Coordinated Disentanglement with Iterative Mode Discovery Under Hidden Correlations

Rong Hu, Ling Chen

State Key Laboratory of Blockchain and Data Security, Zhejiang University

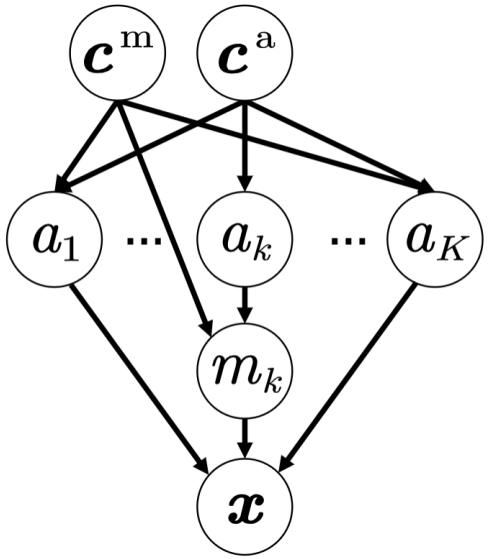
Underlying multi-modal distributions

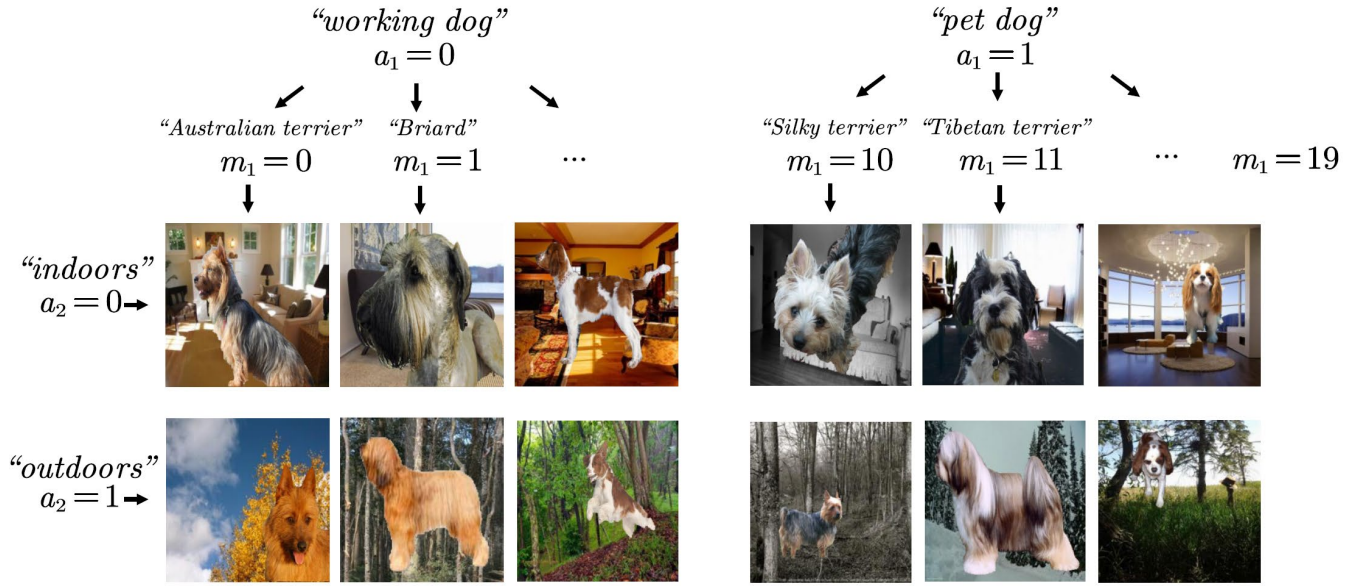
- **Definition of underlying modes**:

Under a value of a certain attribute, the data may form multi-modal distributions due to variations related to this attribute; each high-density region/cluster is referred to as a mode.

- **Why intra-class modes matter for attribute prediction**:

The modes under different attribute values may appear similar, which makes the attribute values hard to predict. Modeling underlying modes helps distinguish them.





Hidden correlations

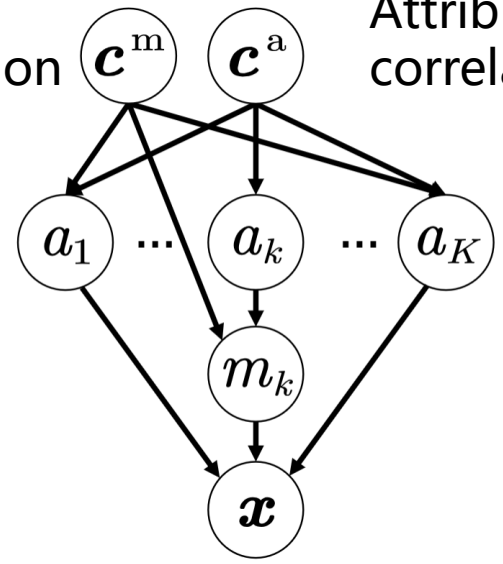
- **Definition of hidden correlations**:

