Overlooked Implications of the Reconstruction Loss for VAE Disentanglement

IJCAI 2023 Paper Presentation

Nathan Michlo, Richard Klein, Steven James

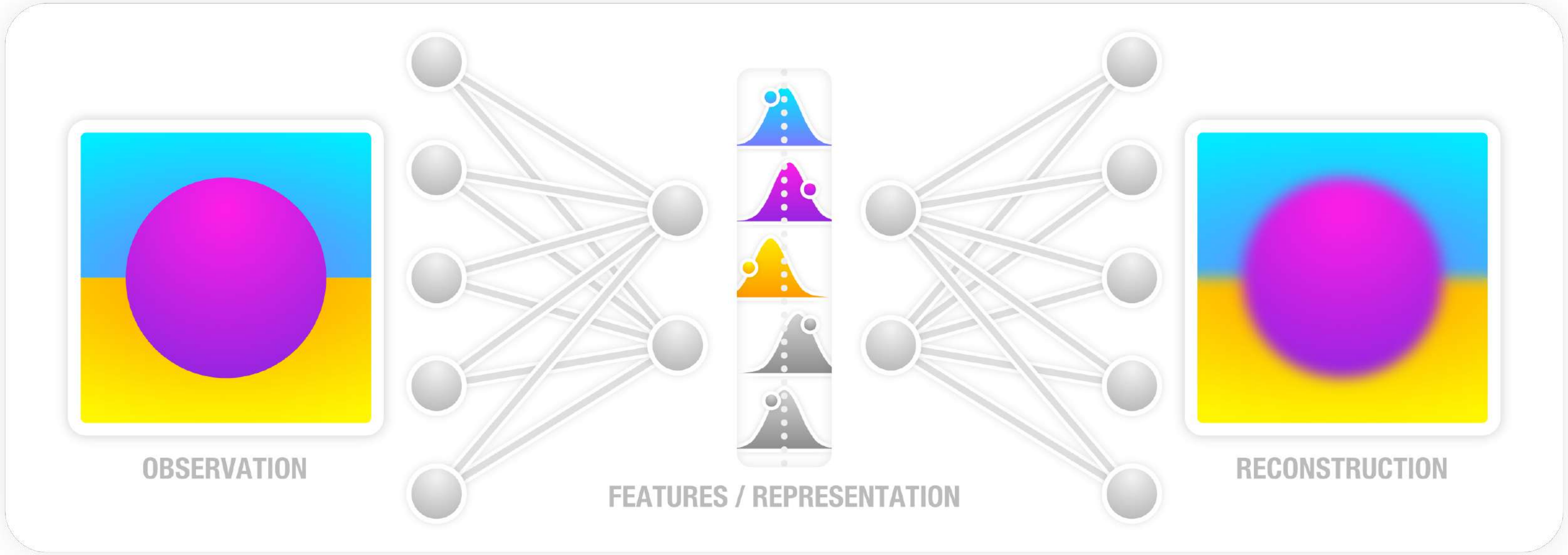
22 Aug 2023

PRIME LAB & RAIL LAB

Background

Variational Auto-Encoders

- Variational Auto-Encoders (VAEs) learn to **compress data** by **reconstructing** the inputs.
- VAEs **encode distributions**, which are sampled from during training.



