Deep Incomplete Multi-view Learning via Cyclic Permutation of VAEs

Seeking PhD 2026 !
Research Interests:
- Generative Modeling
- Multimodal Learning

WeChat

Xin Gao, Jian Pu

gaoxin23@m.fudan.edu.cn / jianpu@fudan.edu.cn

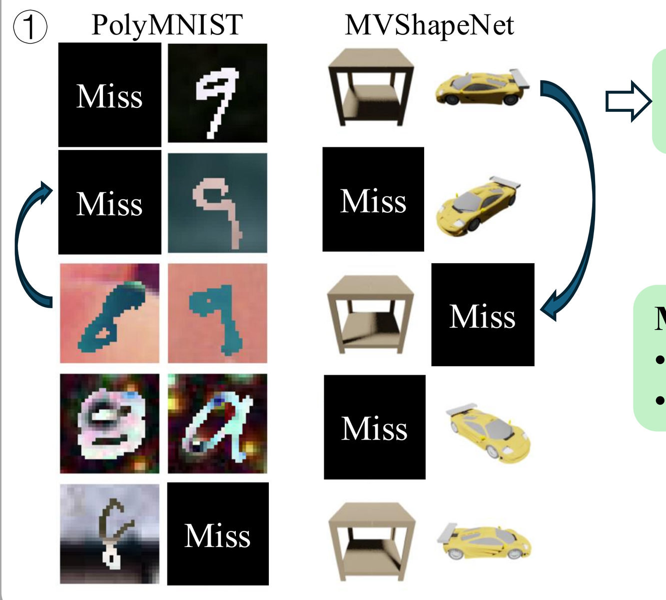
Paper Link

复旦大学

ICLR International Conference On Learning Representations

Motivation

Incomplete multi-view data $\{\mathbb{X}_i\}_{i=1}^{n}$, where each $\mathbb{X}_i = \{x^{(v)}\}_{v\in\mathcal{I}_i}$



Invariant relationships between views ⇨ **Inter-view Correspondences**

⇩ Inferring missing information

Multimodal VAEs
- Distributional Modeling,
- Robustness, …

✚ Convergence of multi-view representations

⇩ 💡**Multimodal VAEs with Inter-view Correspondences in the Latent Space**

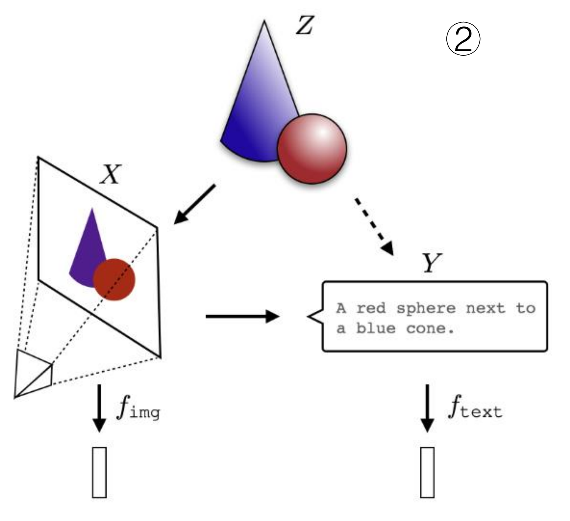


Figure Source: Huh et al., ICML 2024

Method Part (1): Variational Posterior of Incomplete Multi-view Learning for More Complete Information