The correlation between modes under a certain attribute and other attributes

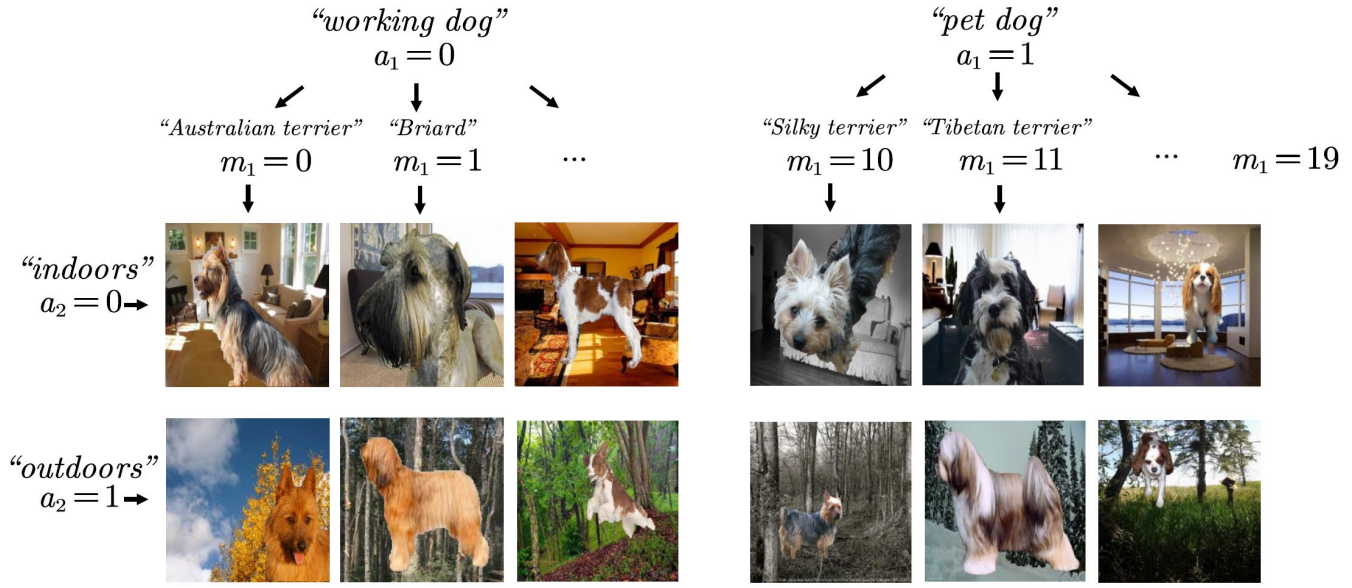
- **Why correlations hurt generalization**:

Under correlations, the model may infer one attribute by encoding others. Under correlation shifts at test time, the model will fail to generalize.

Hidden correlation



Attribute correlation



Disentangled representation learning

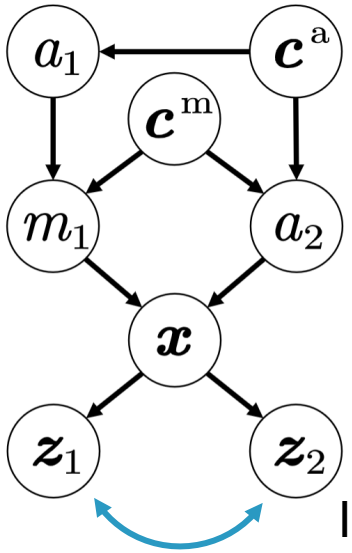
- **Goal of representation disentanglement:**

Learn a representation for each attribute that encodes the information about this attribute and its modes, and removes redundant information from the others.

- **How correlations affect disentanglement:**

Most disentanglement methods ignore correlations and enforce independence between attribute representations.

But when attributes are correlated, forcing the representations to be independent causes information loss.

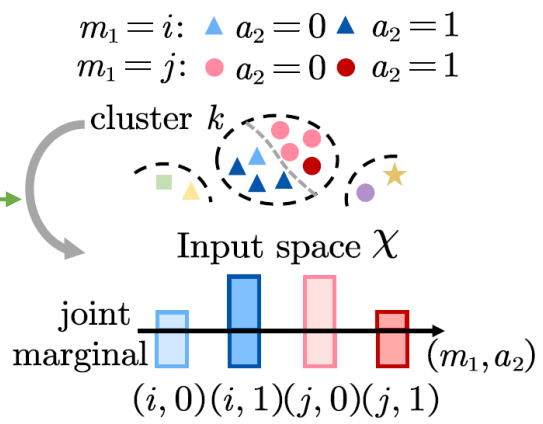


Independence constraint

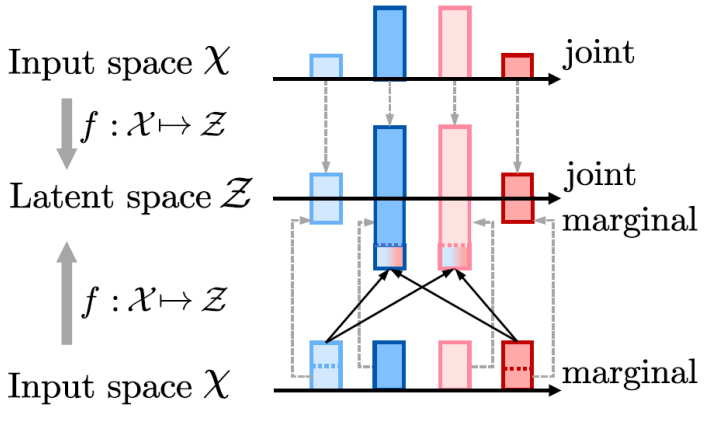
Theoretical insights

	Prior work	CoDID	Mode labels estimated by clustering may contain errors
Correlation type	Attribute correlation	Hidden correlation	Within-cluster correlation
Independence constraint	Unconditional independence	Attribute-based independence	Cluster-based independence
Problem	Attribute information loss	Mode information loss	Mode information loss
Solution / alternatives	Attribute-based independence	Mode-based independence	Weighted CMI minimization; Weight-informed cluster split

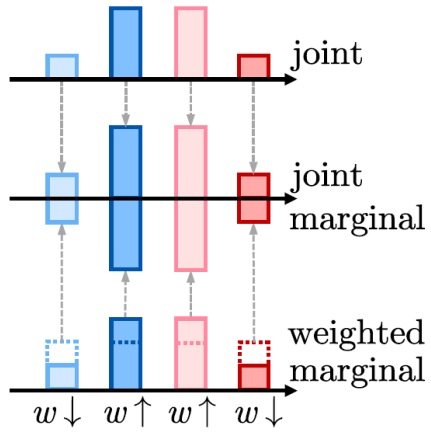
The harm of within-cluster correlations



(a) Mismatched label distributions.

