EBERHARD KARLS UNIVERSITÄT TÜBINGEN

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Disentanglement and Generalization Under Correlation Shifts

Christina Funke*, Paul Vicol*, Kuan-Chieh Wang
Matthias Kümmerer†, Richard Zemel†, Matthias Bethge†

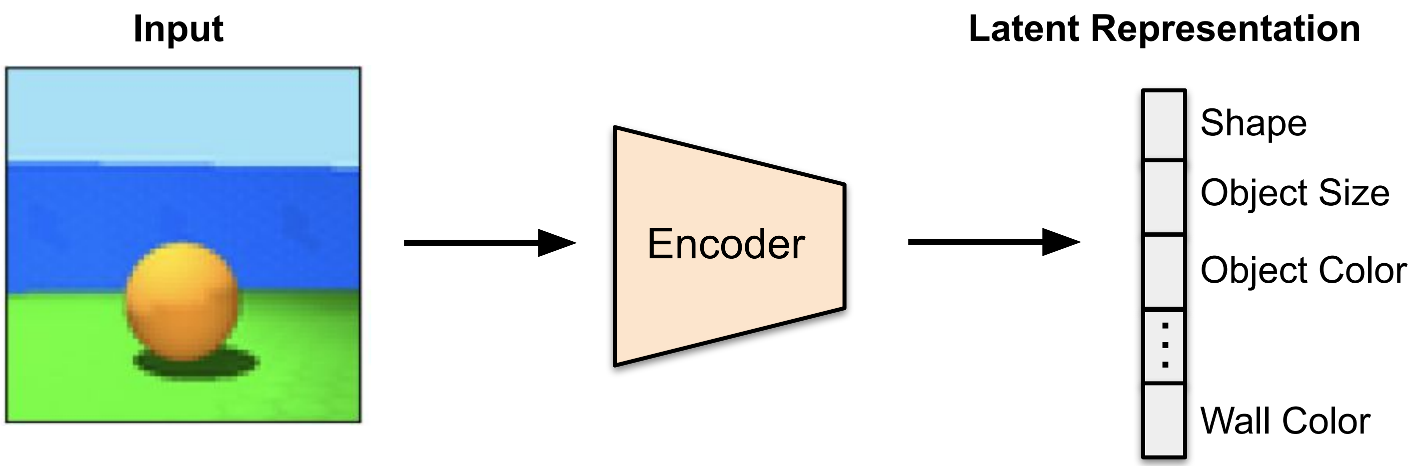
* Equal Contribution   † Shared Senior Authors






Introduction - Disentanglement

- A *disentangled representation* is one in which different factors of variation are represented by *different components of the representation*
  - e.g., different dimensions in the latent space

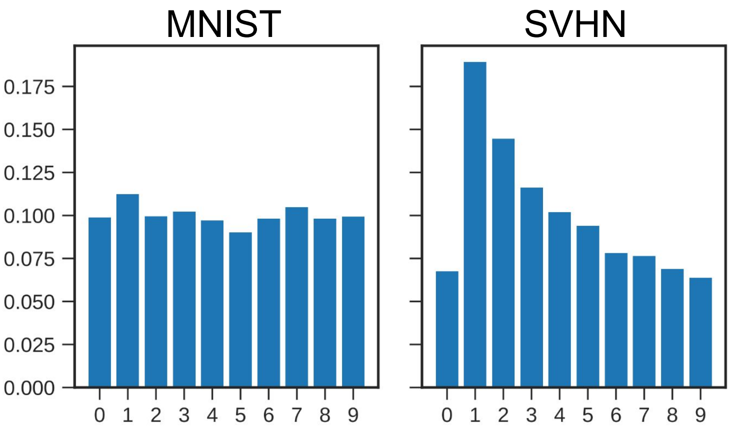


- Disentangled representations are useful for:
  - Improving *fairness* and interpretability
  - Improved *robustness to OOD data* (in domain adaptation & generalization)
  - Controllable generative modeling

Correlations Between Attributes

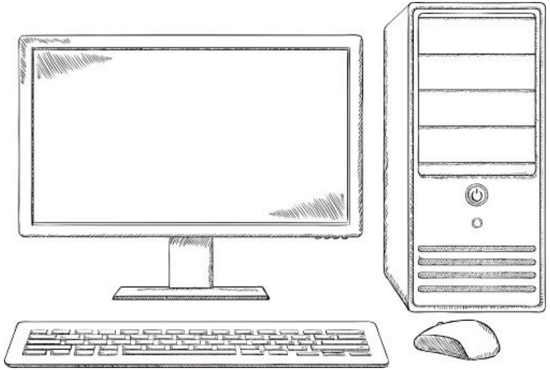
- Most work *assumes that the ground-truth factors of variation are independent*
  - That is, that there are *no correlations between attributes*
  - This holds for simple/synthetic benchmark tasks (e.g., dSprites, Shapes3D)
- But *real data often has correlations* between attributes, breaking this assumption

Class/Domain Correlation



Foreground/Background

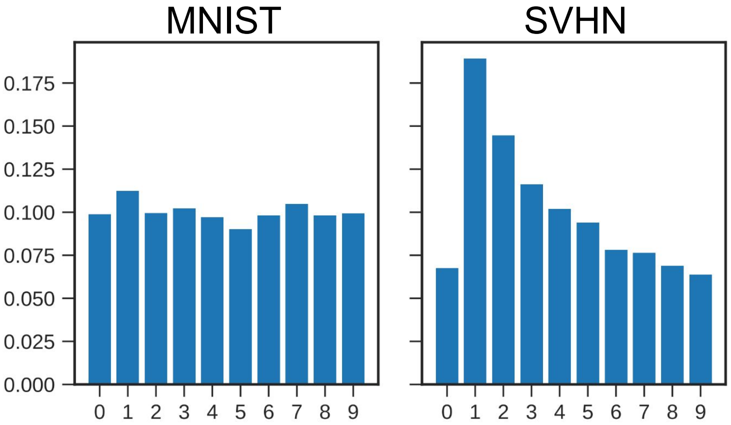
Object Co-Occurrence



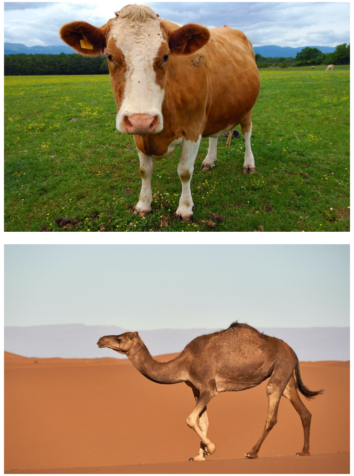
Correlations Between Attributes

- Most work *assumes that the ground-truth factors of variation are independent*
  - That is, that there are *no correlations between attributes*
  - This holds for simple/synthetic benchmark tasks (e.g., dSprites, Shapes3D)
- But *real data often has correlations* between attributes, breaking this assumption