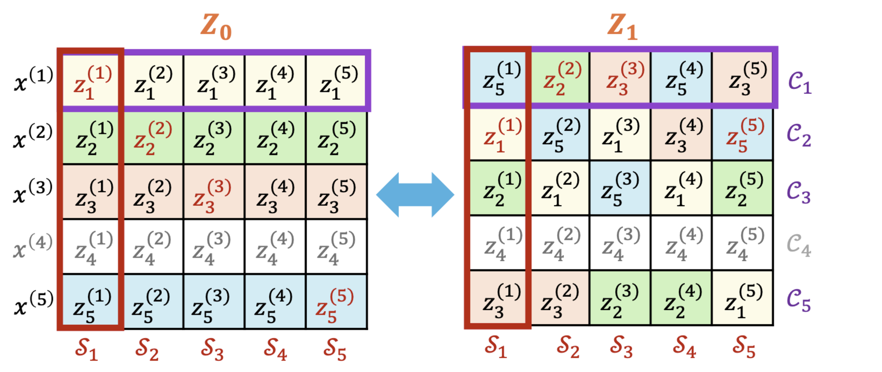
$Z_v^{(l)}$ - Superscript: target view - Subscript: source view

- **Inter-view correspondences:**

  $f_{lv}(\cdot;\ \alpha_{lv})$ from v-th view to l-th view, $\forall\ l \neq v$

- **Multi-View latent variable set:** $\mathcal{Z} = \{z_v^{(l)}\}_{(v,l)\in\mathcal{I}\times[L]}$

- **Operations of a set:**



① **Permutation**: Variables with the same superscript l should represent similar information about the l-th view, regardless of how they are derived.

$\Longrightarrow$ Permutation-invariance

② **Partition**: Group variables by views
- **Single-view Partition (Columns)** $\xrightarrow{\text{Intra-Consistency}}$ Better establishment of Inter-view correspondences
- **Complete-view Partition (Rows)** $\xrightarrow{\text{More Complete Information}}$ Consensus latent variable ω

I. Joint posterior

$$q(\mathcal{Z},\boldsymbol{\Omega} \mid \{x^{(v)}\}_{\mathcal{I}}) \triangleq \prod_{n\in\mathcal{I}} q(\boldsymbol{\omega}_n \mid \mathcal{C}_n, \{x^{(v)}\}_{\mathcal{J}_n}) q(\mathcal{C}_n \mid \{x^{(v)}\}_{\mathcal{J}_n})$$
$$= \prod_{n\in\mathcal{I}} \left[q(\boldsymbol{\omega}_n \mid \mathcal{C}_n, \{x^{(v)}\}_{\mathcal{J}_n}) \prod_{l=1, v\in\mathcal{J}_n}^{L} q(\boldsymbol{z}_v^{(l)} \mid x^{(v)}) \right].$$

II. Generative process: Each view $x^{(n)}$ is reconstructed using the latent variable $z_*^{(n)}$ (represent the n-th view) and a consensus variable ω_n