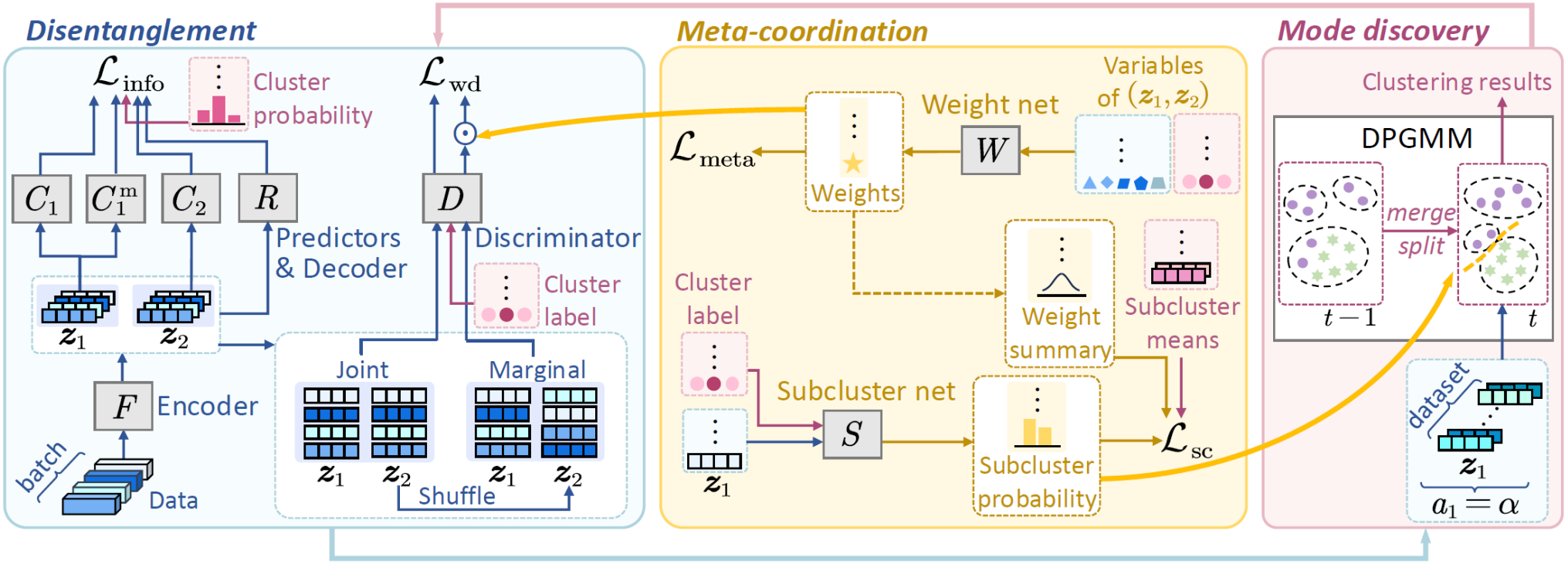


(b) Exact alignment.



(c) Weighted alignment.

CoDID framework



- Learn representations and disentangle them

Prevent error amplification:

- Learn the weights in the independence constraints to preserve informativeness
- Refine sub-clustering for cluster splits

- Estimate the number of modes and labels

Experiments across various data types and tasks

Data types

- Toy data
- Image data
- Time-series data

Hidden correlations

- Synthetic
- Real-world

Generalization tasks

- Correlation shift
- Out-of-distribution

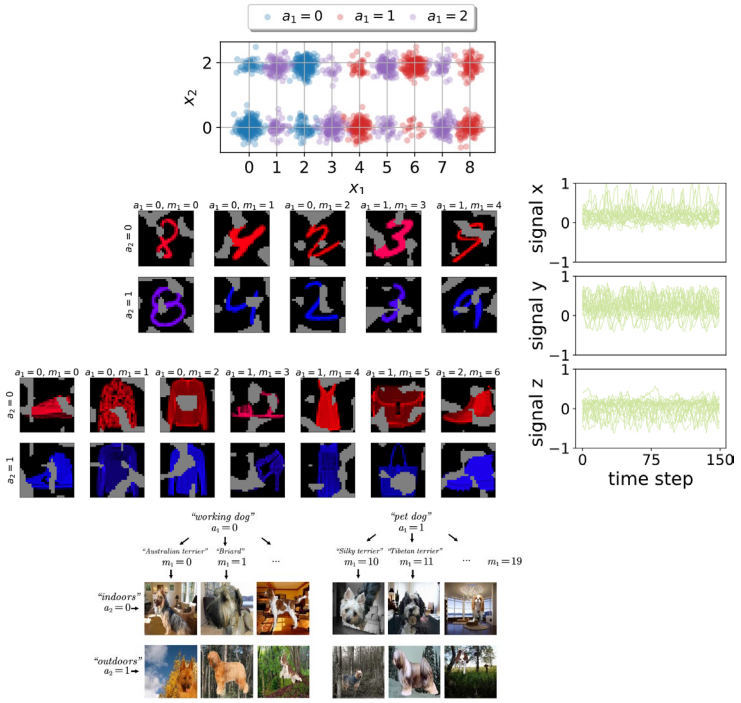


Table 8. Dataset descriptions.