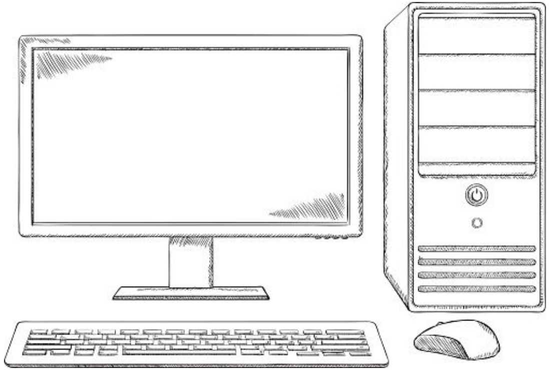
Class/Domain Correlation



Foreground/Background



Object Co-Occurrence



- Correlations also occur in *fairness & healthcare*: demographics differ between hospitals

Introduction - Disentanglement of Correlated Attributes

- *Disentanglement of correlated attributes* is problematic (Träuble et al., 2020)
  - For correlated attributes, the corresponding latent codes *encode a mixture of these attributes.*

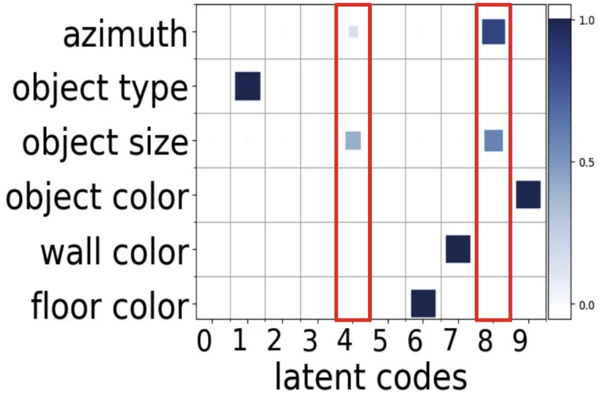
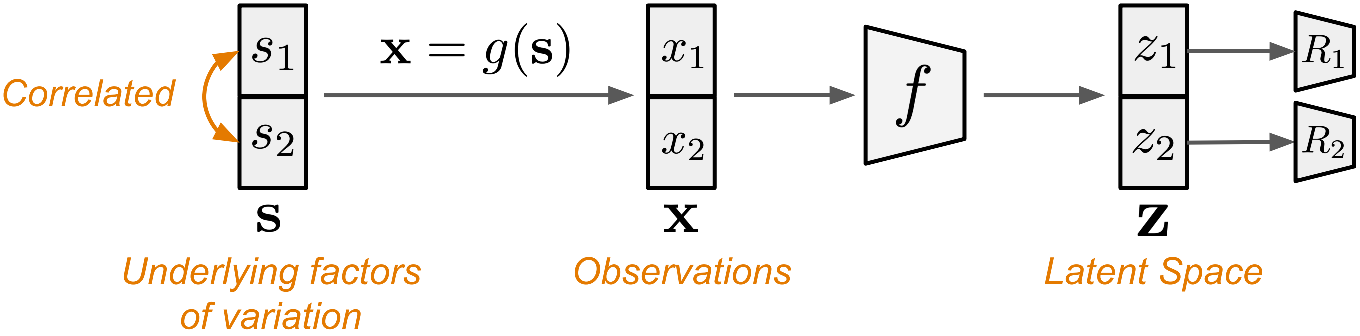


Figure from Träuble et al., 2020

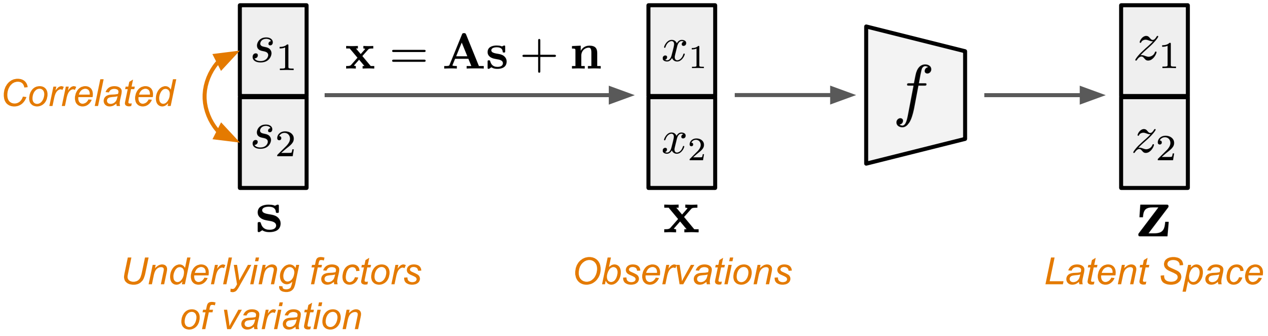
  - Träuble et al. suggest to address this with *weak supervision.*
  - We show that *even under full supervision*, enforcing independence between latent subspaces can fail.

