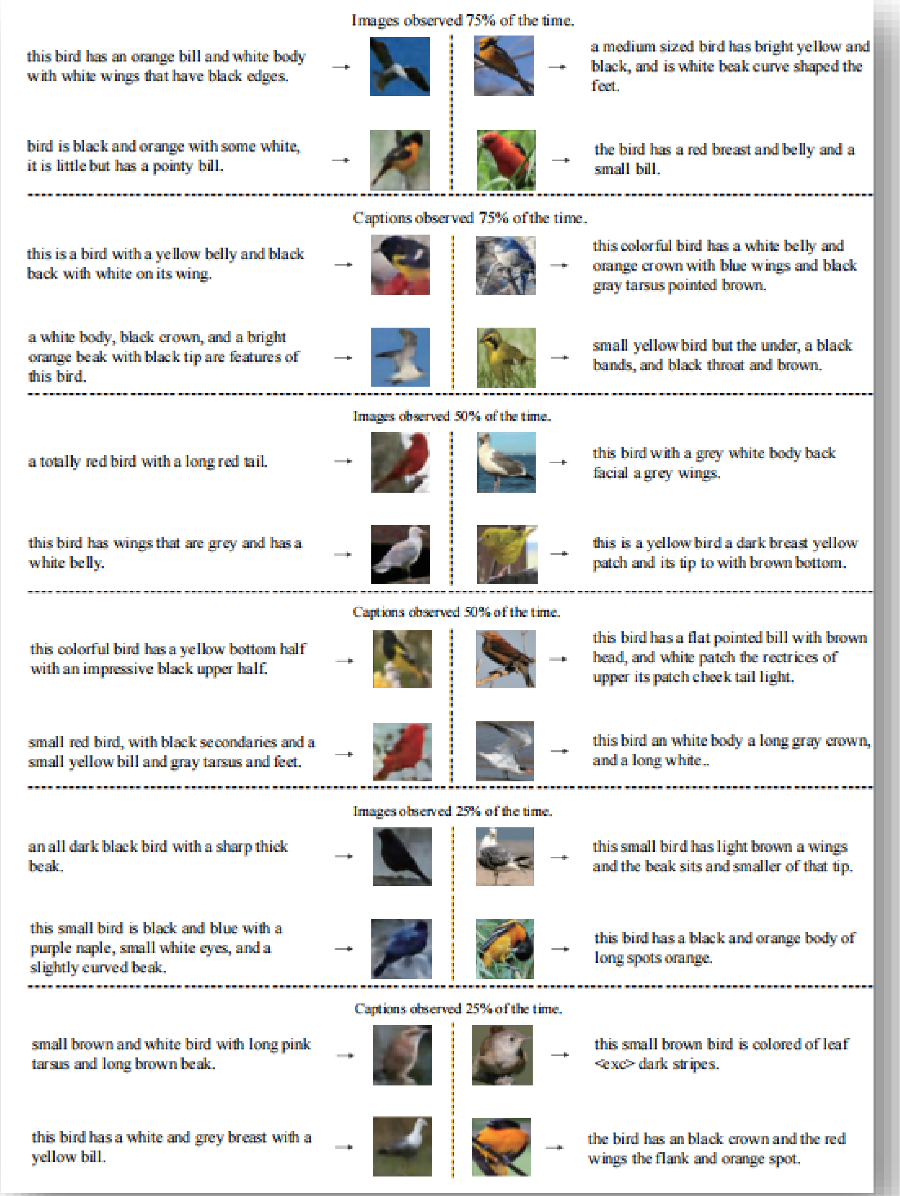


Qualitative Evaluation

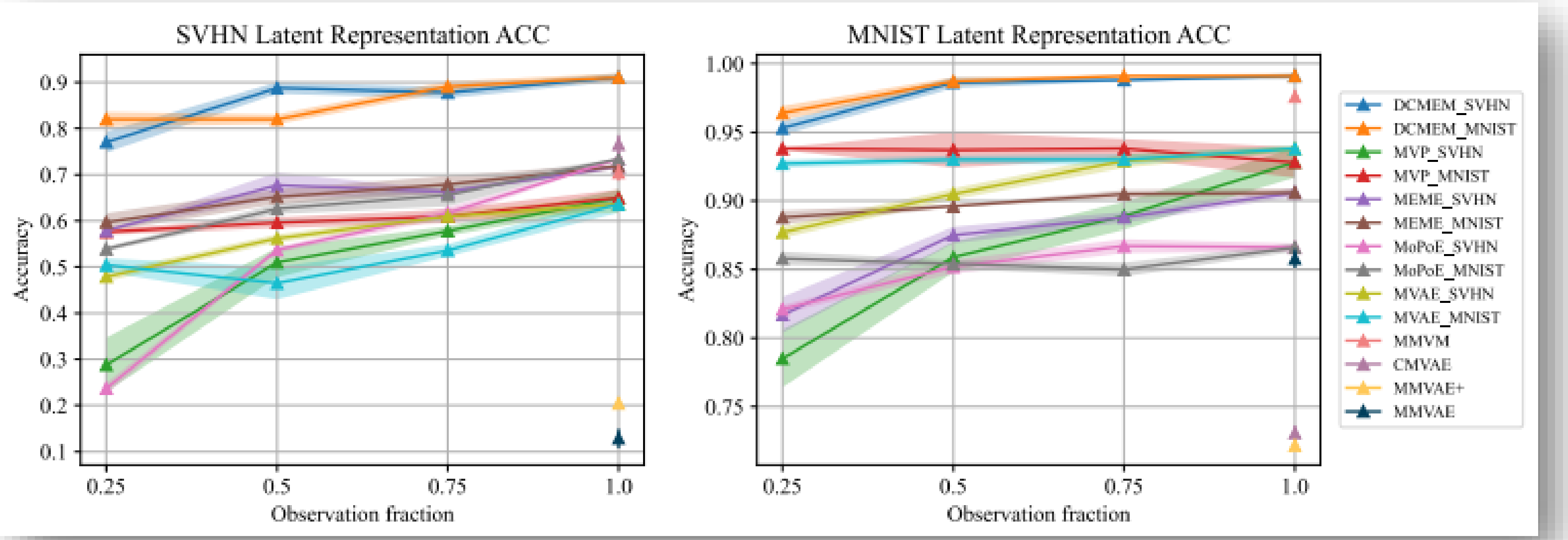
- DCMEM produces visually clear and semantically accurate results that align well with the input modality.
- Compared to baseline methods, it better preserves category information and avoids confusion or blurring in generated samples.