Dataset	CMNIST	CFashion-MNIST	Canine-BG	UCI-HAR	RealWorld	HHAR	MFD
a_1	digit parity	fashion style	functional category	activity	activity	activity	incipient fault type
# values of a_1	2	2	2	6	8	6	3
values of a_1	even, odd	sporty, chic	work, pet	walking, walking upstairs, walking downstairs, sitting, standing, laying	climbing stairs up, climbing stairs down, jumping, lying, standing, sitting, running, walking	biking, sitting, standing, walking, stair up, stair down	healthy, inner-bearing damage, outer-bearing damage
a_2	digit color	clothing color	background environment	user	user	user	operating condition
# values of a_2	2	2	2	30	15	9	4
m_1	digit	clothing	dog breed	activity modes	activity modes	activity modes	fault modes
# values of m_1	5	7	20	unknown	unknown	unknown	unknown
values of m_1	digit 8, 4, 2, 3, 9	sneaker, pullover, shirt, sandal, dress, bag, ankle boot	Australian terrier, Briard, Welsh springer spaniel, Kelpie, Cairn terrier, Bedlington terrier, American Staffordshire terrier, Border collie, Bullmastiff, Irish water spaniel, and Silky terrier, Tibetan terrier, Blenheim spaniel, Toy terrier, West Highland white terrier, Standard poodle, French bulldog, Cardigan Welsh corgi, Pug, Miniature poodle	unknown	unknown	unknown	unknown
# samples	35000	49000	11884	11711	36980	14772	10916
# of groups	-	-	-	5	5	3	4
# channels	-	-	-	3	3	3	1
window length	-	-	-	128	150	128	5120

Experimental result

- CoDID outperforms existing disentanglement and invariant learning methods
- The components regarding meta-coordination and the iterative framework are effective

Table 1. Baseline comparison (mean±std.). "*" indicates that CoDID is statistically superior to the baseline by pairwise t-test at a 95% significance level. The best results are **bold**. The runner-up results are underlined. The improvement over the best baseline is calculated.

Method	CMNIST		CFashion-MNIST		Canine-BG		UCI-HAR		Realworld		HHAR		MFD	
	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
BASE	79.2±0.9*	78.3±0.9*	77.8±1.5*	77.1±1.5*	55.6±2.0*	54.6±2.3*	71.2±2.8*	69.7±3.6*	64.6±1.4*	65.4±1.4*	80.8±1.6*	80.9±2.0*	72.7±1.6*	76.3±0.9*
MMD	57.3±2.7*	41.2±7.0*	63.5±3.6*	62.7±3.7*	54.1±6.7*	52.1±9.4*	70.3±3.7*	66.2±3.5*	66.0±1.9*	65.2±2.3*	80.9±1.2*	80.5±1.7*	78.2±1.9*	79.1±1.6*
DANN	58.3±2.9*	56.9±2.6*	63.8±2.5*	62.5±2.8*	29.8±4.4*	29.6±4.8*	67.8±3.1*	65.1±3.0*	66.0±1.7*	65.1±2.0*	77.1±1.8*	76.7±1.5*	74.0±2.1*	76.3±1.4*
CORAL	58.6±3.2*	57.6±3.0*	68.0±3.4*	67.3±3.0*	51.0±3.5*	33.8±3.2*	74.4±2.8*	72.7±3.3*	64.8±1.9*	65.9±1.6*	81.0±1.4*	81.2±1.7*	77.3±1.8*	77.9±1.5*
IRM	81.0±3.5*	80.3±3.2*	79.3±2.6*	78.9±2.5*	54.3±5.6*	54.3±5.1*	70.9±3.0*	69.6±3.4*	65.4±1.6*	65.4±1.5*	82.5±1.3*	82.5±1.2*	78.4±1.7*	79.9±1.3*
REx	81.7±3.3*	80.9±3.6*	76.3±3.5*	75.9±3.2*	56.0±5.2*	56.0±5.7*	73.7±2.5*	73.2±2.8*	65.1±1.5*	65.3±1.7*	80.6±1.5*	80.4±1.6*	78.8±1.9*	80.3±1.4*
DTS	55.6±3.0*	44.0±1.9*	61.2±2.0*	60.1±2.0*	41.8±4.7*	35.7±2.3*	72.8±3.3*	70.1±2.6*	64.4±2.3*	64.9±1.5*	79.8±2.4*	79.7±1.7*	67.0±2.2*	67.4±1.5*
IDE-VC	54.2±4.0*	49.1±2.9*	58.7±3.8*	57.1±4.3*	49.1±2.2*	45.7±2.0*	73.6±3.1*	73.2±3.4*	65.2±1.3*	65.0±1.7*	80.7±2.0*	80.6±1.4*	74.1±1.8*	76.3±1.1*
MI	56.9±4.0*	45.2±2.8*	62.1±3.0*	60.7±3.0*	52.3±4.8*	51.0±4.4*	74.9±2.1*	74.5±2.7*	66.0±1.8*	65.5±1.6*	80.9±1.7*	80.7±2.1*	76.3±1.2*	77.6±1.6*
A-CMI	58.5±2.3*	41.2±5.9*	61.8±5.3*	60.0±6.5*	53.0±8.2*	52.8±8.1*	71.4±3.4*	70.0±3.0*	65.4±1.5*	65.5±1.2*	80.2±1.8*	80.3±2.3*	78.8±1.4*	79.8±0.7*
HFS	66.5±2.1*	64.9±2.1*	66.2±3.6*	65.3±3.1*	46.8±3.7*	46.2±4.1*	67.1±3.5*	65.1±4.0*	48.9±1.8*	39.8±1.5*	78.2±1.2*	78.3±1.5*	75.4±1.7*	71.0±1.3*
CODA	69.8±1.9*	68.3±2.3*	72.5±1.5*	71.3±1.9*	43.9±2.2*	43.9±2.3*	71.1±4.2*	70.4±5.0*	65.3±0.7*	66.7±0.6*	77.9±1.7*	77.6±1.8*	62.5±0.8*	54.6±1.2*
ID-FaceVC	58.1±2.2*	56.2±2.4*	63.6±1.4*	62.5±1.7*	42.1±6.2*	41.3±6.7*	68.0±8.1*	68.0±9.3*	65.2±1.5*	66.2±1.8*	78.6±1.2*	77.8±1.3*	71.1±1.9*	71.6±1.6*
DIOSC	73.4±0.7*	72.6±1.0*	73.2±1.6*	72.8±1.4*	42.5±5.4*	42.0±5.3*	74.6±3.0*	74.5±3.7*	65.2±1.3*	66.3±1.2*	79.0±1.4*	78.6±1.5*	68.9±1.6*	67.9±1.4*
CoDID	**85.2**±2.7	**84.6**±3.0	**84.1**±3.2	**83.5**±3.3	**68.3**±5.2	**68.2**±5.7	**86.2**±0.5	**86.7**±0.6	**70.9**±1.3	**71.7**±1.0	**88.4**±1.1	**88.3**±1.1	**90.7**±1.7	**91.9**±1.7
Improvement	+3.5 %	+3.7 %	+4.8 %	+4.6 %	+12.3 %	+12.2 %	+11.3 %	+12.2 %	+4.9 %	+5.0 %	+5.9 %	+5.8 %	+11.9 %	+11.6 %