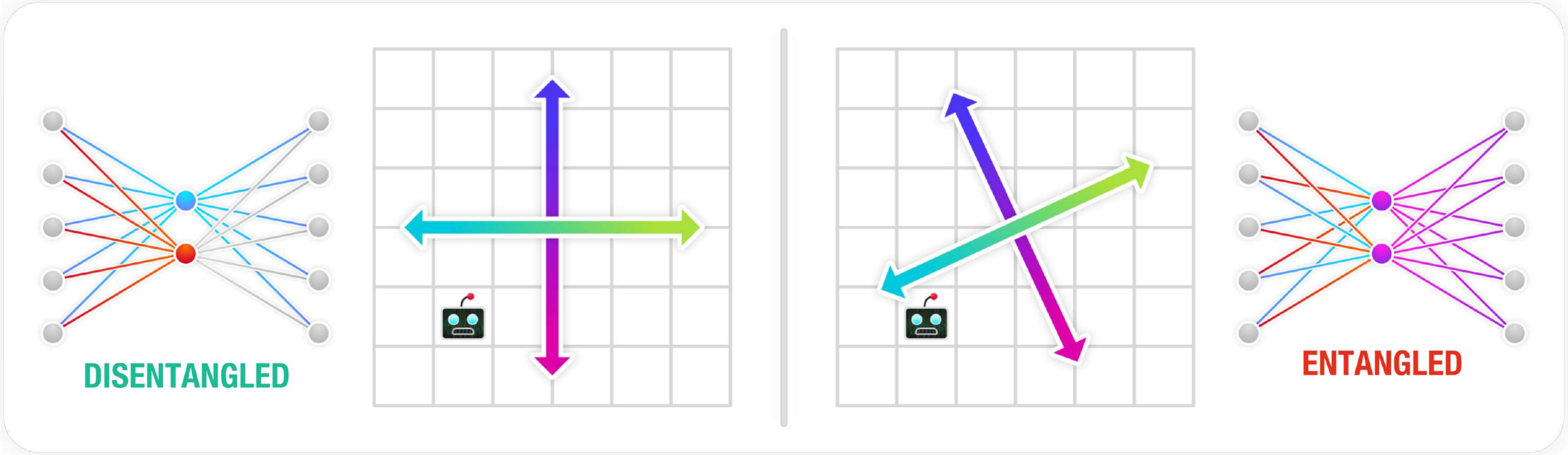
$$L_{\beta\mathrm{VAE}} = L_{\mathrm{rec}} + L_{\mathrm{reg}}$$

Regularisation is usually the focus of disentanglement research

Background

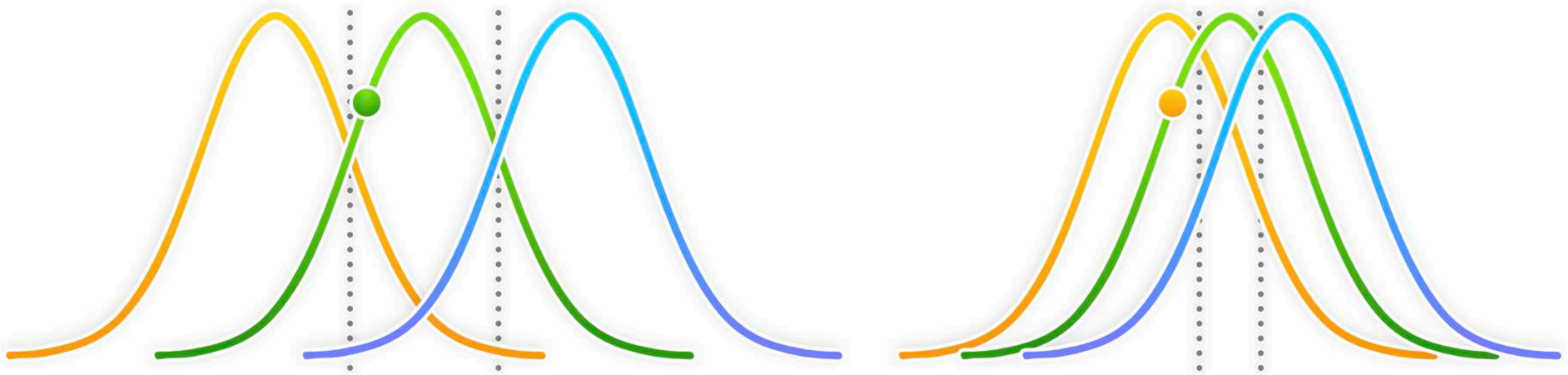
Disentanglement (with unsupervised VAEs)

- VAEs are known to produce human interpretable or **disentangled** representations from data.
- VAEs may fail and produce **entangled** representations.
  - **Why?**