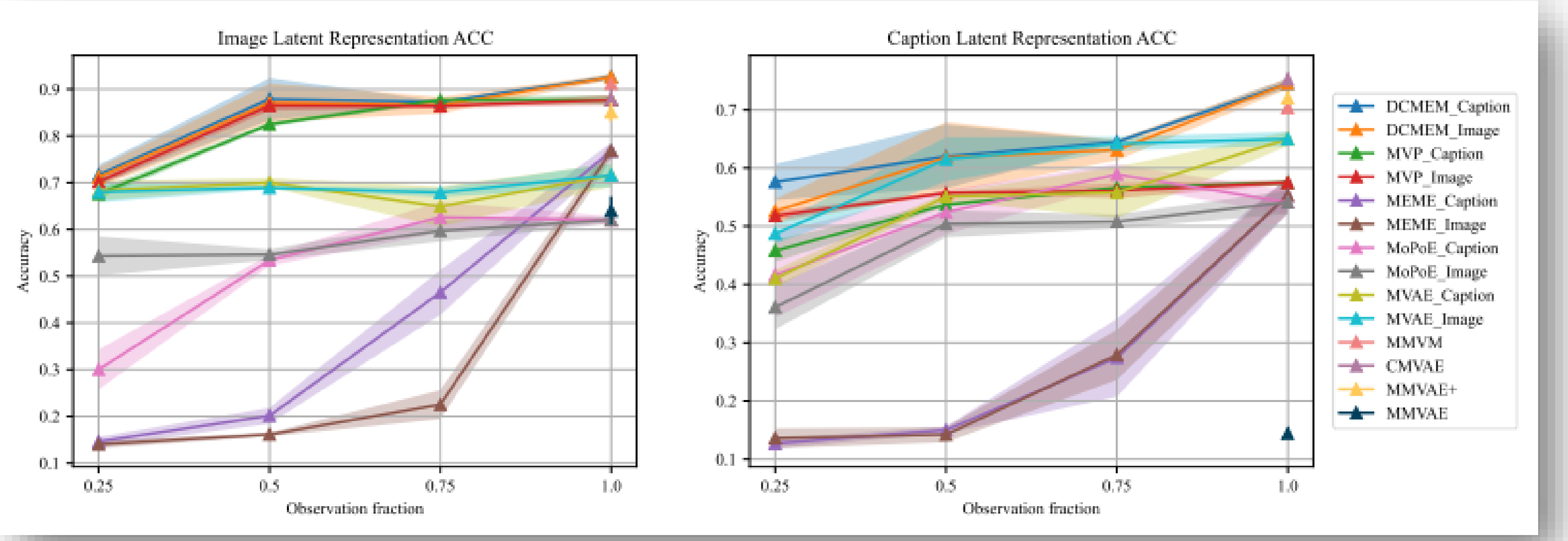
Diversity Evaluation

- DCMEM maintains semantic consistency while generating diverse outputs within each category.
- In contrast, MMVAE+ and CMVAE show coupled category and modality-specific variables, and MVP exhibits limited variability.



Classification performance





Result Analysis

- DCMEM: Achieves the best accuracy across all observation ratios, showing strong robustness and stability.
- MMVAE+: Performs poorly on MNIST-SVHN, as its shared latent space fails to capture discriminative features.
- MEME / MVP: Designed for incomplete modalities but accuracy drops sharply with higher missing rates.

Clustering performance

- DCMEM achieves consistently superior performance across all three types of representations.
- Traditional multimodal VAEs (e.g., MVAE, MMVAE, MoPoE) show clear limitations in cross-modal representation alignment.

Methods	SVHN Representation			MNIST Representation			Joint Representation		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MVAE	27.9	16.0	13.1	79.2	65.5	62.6	42.7	35.3	24.5
MMVAE	22.0	10.4	10.1	21.8	10.3	10.1	22.6	10.7	10.1
MoPoE	37.9	27.2	18.5	50.5	45.6	33.0	64.1	60.5	50.7
MEME	21.9	10.3	10.0	36.5	32.1	20.4	22.4	10.6	10.1
MMVAE+	23.9	11.4	11.1	21.3	10.4	10.0	22.9	11.9	10.8
CMVAE	42.2	36.3	25.4	28.1	15.9	14.5	32.3	19.5	15.4
MMVM	42.2	27.1	20.7	88.1	82.1	80.4	77.5	72.2	67.5
MVP	53.6	38.7	30.1	81.4	79.6	73.6	84.8	76.4	70.6
DCMEM	**91.5**	**80.6**	**82.0**	**99.1**	**97.3**	**98.0**	**99.5**	**98.4**	**98.9**

- DCMEM achieves the best performance on most metrics, showing strong feature alignment and discriminative capability.

Methods	Image Representation			Caption Representation			Joint Representation		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MVAE	26.2	12.4	7.5	18.1	2.4	0.9	38.7	26.8	18.0
MMVAE	23.1	12.1	6.1	14.5	1.3	0.1	15.8	1.5	0.2
MoPoE	33.4	17.6	11.5	43.5	27.1	19.9	40.8	30.4	20.2
MEME	44.8	43.4	28.4	36.3	29.5	18.6	19.8	4.8	2.1
MMVAE+	27.7	11.9	7.1	48.7	36.4	26.8	64.4	52.6	44.1
CMVAE	67.7	58.3	47.4	65.1	**53.3**	42.7	73.7	67.4	57.2
MMVM	58.9	56.9	44.5	23.9	9.4	5.4	66.8	67.0	55.5
MVP	64.1	53.8	41.8	48.5	34.4	26.1	61.1	55.6	44.0
DCMEM	**86.9**	**77.4**	**72.4**	**69.7**	52.2	**44.2**	**86.3**	**76.8**	**71.5**