Table 2. Variant comparison (mean±std.). The notations follow Table 1.

Method	Design choices							CMNIST		CFashion-MNIST		Canine-BG		UCI-HAR		Realworld		HHAR		MFD	
	$\mathcal{L}_{mp}$	$\mathcal{L}_{d}$	IC	IN	$\mathcal{L}_{wd}$	$\mathcal{L}_{sc}$	UP	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1	Acc.	Mac. F1
BASE	-	-	-	-	-	-	-	79.2±0.9*	78.3±0.9*	77.8±1.5*	77.1±1.5*	55.6±2.0*	54.6±2.3*	71.2±2.8*	69.7±3.6*	64.6 ±1.4*	65.4 ±1.4*	80.8±1.6*	80.9±2.0*	72.7±1.6*	76.3±0.9*
CoDID-km-MP	✓	✗	✗	✗	✗	✗	✗	74.2±1.9*	73.6±1.5*	74.2±2.0*	73.7±2.1*	63.2±4.6*	62.2±4.2*	77.6±5.3*	77.5±4.5*	63.7±0.9*	63.3±0.8*	83.5±1.2*	83.4±1.2*	78.4±2.6*	80.1±2.5*
CoDID-km-MC	✗	✓	✗	✗	✗	✗	✗	73.0±1.4*	72.1±1.4*	71.6±1.5*	70.2±1.3*	60.9±3.2*	60.2±3.4*	80.2±3.3*	79.7±4.4*	68.4±0.8	68.0±0.9	83.8±1.4*	83.2±1.5*	81.1±2.1*	80.2±2.3*
CoDID-km-ID	✓	✓	✗	✗	✗	✗	✗	77.1±1.0*	76.7±1.2*	73.6±1.4*	72.8±1.3*	62.6±4.9*	62.5±4.7*	77.6±1.8*	76.8±2.3*	68.3±1.2*	67.8±1.1*	77.2±1.9*	75.5±1.5*	80.6±1.7*	80.9±1.2*
CoDID-km-SD	✓	✓	✗	✗	✗	✗	✗	76.3±1.7*	75.6±1.5*	72.9±1.6*	72.3±1.6*	61.5±3.5*	61.4±3.6*	77.4±1.5*	76.8±1.8*	66.2±1.3*	66.6±1.8*	81.0±2.4*	81.2±1.8*	79.2±1.8*	79.2±1.3*
CoDID-km	✓	✓	✗	✗	✗	✗	✗	79.0±1.8*	78.3±1.6*	77.4±2.3*	76.9±2.2*	64.9±2.3*	64.5±2.2*	83.0±3.0*	83.3±3.6*	69.8±1.9	69.9±1.4	84.5±2.3*	84.2±1.5*	82.5±2.0*	82.5±1.5*
CoDID-itkm	✓	✓	✓	✗	✗	✗	✗	76.5±1.0*	75.1±1.2*	76.6±1.5*	76.0±1.5*	60.5±5.5*	60.0±5.6*	79.4±1.6*	79.1±1.6*	66.1±1.8*	65.2±1.8*	80.5±0.7*	80.4±0.8*	79.6±0.9*	80.0±0.5*
CoDID-itdpm	✓	✓	✓	✓	✗	✗	✓	77.6±2.0*	77.0±2.1*	77.4±1.3*	76.8±1.4*	61.4±3.2*	61.1±3.5*	83.2±0.5*	83.7±0.6*	65.7±1.3*	66.0±1.6*	83.8±0.3*	83.5±0.3*	82.2±0.9*	84.5±1.2*
CoDID-w/o-SC	✓	✓	✓	✓	✓	✗	✓	79.1±0.8*	78.4±0.7*	78.4±1.6*	77.9±1.6*	64.5±3.7*	64.4±3.7*	83.5±0.3*	84.0±0.4*	65.4±1.9*	65.3±1.9*	85.3±0.8*	85.0±0.8*	82.9±1.3*	85.1±1.3*
CoDID-UA	✓	✓	✓	✓	✓	✓	✗	82.0±0.7*	81.3±0.5*	79.4±1.9*	78.9±1.9*	66.2±3.9*	65.9±3.4*	84.2±3.2*	84.6±3.6*	69.6±1.4	70.9±1.5	86.0±1.2*	85.7±1.3*	85.8±2.4*	86.9±2.0*
CoDID	✓	✓	✓	✓	✓	✓	✓	**85.2**±2.7	**84.6**±3.0	**84.1**±3.2	**83.5**±3.3	**68.3**±5.2	**68.2**±5.7	**86.2**±0.5	**86.7**±0.6	**70.9**±1.3	**71.7**±1.0	**88.4**±1.1	**88.3**±1.1	**90.7**±1.7	**91.9**±1.7