Background

Random sampling reorganises VAE embeddings



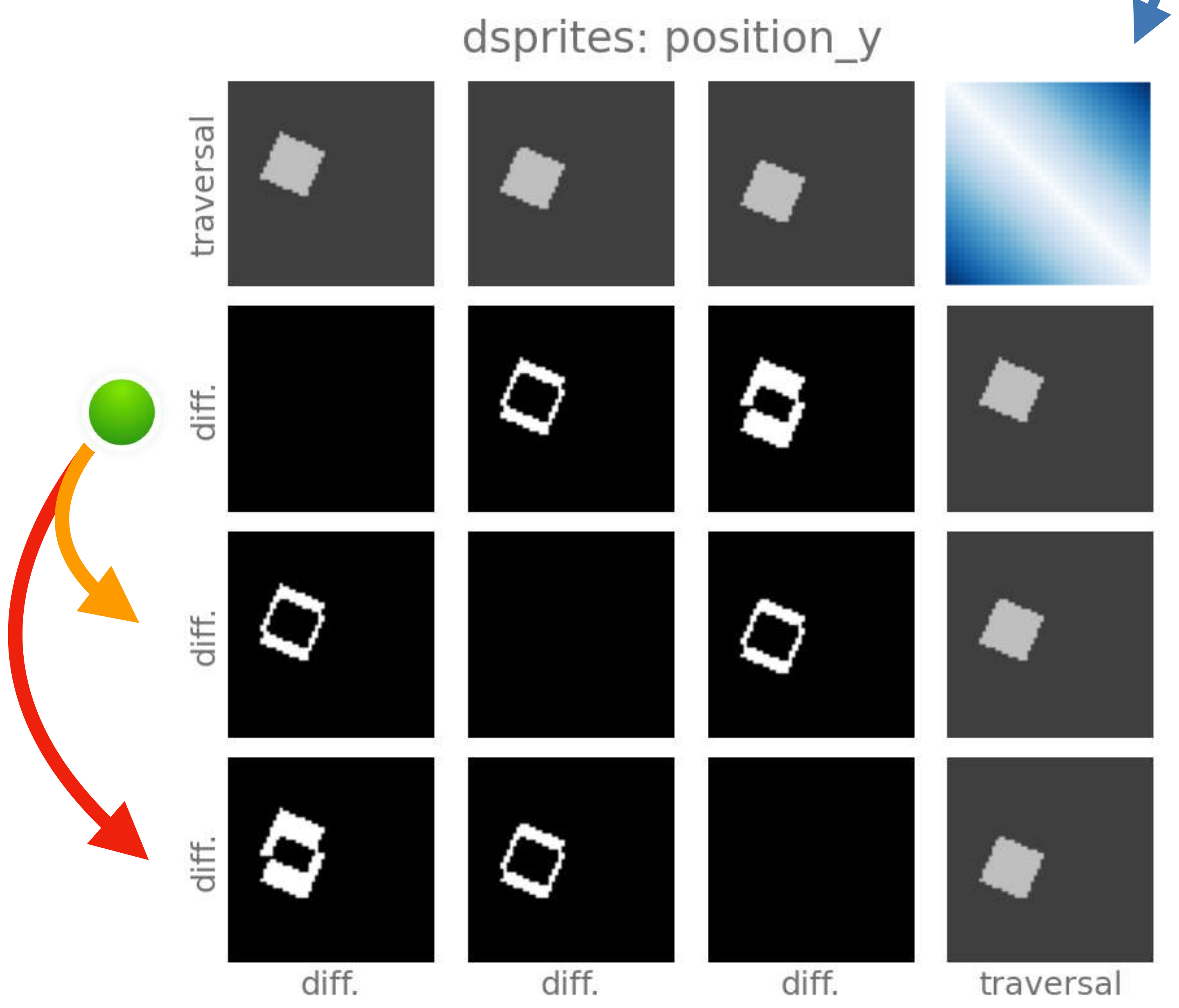
a: **Correct Distribution** b: **Confused Distribution**