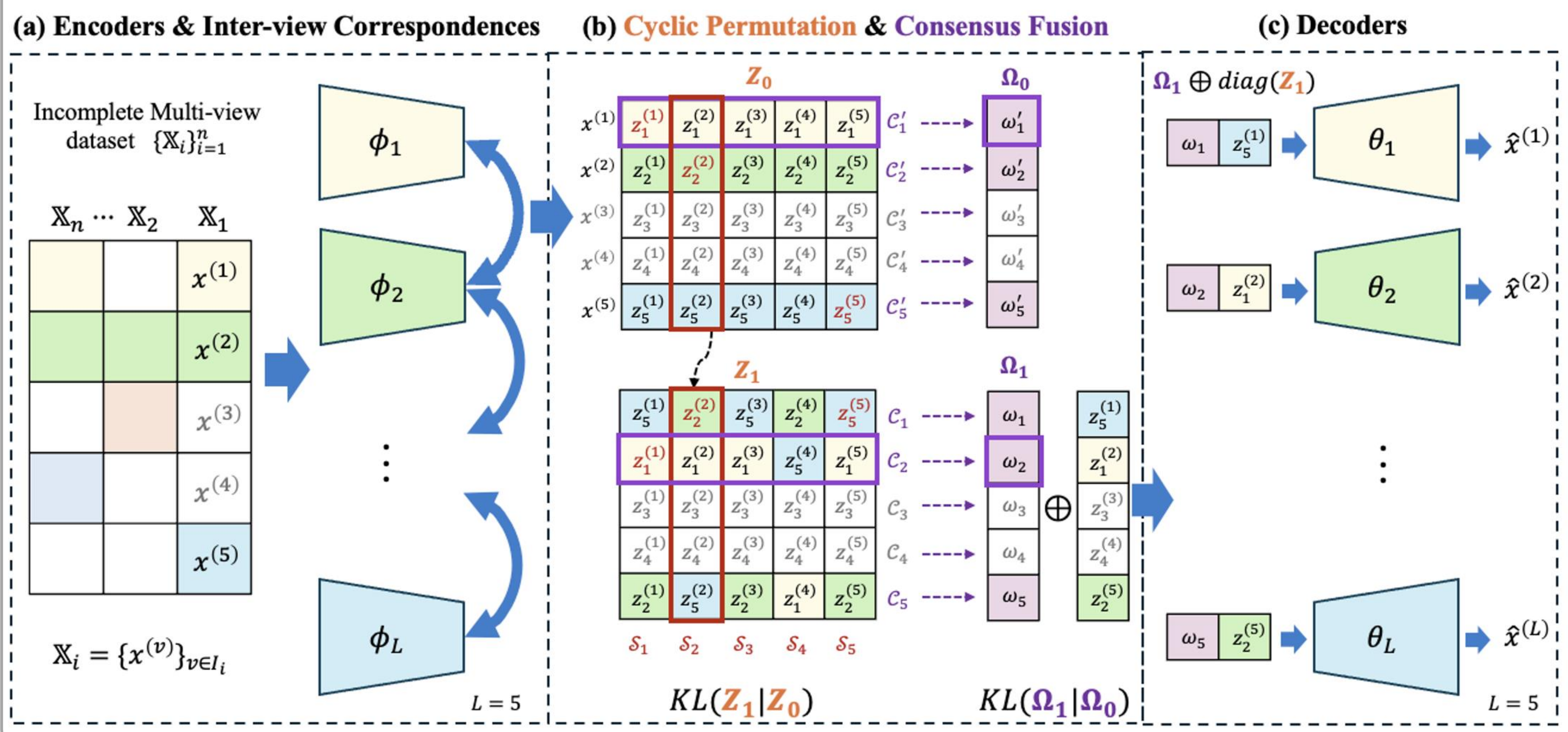
$$p(x^{(n)} | \mathcal{C}_n, \omega_n) = p\left(x^{(n)} | \mathcal{C}_n \cap \mathcal{S}_n, \omega_n; \theta_n\right)$$

III. ELBO: Reconstruction loss and two regularization terms for the latent space

$$\mathcal{L}_{\text{ELBO}}(\{x^{(v)}\}_{\mathcal{I}}) = \sum_{n\in\mathcal{I}} \mathbb{E}_{q(\mathcal{C}_n,\omega_n | \{x^{(v)}\}_{\mathcal{J}_n})} \left[\log p(x^{(n)} \mid \mathcal{C}_n \cap \mathcal{S}_n, \omega_n)\right]$$
$$- \sum_{l=1}^{L} \sum_{v\in\mathcal{I}} KL\left[q(z_v^{(l)} \mid x^{(v)}) \,\|\, \boxed{p(z_v^{(l)})}\right] - \sum_{n\in\mathcal{I}} KL\left[q(\omega_n \mid \mathcal{C}_n, \{x^{(v)}\}_{\mathcal{J}_n}) \,\|\, \boxed{p(\omega_n)}\right].$$