Toy data classification boundary

- The classification boundary of CoDID separates different modes along the x1 axis, while excluding redundant information about the other attribute along the x2 axis.

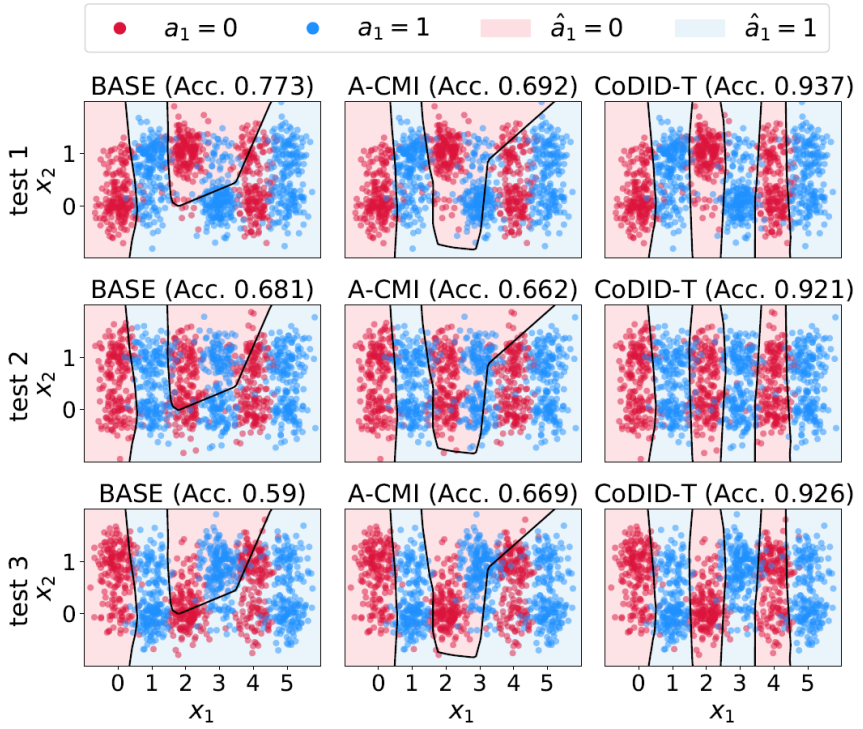
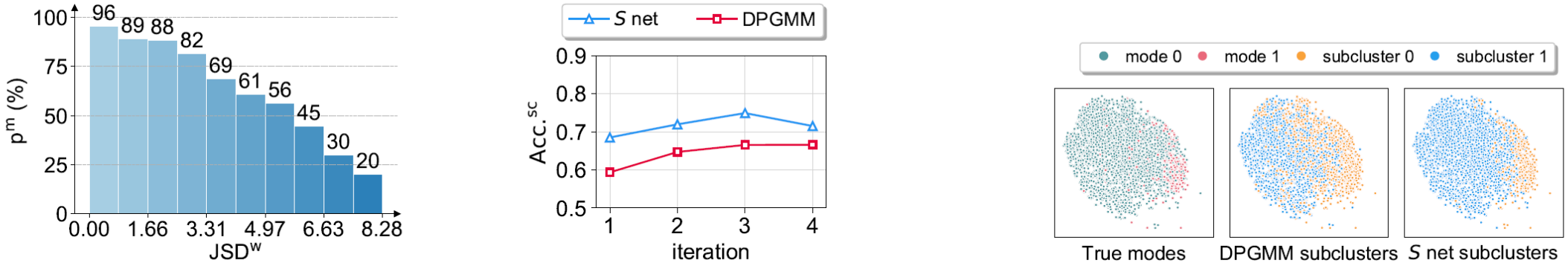


Figure 5. Toy decision boundary. Attributes a_1, a_2 control data dimensions $\boldsymbol{x}_1, \boldsymbol{x}_2$, respectively. Clusters centered at $\boldsymbol{x}_1 = 0, 2, 4$ and $1, 3, 5$ represent modes under $a_1 = 0$ and 1, respectively.

Effectiveness of meta-learned weights and cluster split

- A larger weight divergence corresponds to a lower chance of sharing the same mode, which supports cluster splits and latent mode discovery.