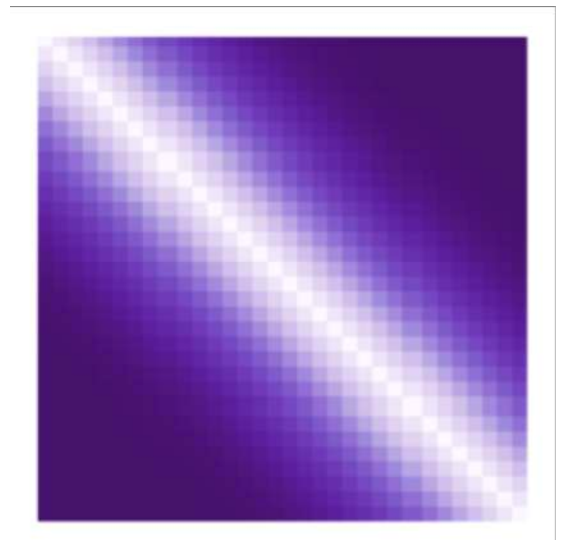
A VAE prefers to minimise sampling errors by placing **embeddings** close together that it also perceives as **close in the data space**

We use this idea to measure the similarity of observations directly using the reconstruction loss "**Perceived distances**"

$$\mathrm{d}_{\mathrm{pcv}}(\bullet, \bullet) = \lim_{\hat{\boldsymbol{x}} \to \boldsymbol{x}} \mathcal{L}_{\mathrm{rec}}(\bullet, \bullet)$$