Method: Overview

💡**Multimodal VAEs with Inter-view Correspondences**



Method Part (2): A New Informational Prior for Better Consistency

Cyclic Permutation

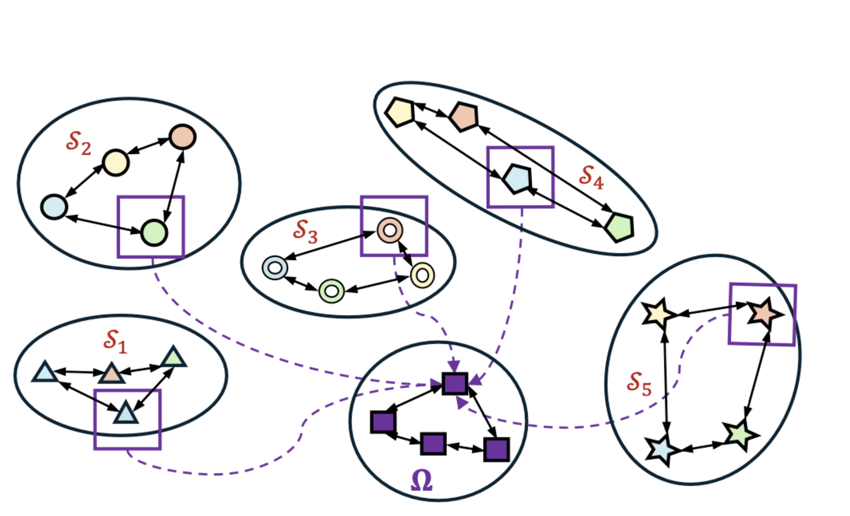
- A cyclic permutation is a way of rotating the elements in a set so that they all shift positions in a cycle and return to their original spots after a full round.

1 2 3
3 1 2
1=>3=>2=>1

$$KL[P_1||P_3] + KL[P_2||P_1] + KL[P_3||P_2]$$
$$P_1 = P_3 = P_2$$

Definition 4 (Permutation Divergence). *Let $N \geq 2$ be a fixed natural number. Given a cyclic permutation σ on the index set $[N]$ and a set $\mathcal{P}$ of probability distributions on the same measure space, the Permutation Divergence of order N is a mapping d from $\mathcal{P}^N$ to the extended real line $\mathbb{R}\cup\{+\infty\}$, defined as follows for any $P_1, P_2, \ldots, P_N \in \mathcal{P}$:*

$$d(P_1, P_2, \ldots, P_N; \sigma) = \sum_{i=1}^{N} KL[P_i \,\|\, P_{\sigma(i)}].$$



Inter-View Translatability

$$\sum_{l=1}^{L}\sum_{v\in\mathcal{I}} KL\left[q_v^{(l)}(z) \,\|\, q_{\sigma_l^{-1}(v)}^{(l)}(z)\right] = \sum_{l=1}^{L} d(q_{k_1}^{(l)}, \ldots, q_{k_n}^{(l)}; \sigma_l^{-1}).$$

Consensus Concentration

$$\sum_{n\in\mathcal{I}} KL\left(q(\omega \mid \mathcal{C}_n) \,\|\, q(\omega \mid \mathcal{C}_n^0)\right)$$

Experiments

Three datasets are selected for displaying 😬

- **Clustering**
- Compared with eight incomplete multi-view learning methods.
- Missing rate $\eta = \{0.1, 0.3, 0.5, 0.7\}$
- Our method consistently achieved the best (**bold red**) or second-best (**bold blue**) performance across different missing ratios and datasets

Table 3: Complete clustering results of nine methods on five multi-view datasets with missing rates of $\eta = 0.1, 0.3, 0.5$, and 0.7. The first and second best results are indicated in **bold red** and **blue**, respectively. Each experiment was run five times using different random seeds.