Problem Setup



- We have noisy data $\mathbf{x} = g(\mathrm{s})$ where $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ are the *underlying factors of variation*, which may be correlated
- **Goal:** Find a mapping to a latent space $f(\mathbf{x}) = \mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K)$ such that we can recover the ground-truth attributes via *linear functions* $\hat{s}_k = \mathbf{R}_k \mathbf{z}_k \approx s_k$
- **Goal:** *Learn a model robust to correlation shifts*
  - If we train on data where $\mathrm{corr}(s_i, s_j) > 0$, then we want the resulting model to perform well on *uncorrelated data* $\mathrm{corr}(s_i, s_j) = 0$, or *anticorrelated data*, $\mathrm{corr}(s_i, s_j) < 0$

Problem Setup



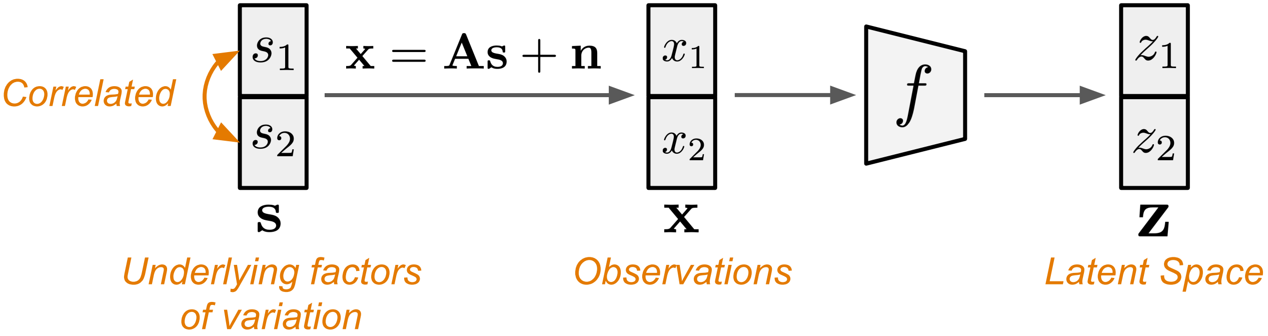
- *Linear generative model* with *correlated* Gaussian source signals

$$\underbrace{\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}}_{}\quad , \quad \underbrace{\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s})}_{\text{Gaussian Source Signals}} \quad , \quad \underbrace{\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})}_{\text{Gaussian Noise Variables}}$$

Ground-Truth Mixing Matrix

- **Goal:** Recover a mapping that *inverts the data-generating process*, $f(\mathbf{x}) = \mathbf{z} = \mathbf{A}^{-1}\mathbf{x}$

Supervised Learning Does Not Yield Disentanglement



- *Linear generative model* with *correlated* Gaussian source signals

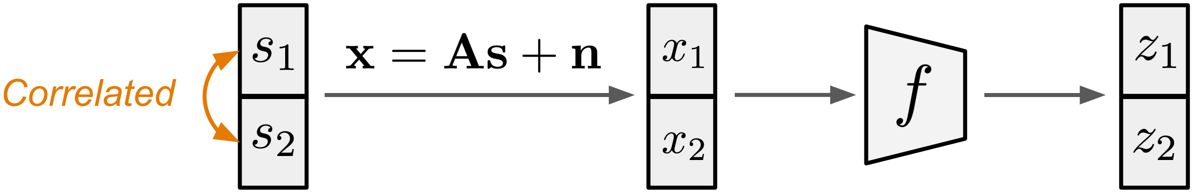
$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \quad , \quad \mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s}) \quad , \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$$

- The *optimal linear regression* solution is given by:

$$\boxed{\hat{\mathbf{s}}(\mathbf{x}) = \mathbf{C_{sx}} \mathbf{C_x}^{-1} \mathbf{x}} \quad \text{where} \quad \mathbf{C_{xs}} = \mathbf{C_s}\mathbf{A}^\top \quad \text{and} \quad \mathbf{C_x} = \mathbf{A}\mathbf{C_s}\mathbf{A}^\top + \mathbf{C}$$

$\neq \mathbf{A}^{-1}$ because it is *biased by the correlation structure* $\mathbf{C_s}$ and $\mathbf{C_n}$ towards directions of maximal signal to noise ratio

Supervised Learning Does Not Yield Disentanglement



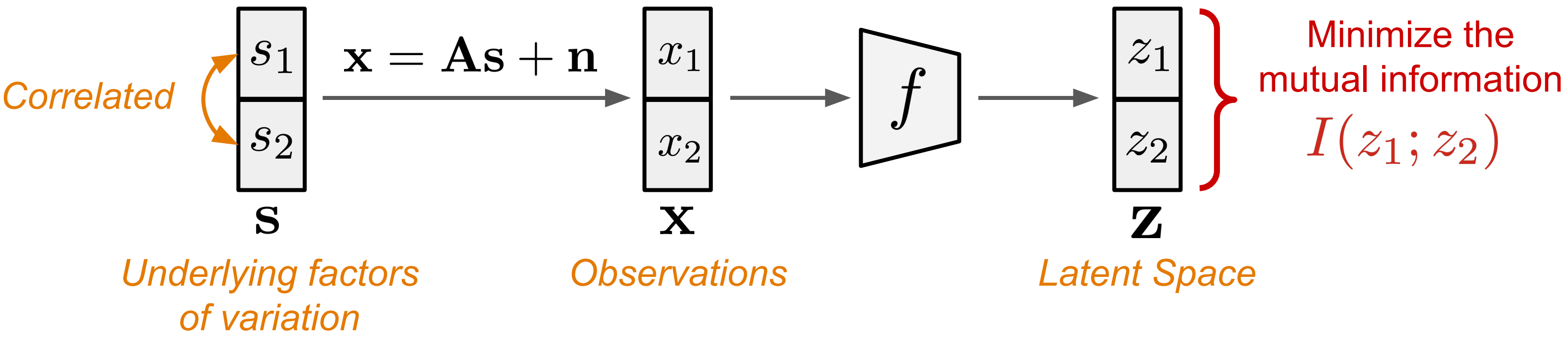
- **Problem:** Linear regression is *sensitive to noise*

	Base	Base + MI	Base + CMI
VE, Train (Corr = 0.8)	91.9%	69.8%	90.9%
VE, Test (Corr = 0)	87.6%	65.0%	90.9%

The estimator $\hat{\mathbf{s}}$ tries to make use of the *assumed correlation* between s_1 and s_2 to *counteract the information lost due to noise*, but this correlation is *no longer present in the test data*.