Representation Analysis

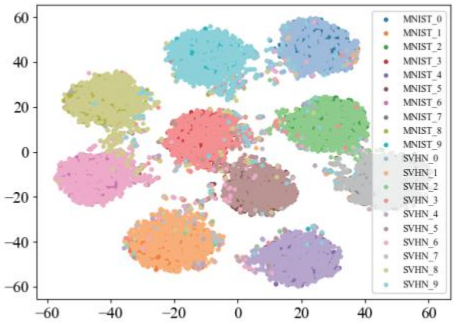
Class-level Semantic Alignment Analysis

Objective: To examine whether different modalities maintain consistent semantic representations at the class level.
Method: Compute the distance matrix K_{ij} conditioned on class, where the diagonal entries represent intra-class distances and the off-diagonal entries represent inter-class distances.

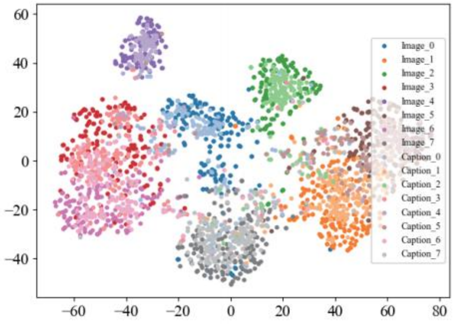
Sample-level Semantic Alignment Analysis

Objective: To determine whether semantically related cross-modal samples have closer latent distributions.
Method: Compute the 2-Wasserstein distance between all sample pairs and use histograms to distinguish paired from unpaired samples.

Shared Representation t-SNE Visualization

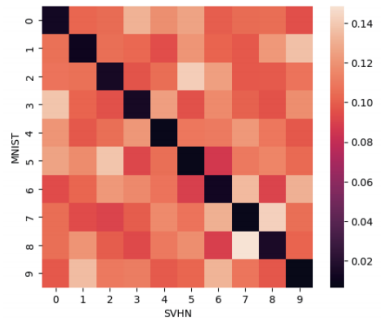


(a) MNIST-SVHN

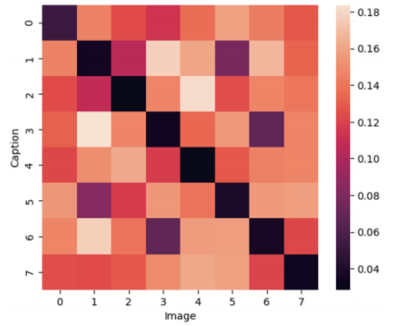


(b) CUBICC

Class-level Semantic Alignment Analysis

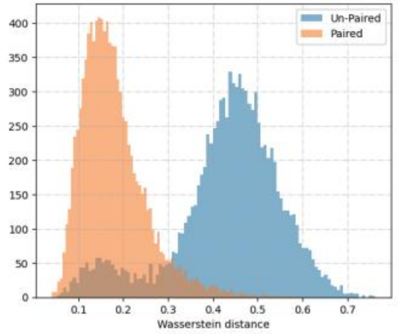


(a) MNIST-SVHN

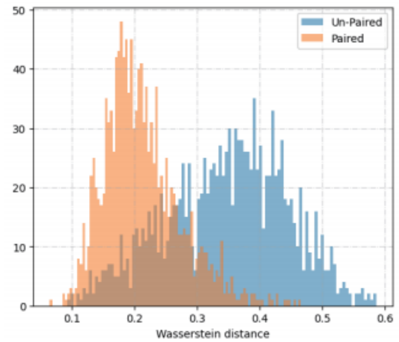


(b) CUBICC

Sample-level Semantic Alignment Analysis



(a) MNIST-SVHN



(b) CUBICC

04 Conclusion

Conclusion

Contributions

01
- DCMEM separates shared and modality-specific information and enforces consistency through mutual supervision.

02
- DCMEM leverages the information bottleneck principle to promote compact and complementary feature encoding.

03
- DCMEM supports learning from both complete and partially missing multimodal data through a valid variational objective.

04
- Experimental results show superior performance on diverse tasks, including cross-modal generation, clustering, and classification.

中南大學
CENTRAL SOUTH UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

山东师範大学
SHANDONG NORMAL UNIVERSITY

Thanks!

Disentangled Cross-Modal Representation Learning with Enhanced Mutual Supervision