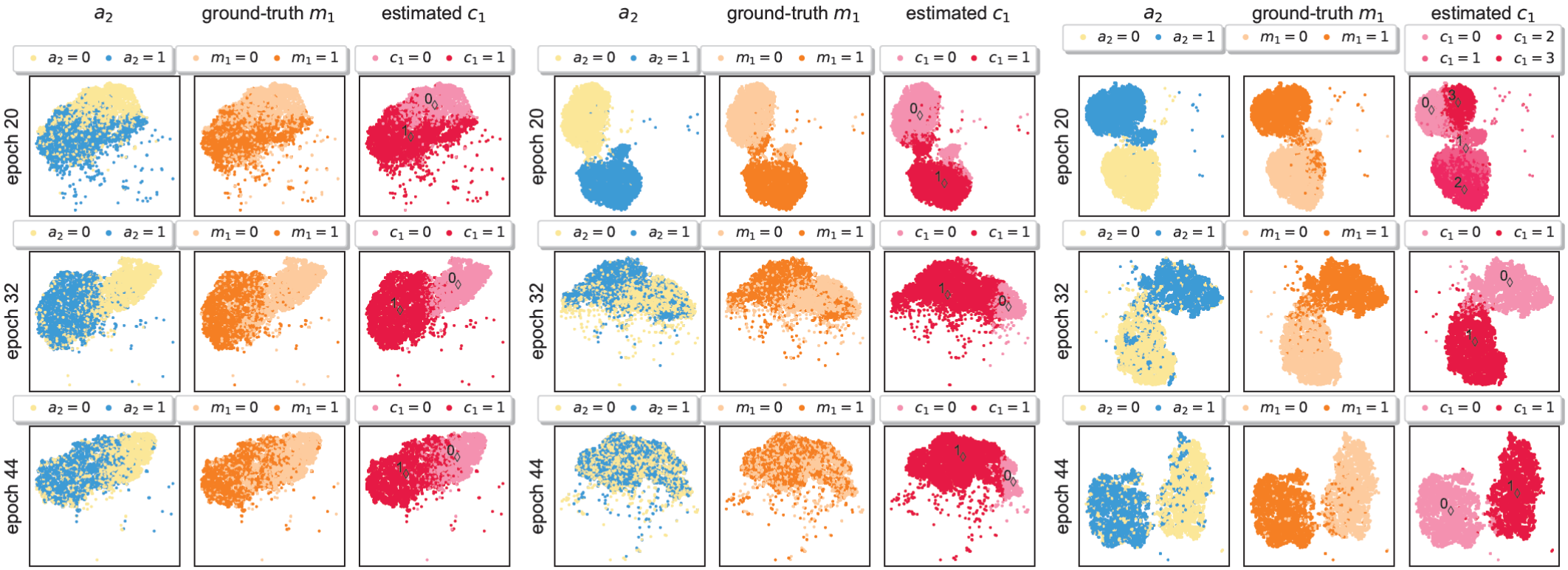


(a) Weight patterns *(b)* Subclustering accuracy *(c)* Subclustering visualization

Figure 4. Weight and subclustering analysis on CMNIST. The cluster k in (c) exhibits statistically significant within-cluster correlations based on the χ^2-test ($p = 4.22 \times 10^{-15}$), and DPGMM and S net achieve subcluster accuracies of 0.678 and 0.892, respectively.

Representation distribution on synthetic data

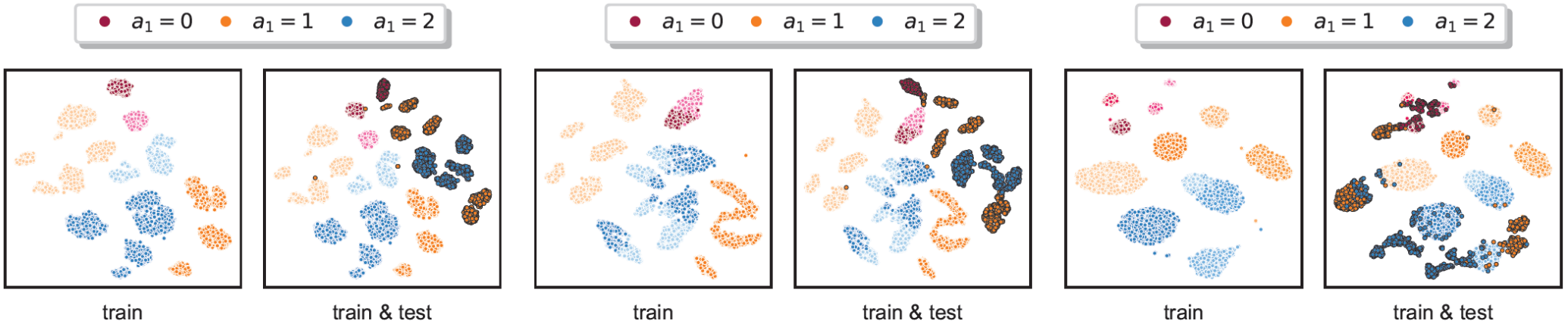
- Through iterative clustering and disentanglement, CoDID gradually removes redundant information while preserving mode information.



(a) CoDID-km *(b)* CoDID-itkm *(c)* CoDID

Representation distribution on real time-series data

- Modeling intra-class modes improves discriminability on real time-series data and boosts generalization to out-of-distribution samples.