- There is *no constraint* preventing the model from encoding both s_1 and s_2 into *each of* z_1 and z_2

Unconditional Independence Constraint Does Not Help

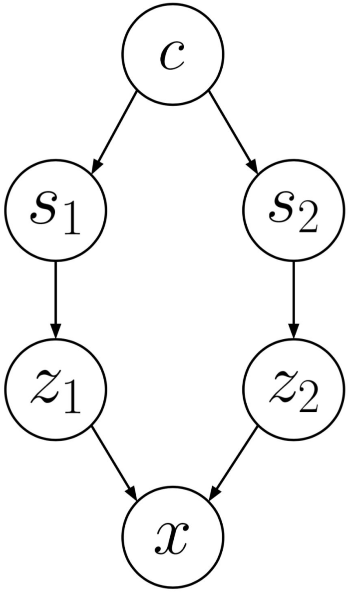


- **Common approach:** Enforce independence by minimizing the MI between latent subspaces $I(z_1; z_2) = 0$
- **Issue:** Because s_1 and s_2 are correlated, $I(s_1; s_2) > 0$
  - By enforcing $I(z_1; z_2) = 0$, *at least one of the subspaces cannot contain all relevant information about its attribute*
  - This leads to *poor performance on the in-distribution (correlated) training data*

	Base	Base + MI	Base + CMI
VE, Train (Corr = 0.8)	91.9%	69.8%	90.9%
VE, Test (Corr = 0)	87.6%	65.0%	90.9%

Conditional Independence is the Correct Objective

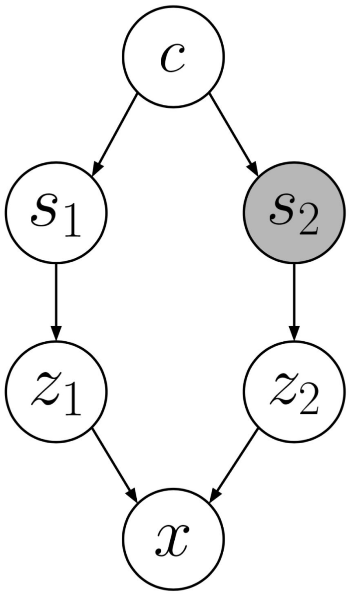
- Assuming a *common cause* for the correlation between s_1 and s_2, there is a connection in the graphical model between z_1 and z_2 introducing the statistical dependence.



$$I(z_1; z_2) > 0$$

Conditional Independence is the Correct Objective

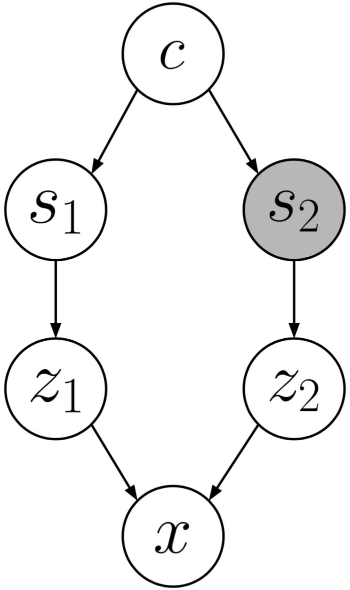
- Assuming a *common cause* for the correlation between s_1 and s_2, there is a connection in the graphical model between z_1 and z_2 introducing the statistical dependence.
- Observing either s_1 or s_2 *disconnects* z_1 and z_2



$$I(z_1; z_2 | s_2) = 0$$

Conditional Independence is the Correct Objective

- Assuming a *common cause* for the correlation between s_1 and s_2, there is a connection in the graphical model between z_1 and z_2 introducing the statistical dependence.
- Observing either s_1 or s_2 *disconnects* z_1 and z_2
- We desire that z_1 and z_2 share *as little information as possible* (given the ground truth correlation)

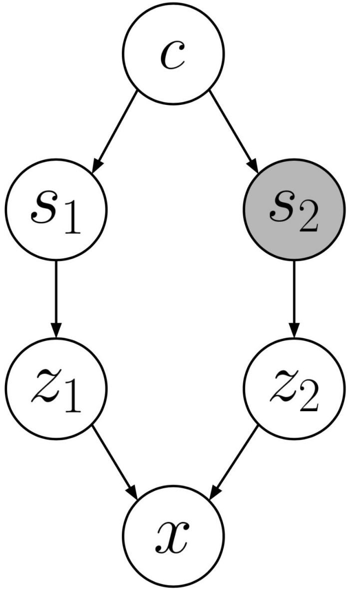


$$I(z_1; z_2 | s_2) = 0$$

Conditional Independence is the Correct Objective

- Assuming a *common cause* for the correlation between s_1 and s_2, there is a connection in the graphical model between z_1 and z_2 introducing the statistical dependence.
- Observing either s_1 or s_2 *disconnects* z_1 and z_2
- We desire that z_1 and z_2 share *as little information as possible* (given the ground truth correlation)
- We *minimize the MI* between latent subspaces *conditioned on the attributes*:

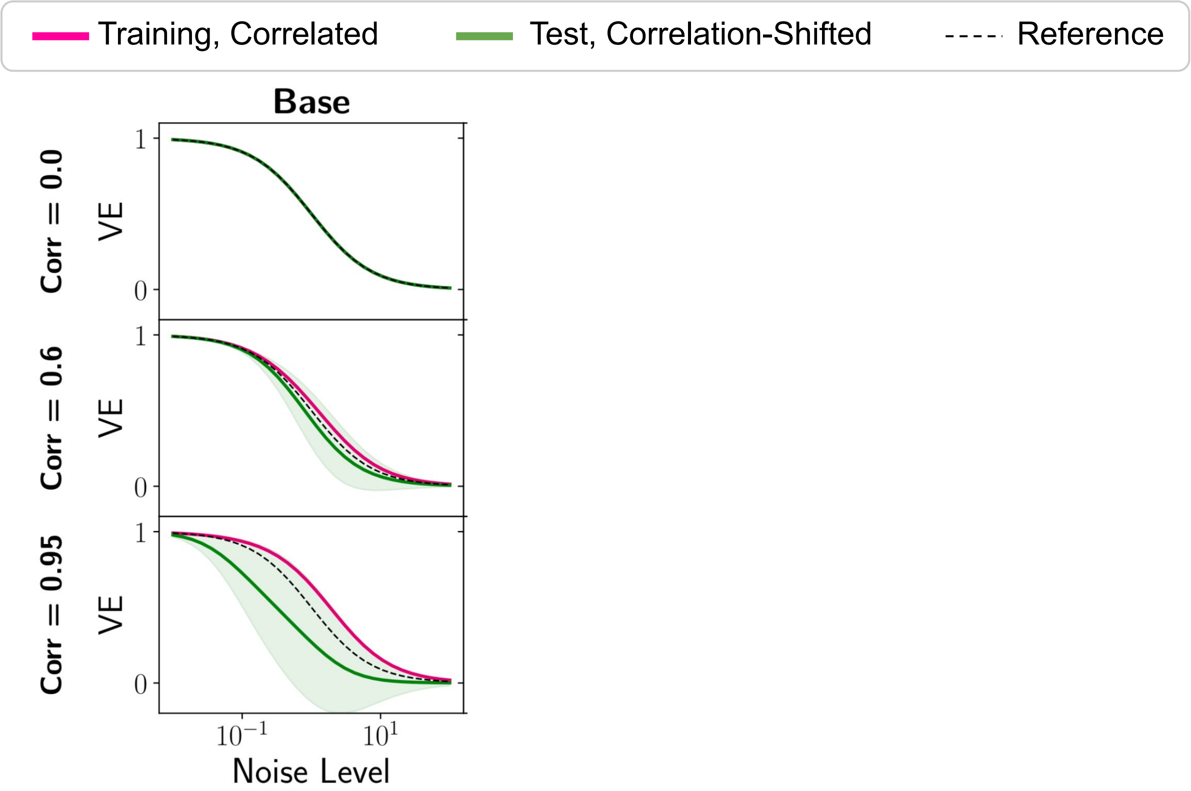
$$I(z_1; z_2 \mid s_1) = 0 \quad \text{and} \quad I(z_1; z_2 \mid s_2) = 0$$