	Missing rate	0.1			0.3			0.5			0.7		
	Metrics	ACC↑	NMI↑	ARI↑	ACC↑	NMI↑	ARI↑	ACC↑	NMI↑	ARI↑	ACC↑	NMI↑	ARI↑
Handwritten	DCCA	73.83±0.66	71.17±1.92	55.13±2.01	68.02±5.87	64.93±6.40	44.76±11.45	63.25±8.30	60.27±8.42	38.93±12.49	59.07±10.23	56.44±9.86	33.58±14.25
	DCCAE	73.61±0.65	72.22±1.62	54.61±2.49	68.12±5.53	65.76±6.62	43.42±12.07	63.36±8.13	60.72±8.97	37.32±13.48	59.31±9.94	56.89±10.22	32.55±14.33
	DIMVC	**89.13±1.11**	80.06±1.36	77.96±2.06	85.24±1.20	74.67±1.47	70.83±2.10	82.76±0.71	71.17±1.06	66.81±1.19	79.66±4.28	68.94±2.20	63.12±4.17
	DSIMVC	81.27±1.48	79.47±1.30	71.59±1.92	81.82±3.27	80.27±2.95	73.36±3.40	81.39±2.83	79.23±2.47	71.88±3.94	77.38±4.35	74.80±3.12	66.84±4.35
	Completer	82.18±2.06	77.73±1.03	68.92±1.98	78.70±4.22	69.08±4.17	58.14±4.50	74.73±4.51	67.49±4.05	50.21±6.37	68.86±2.41	62.41±1.61	41.06±1.88
	CPSPAN	80.30±7.13	78.43±3.90	71.84±6.71	79.80±6.76	79.08±3.61	72.28±6.09	84.64±5.47	80.23±3.10	75.34±5.62	**83.90±5.26**	80.60±1.98	**75.25±4.26**
	ICMVC	88.86±5.39	82.19±3.94	79.51±6.58	80.95±2.75	74.53±1.15	68.15±1.81	73.83±0.46	67.95±0.23	58.72±0.44	67.48±3.83	61.99±2.97	48.66±3.19
	DVIMC	87.89±3.28	**83.51±1.98**	**79.66±2.81**	**85.36±5.55**	**82.82±3.82**	**78.50±5.83**	**85.01±4.49**	**82.96±2.41**	**78.50±4.06**	81.66±5.92	**80.78±1.84**	74.85±4.66
	MVP(Ours)	**90.55±5.29**	**87.08±2.48**	**84.46±5.16**	**88.69±5.89**	**84.86±2.73**	**81.59±5.77**	**90.76±4.69**	**85.44±1.55**	**83.53±3.69**	**86.74±5.15**	**82.56±2.31**	**78.75±4.77**
CUB	DCCA	58.93±3.21	59.59±1.08	40.87±1.23	55.65±4.49	54.23±5.71	35.63±5.86	48.60±10.85	46.71±11.92	27.88±12.17	42.35±14.35	39.42±16.37	21.77±14.94
	DCCAE	57.27±3.22	61.04±2.59	43.09±3.32	53.43±4.53	54.59±6.82	36.48±7.11	47.21±9.68	47.05±12.10	28.37±12.93	41.52±13.12	40.50±15.76	22.66±15.03
	DIMVC	66.03±5.04	61.70±4.06	48.96±5.55	57.20±3.98	56.00±2.89	41.25±3.74	60.65±3.62	55.75±3.62	42.81±4.77	56.08±1.16	51.07±1.54	36.65±2.23
	DSIMVC	**72.93±5.28**	**67.82±3.07**	**55.89±4.56**	**66.83±4.74**	61.78±2.21	47.71±3.84	**68.37±5.67**	61.55±3.00	**48.21±4.70**	**67.33±2.29**	59.89±0.95	**46.31±1.97**