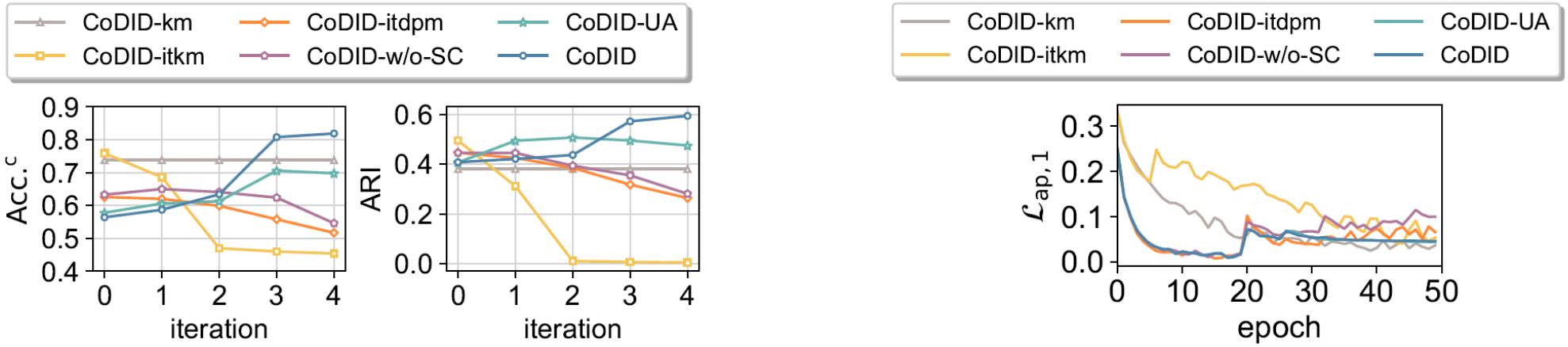


(a) CoDID-km (Sil. = 0.574, $d_{\mathcal{A}}$ = 1.99) *(b)* CoDID-itkm (Sil. = 0.401, $d_{\mathcal{A}}$ = 1.97) *(c)* CoDID (Sil. = 0.629, $d_{\mathcal{A}}$ = 1.78)

Figure 7. Machine fault representation distributions on MFD. Different colors, i.e., red, orange, and blue, indicate different machine fault types a_1, i.e., healthy, inner-bearing damage, and outer-bearing damage, respectively. Different shades of a color indicate different modes under the a_1 value. Points with white edges indicate training data, and points with black edges indicate test data.

Convergence analysis

- During training, the clustering performance of CoDID steadily improves, and the prediction loss curves remain stable even as the number of clusters changes.



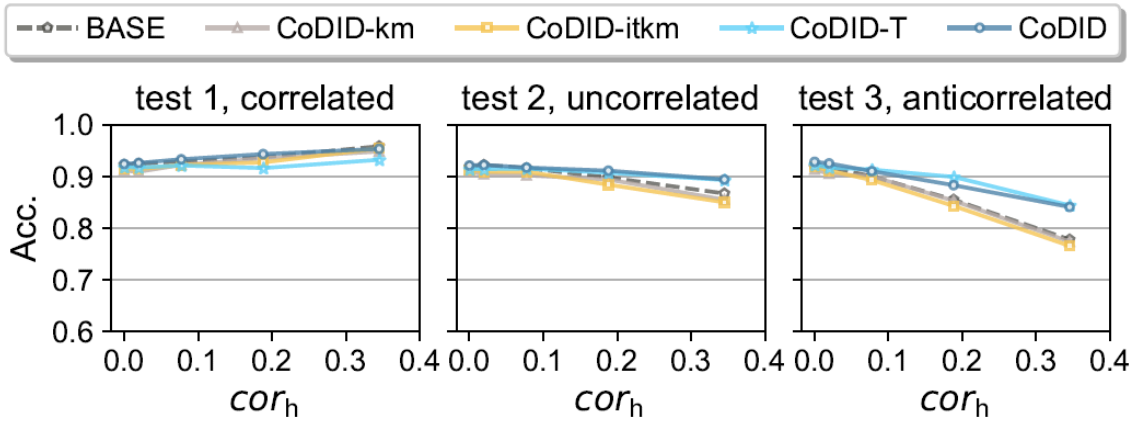
(a) Clustering metrics

(b) Attribute prediction loss for a_1

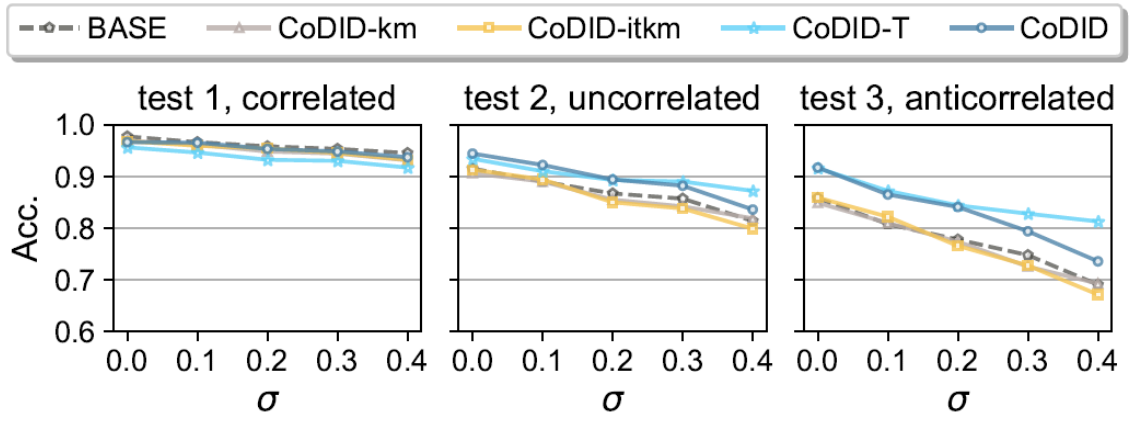
Figure 6. Learning curves on CMNIST.

Robustness against noise levels and correlation strengths

- CoDID maintains a robust advantage across different noise levels and hidden correlation strengths.



(a) Hidden correlation cor_h



(b) Noise level σ

Figure 10. Comparison under varying correlation strengths and noise levels on CFashion-MNIST.