Ground-truth distance matrix over factor traversals

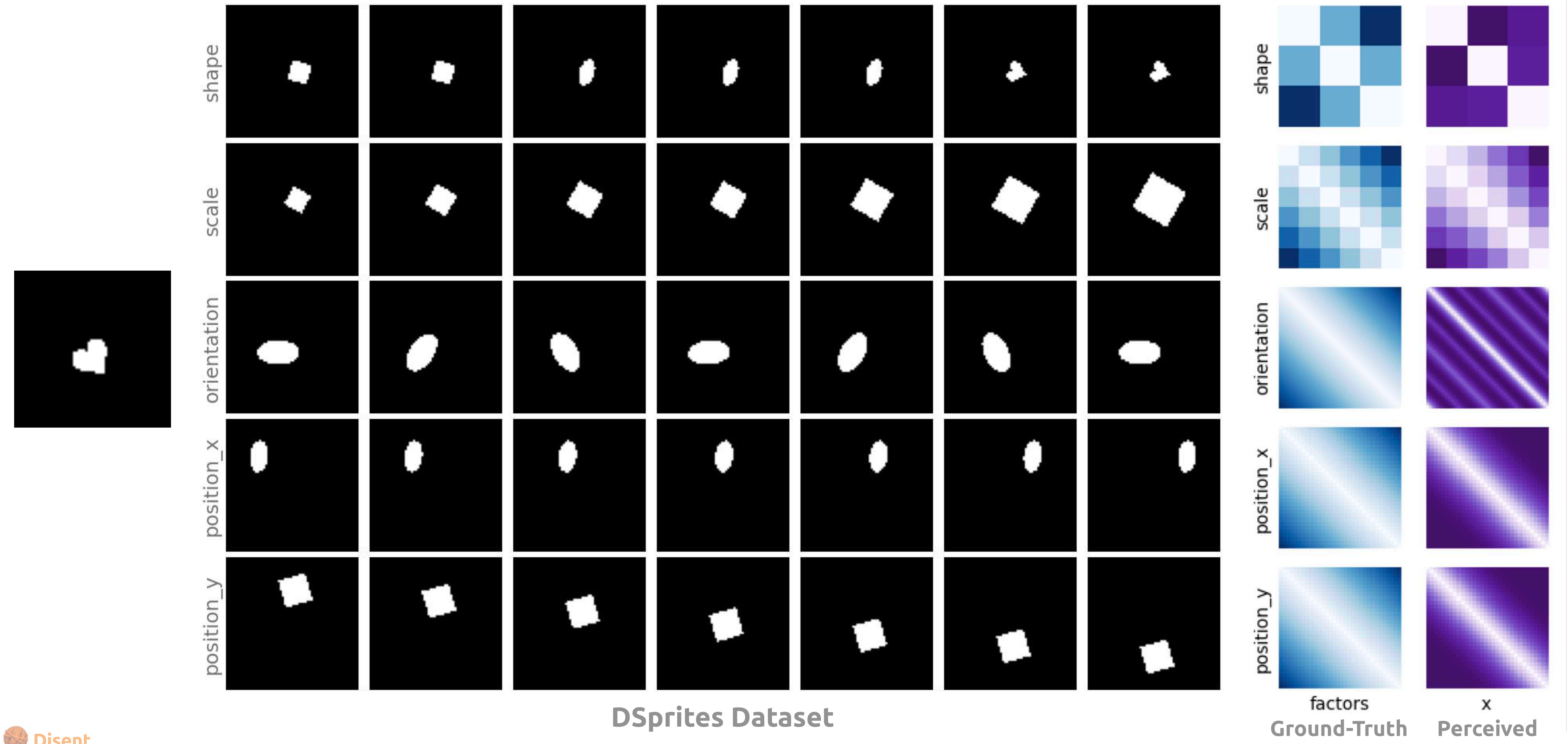




Perceived distance matrix between obs. over factor traversals

Characterising Existing Datasets

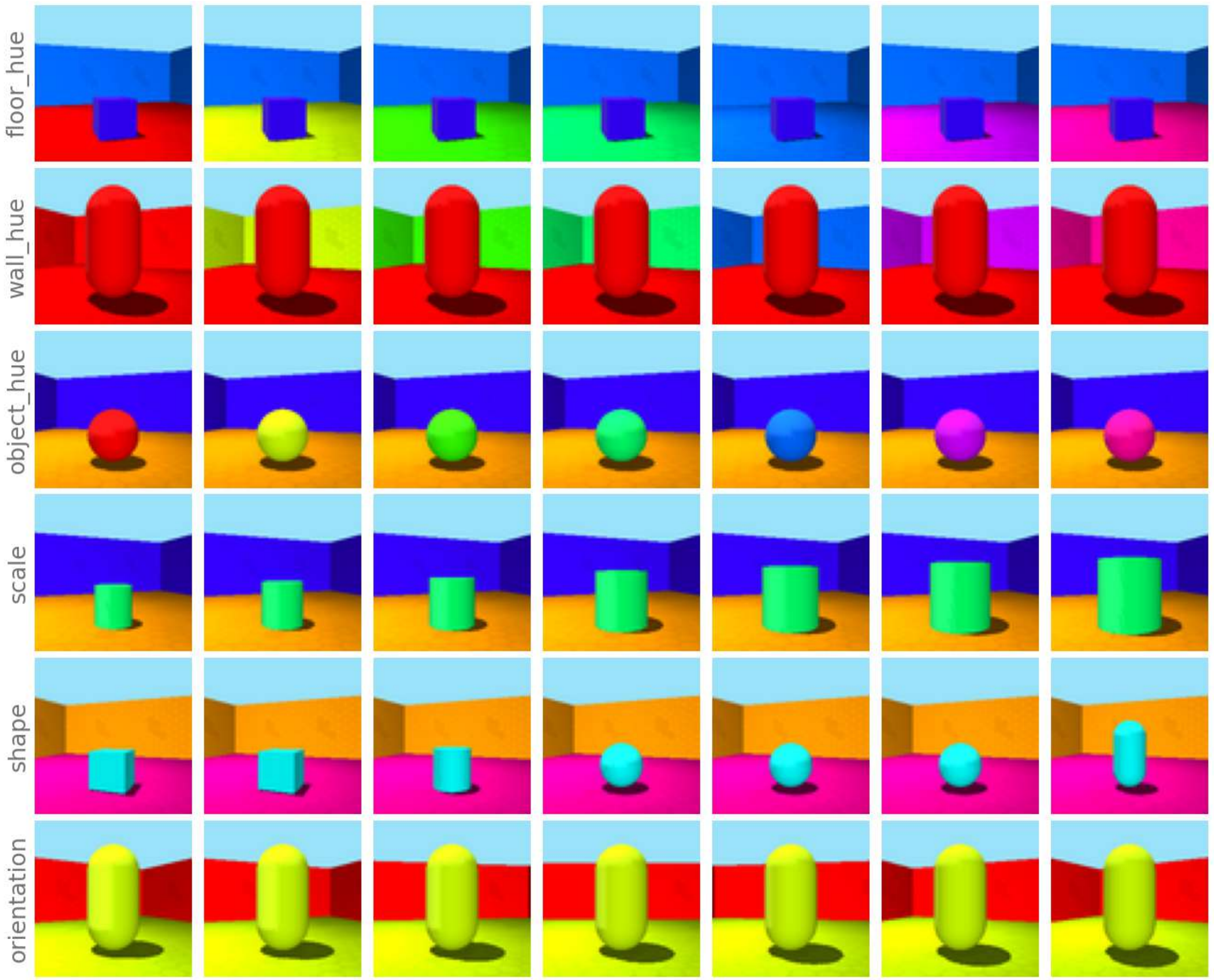
Ground-truth distances usually correspond to VAE perceived distances