	Base	Base + MI	Base + CMI
VE, Train (Corr = 0.8)	91.9%	69.8%	90.9%
VE, Test (Corr = 0)	87.6%	65.0%	90.9%

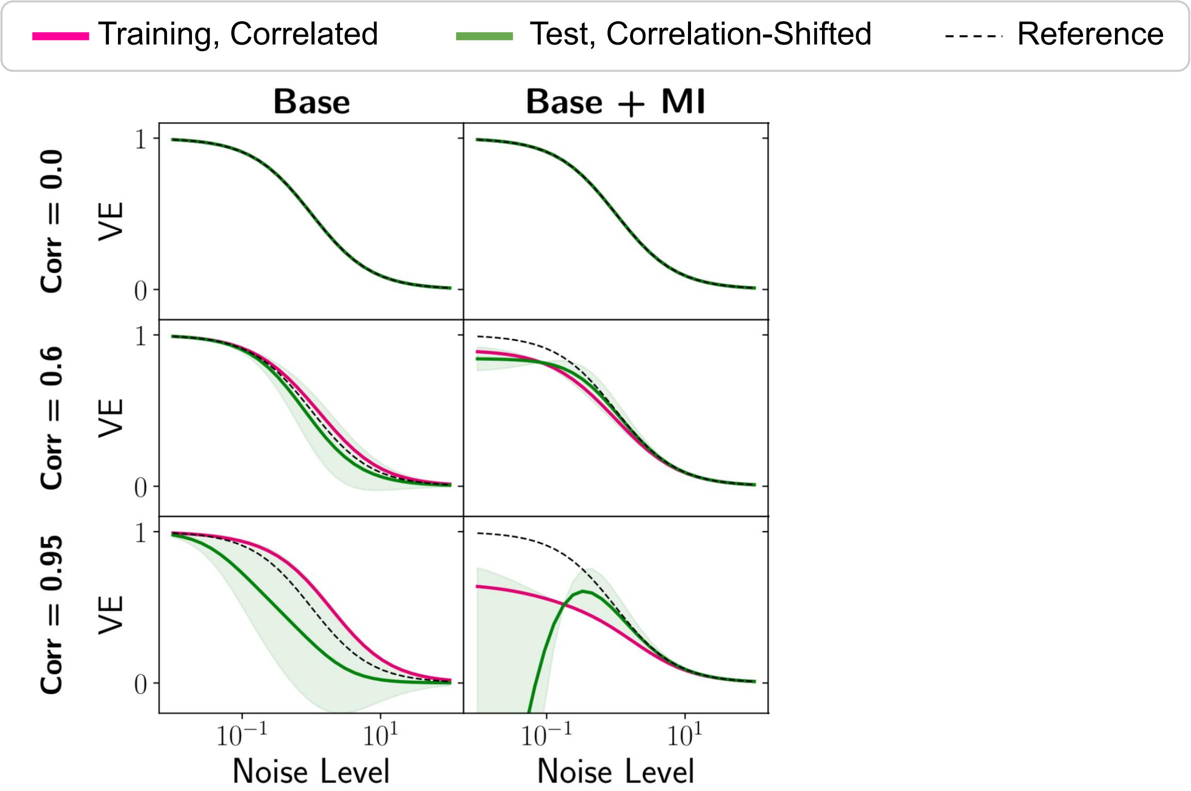
Training Set Correlation and Noise Level

Linear regression: $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$, $\underbrace{\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s})}_{\text{Train correlation}}$, $\underbrace{\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})}_{\text{Noise level}}$



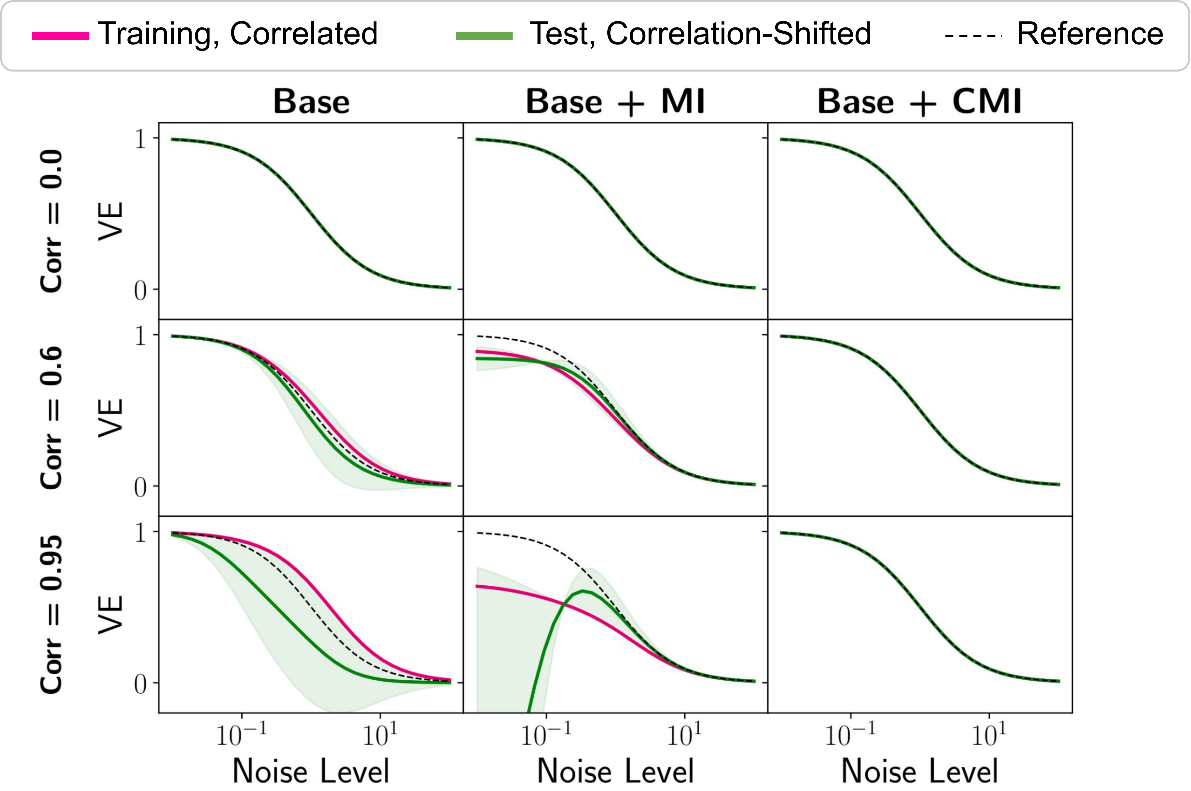
Training Set Correlation and Noise Level