	Completer	52.97±4.56	65.47±2.41	45.98±4.45	60.73±2.49	**68.88±1.94**	**52.78±2.71**	51.90±2.47	61.84±0.89	45.11±1.28	19.43±1.15	17.37±1.23	0.73±0.31
	CPSPAN	58.77±3.12	62.27±2.18	45.35±2.40	61.30±1.41	64.21±2.32	48.93±2.94	60.07±4.51	**64.18±2.50**	46.42±3.41	58.60±1.41	**62.16±1.43**	45.23±1.64
	ICMVC	29.23±0.37	38.31±1.87	21.55±0.88	24.33±4.17	25.74±5.40	14.17±4.28	22.90±2.83	19.87±1.93	10.52±1.76	19.43±2.62	16.10±1.80	7.79±1.67
	DVIMVC	44.53±4.00	41.83±3.49	23.78±2.59	43.37±2.57	45.18±1.85	28.29±1.80	39.57±5.54	34.39±4.68	20.52±3.40	39.47±2.61	36.71±3.60	22.41±2.82
	MVP(Ours)	**78.67±2.39**	**77.67±0.66**	**66.73±0.71**	**74.97±4.58**	**73.09±2.55**	**61.35±3.60**	**74.20±3.21**	**71.40±1.30**	**59.11±2.53**	**66.53±3.73**	**63.11±1.21**	**50.59±1.84**
Scene15	DCCA	38.22±1.00	41.20±0.52	19.89±0.42	36.16±2.17	39.46±1.79	17.12±2.80	34.05±3.50	37.26±3.49	14.48±4.45	30.84±6.35	33.93±6.55	12.60±5.08
	DCCAE	39.46±0.84	42.08±0.55	20.36±0.27	36.73±2.86	39.80±2.35	17.06±3.32	34.49±3.96	37.66±3.61	14.57±4.44	31.16±6.75	34.21±6.84	12.64±5.13
	DIMVC	42.51±2.42	41.53±1.60	24.45±2.23	40.37±1.85	38.57±1.30	20.84±1.54	40.17±2.14	35.95±2.50	20.59±2.68	36.01±1.27	32.57±1.08	16.29±2.24
	DSIMVC	29.43±1.21	30.38±0.95	14.86±0.59	31.38±1.02	32.54±1.33	16.29±0.90	27.24±1.21	28.68±0.85	13.38±0.64	28.42±0.78	29.09±0.81	13.85±0.21
	Completer	37.00±1.82	41.89±0.41	23.60±0.84	40.04±0.67	42.41±0.32	24.22±0.24	36.64±1.91	38.99±1.05	19.70±1.35	35.37±0.87	37.05±1.03	17.58±0.99
	CPSPAN	42.69±1.77	38.79±2.49	24.56±2.15	43.21±1.51	39.42±0.56	24.94±0.85	43.44±1.43	39.19±1.85	24.96±1.46	**42.53±2.56**	**38.41±2.43**	**24.38±1.97**
	ICMVC	43.88±2.35	40.03±1.14	25.80±1.54	43.14±1.02	38.06±0.51	24.74±0.89	37.96±1.87	33.45±0.93	20.34±0.78	36.70±2.22	35.80±1.30	18.35±1.32
	DVIMC	**45.16±2.40**	**45.06±1.33**	**28.64±1.52**	**43.68±1.54**	**42.32±2.11**	**26.68±1.05**	**41.13±3.32**	**39.58±2.25**	**25.03±2.42**	39.59±4.41	36.66±4.62	21.56±4.04
	MVP(Ours)	**45.70±1.63**	**43.77±0.93**	**27.90±1.38**	**45.81±2.75**	**42.54±1.02**	**27.53±1.86**	**45.28±1.44**	**41.84±0.79**	**26.80±1.20**	**43.14±2.20**	**39.53±0.67**	**24.68±1.58**