Linear regression: $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$, $\underbrace{\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s})}_{\text{Train correlation}}$, $\underbrace{\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})}_{\text{Noise level}}$



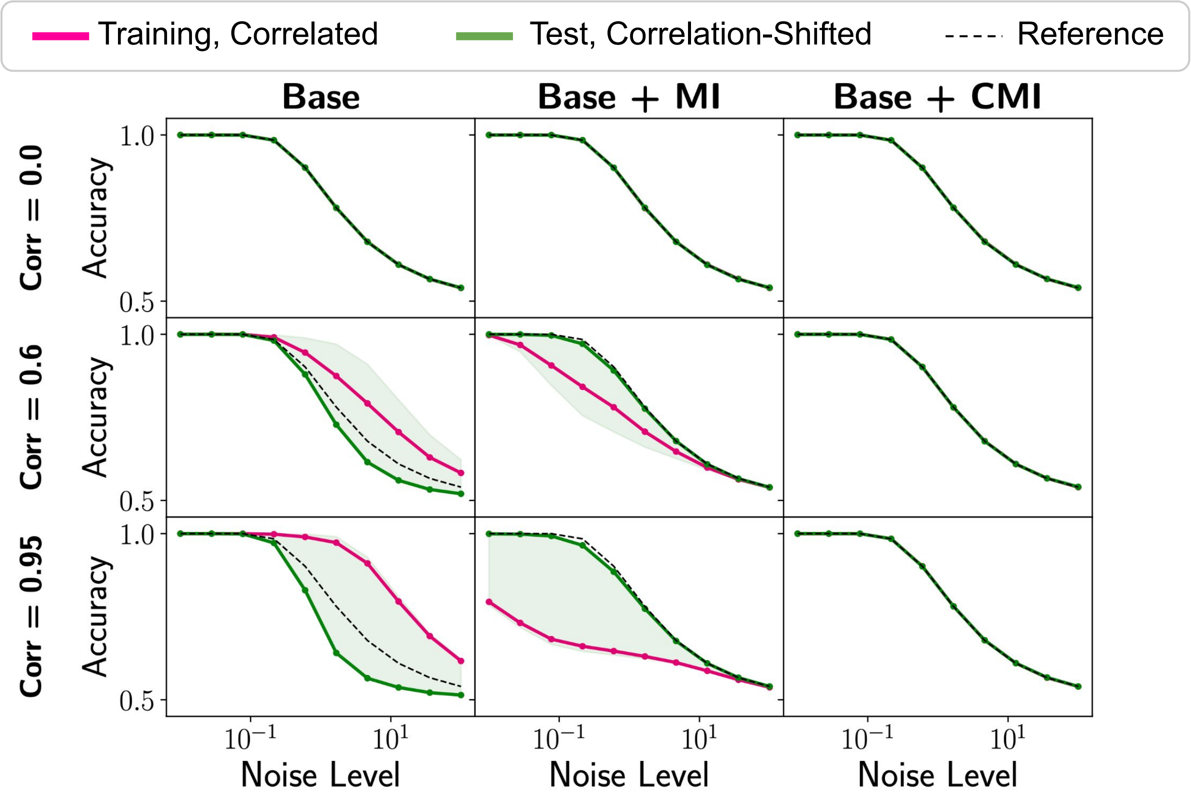
Training Set Correlation and Noise Level

Linear regression: $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$, $\underbrace{\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s})}_{\text{Train correlation}}$, $\underbrace{\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})}_{\text{Noise level}}$



Multi-Attribute Classification Results

- Synthetic *classification task* with multiple attributes. Observed data $\mathbf{x} = \mathbf{A}\mathbf{a} + \mathbf{n}$ is generated from noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$ and correlated source attributes $a_k = \pm 1, \forall k \in \{1, \ldots, K\}$



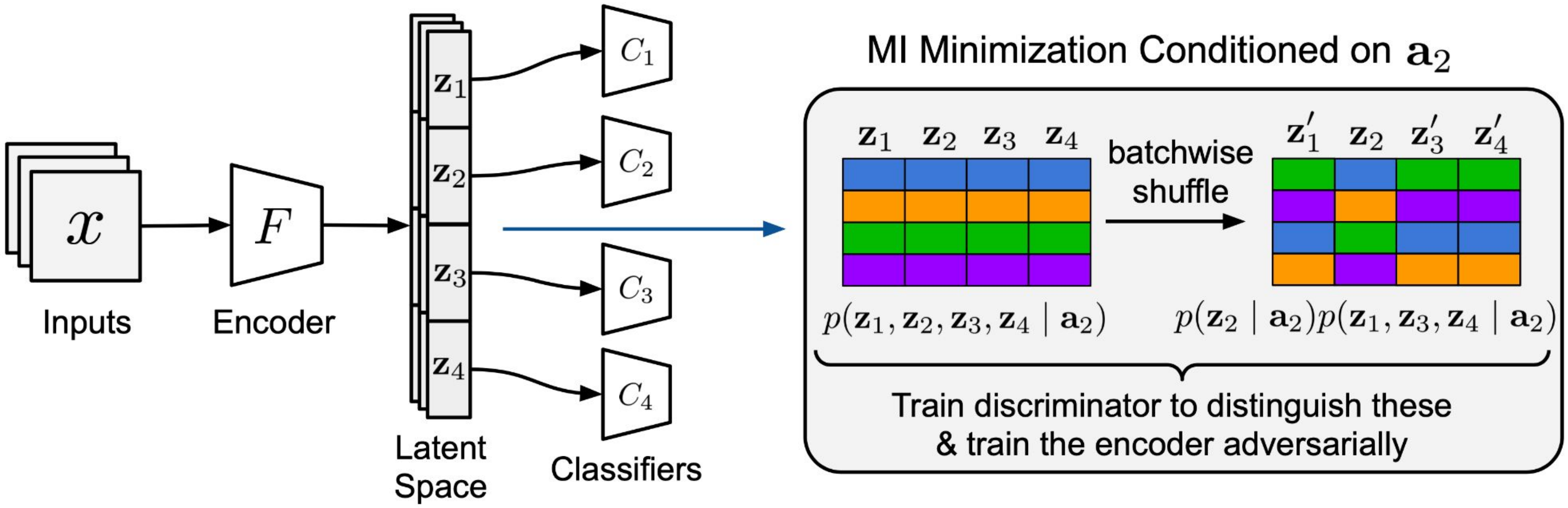
Adversarial Disentanglement

$$I(x; y \mid z) = 0 \quad \text{if} \quad \underbrace{p(x, y \mid z) = p(x \mid z)p(y \mid z)}_{\text{We align these distributions adversarially}}$$

Adversarial Disentanglement

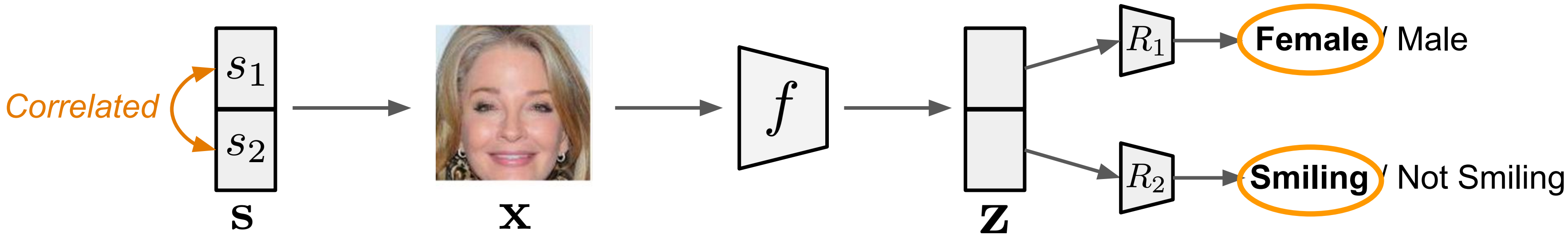
$$I(x; y \mid z) = 0 \quad \text{if} \quad \underbrace{p(x, y \mid z) = p(x \mid z)p(y \mid z)}_{\text{We align these distributions adversarially}}$$

- We use an *adversarial approach to minimize CMI*, based on *batchwise shuffling of latent subspaces*



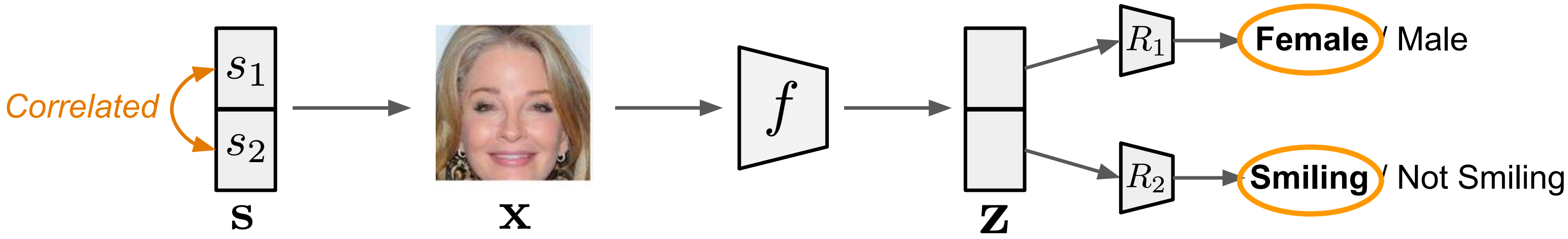
Correlated CelebA

- We used attributes **Female/Male** and **Smiling/Not Smiling** that we know *a priori* are *not causally related*.



Correlated CelebA

- We used attributes **Female/Male** and **Smiling/Not Smiling** that we know *a priori* are *not causally related*.