- **PolyMNIST: Preserving Consistent Semantics Across Diverse Styles**

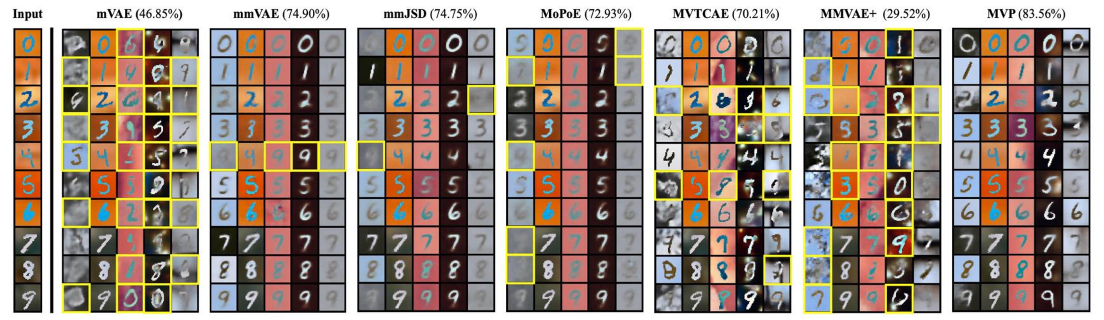
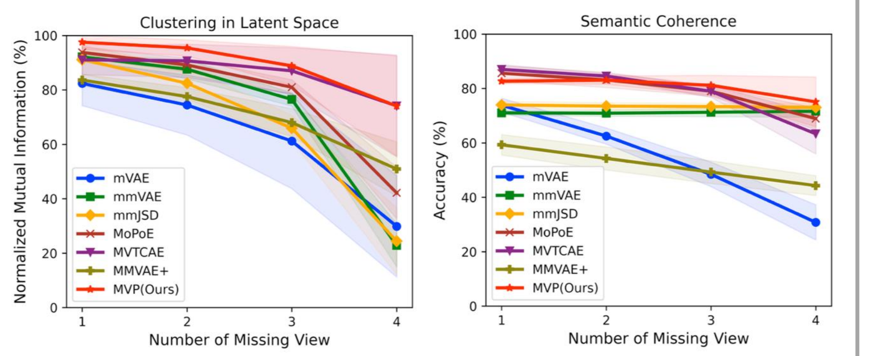


Figure 3: **Multi-view sample generation conditioned on view 2.** The leftmost column shows input images of view 2, randomly selected from digit classes 0 to 9. The following columns display multi-view samples (five views per sample) generated by various models. Ideally, the conditional generated digits should match the input digit, with yellow boxes highlighting inconsistencies. Accuracy scores, shown in parentheses, are derived from pre-trained classifiers on the generated images.

- **MoE-based methods**: Limited in effectively aggregating information
- **PoE-based methods**: Struggle with precision miscalibration of each view
- **MVTCAE**: Penalizes latent information not inferred from other views, retaining only highly correlated details, but struggles with consistency when views are missing.
- **MVP (Ours)**: Consistently balances high semantic coherence with diverse background styles.



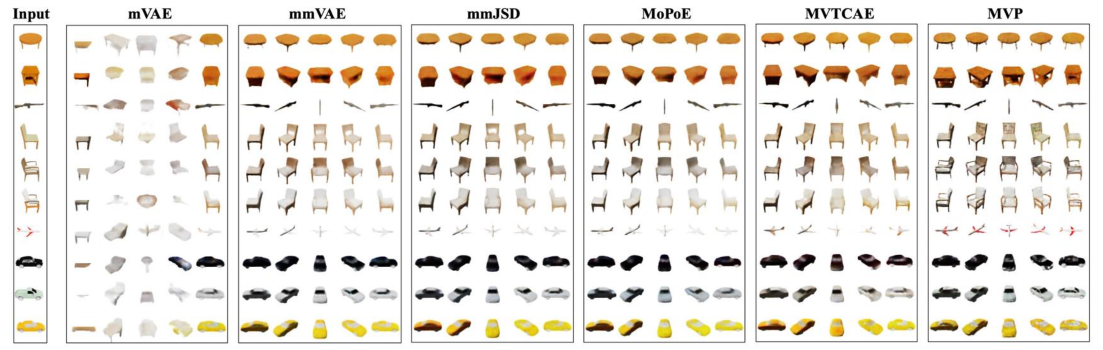


Figure 9: **Multi-view sample generation conditioned on view 5.** The leftmost column shows input images from view 5, randomly selected from five categories: table, rifle, chair, airplane, and car. The following columns display five-view samples generated by different models.

- **MVShapeNet: Capturing Detailed Information From Various Angles**
- Our method clearly infers more accurate details in missing views, such as the placement of table legs at different angles, changes in light and shadow, and hollowed-out armrests on chairs.

References

[1] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. International Conference on Learning Representations, 2021.

[2] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. Advances in Neural Information Processing Systems, 34:12194–12207, 2021.