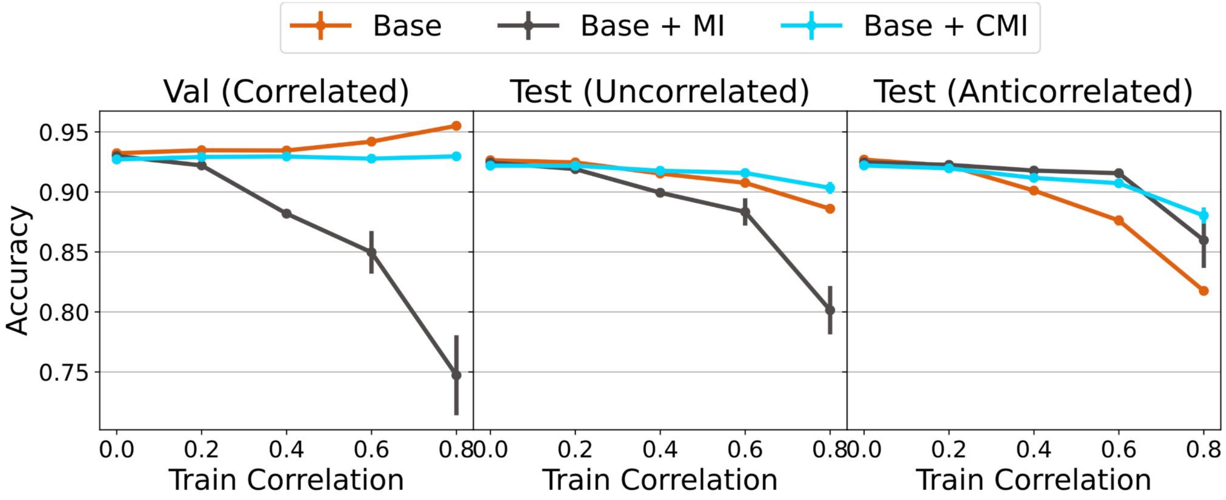
Correlated Train Data

	Smiling? Yes	Smiling? No
Male	0.4	0.1
Female	0.1	0.4

Anti-Correlated Test Data

	Smiling? Yes	Smiling? No
Male	0.1	0.4
Female	0.4	0.1

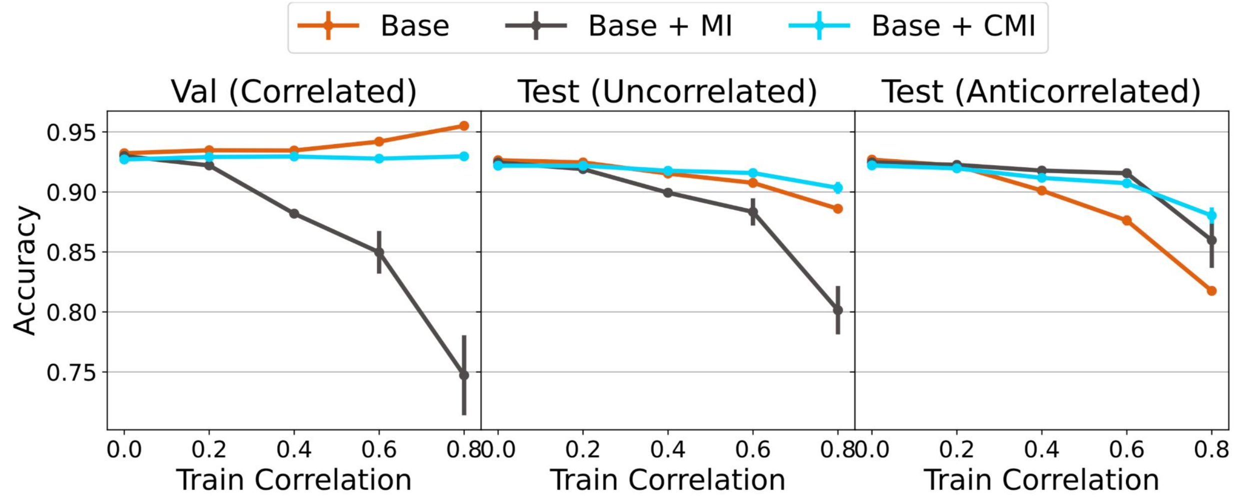
Correlated CelebA



Base: ✔ Performs well on correlated validation data

✖ Performance drops on correlation-shifted test data

Correlated CelebA

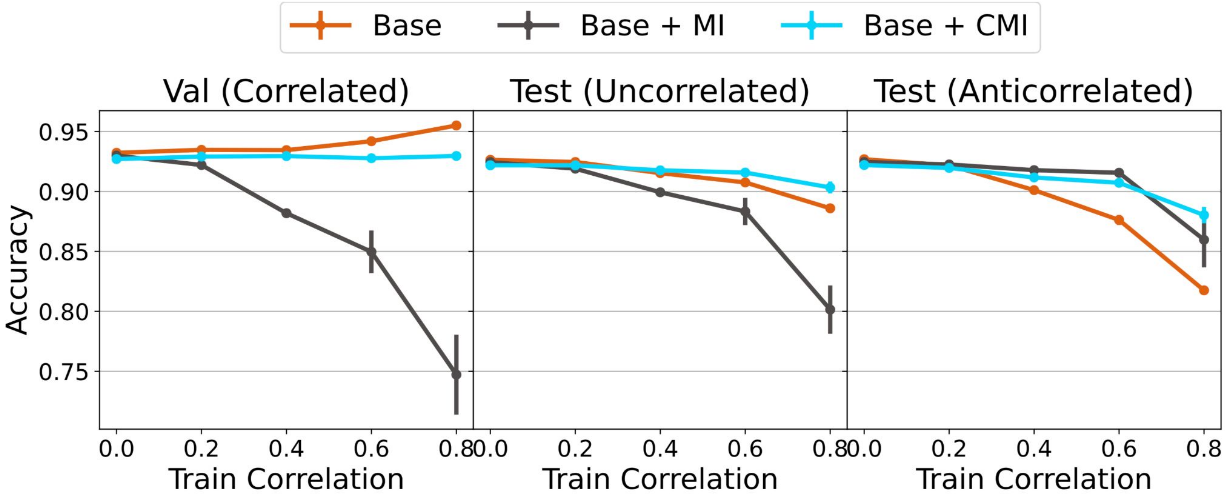


Base: ✔ Performs well on correlated validation data

✖ Performance drops on correlation-shifted test data

Base + MI: ✖ Fails to model the in-distribution, correlated validation data

Correlated CelebA



Base:
- ✔ Performs well on correlated validation data
- ✖ Performance drops on correlation-shifted test data

Base + MI:
- ✖ Fails to model the in-distribution, correlated validation data

Base + CMI:
- ✔ Achieves the most consistent performance across correlated and correlation-shifted datasets

  Has a *larger effect for stronger correlations*, but *does not harm performance for low correlation* strengths.

Correlated CelebA

- Even without constructing correlated datasets from CelebA by subsampling, we can see *detrimental effects due to correlations* if we evaluate performance on *subpopulations* of correlated (in-distribution) data.
- *Some combinations of attributes are more common* than others, and models that exploit these correlations for prediction may treat rare combinations unfairly
  - **Base** fails on rare attribute combinations
  - **Base + MI** does not succeed even on the common attribute combinations
  - ***Base + CMI*** *improves accuracy on rare attribute combinations*

	Common Combinations		**Rare Combinations**	
	Female + Non-Smiling	Male + Smiling	Female + Smiling	Male + Non-Smiling
Base	**4%**	**5%**	33%	49%
Base + MI	24%	28%	**11%**	26%
Base + CMI	9%	9%	19%	**25%**

Summary

- Learning robust disentangled representations can be challenging in the presence of correlations, even with full supervision
- We introduced and motivated *subspace independence conditioned on available attributes* as the *correct objective for disentanglement* in correlated settings.
- We described an *algorithm* to achieve conditional independence for general classification tasks.
- We showed that CMI minimization *improves robustness to correlation shifts* on both synthetic tasks and real-world datasets.

Thank you!