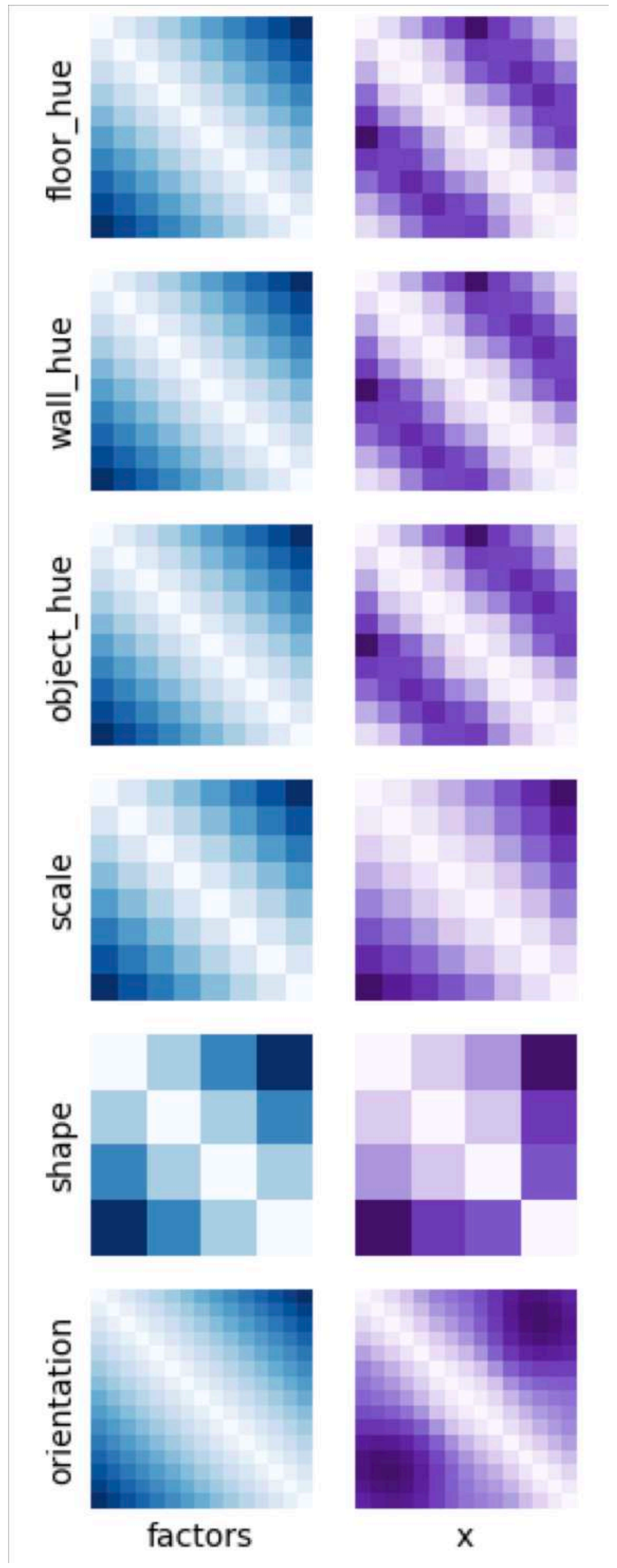
Disent

Characterising Existing Datasets

Ground-truth distances usually correspond to VAE perceived distances



Shapes3D Dataset

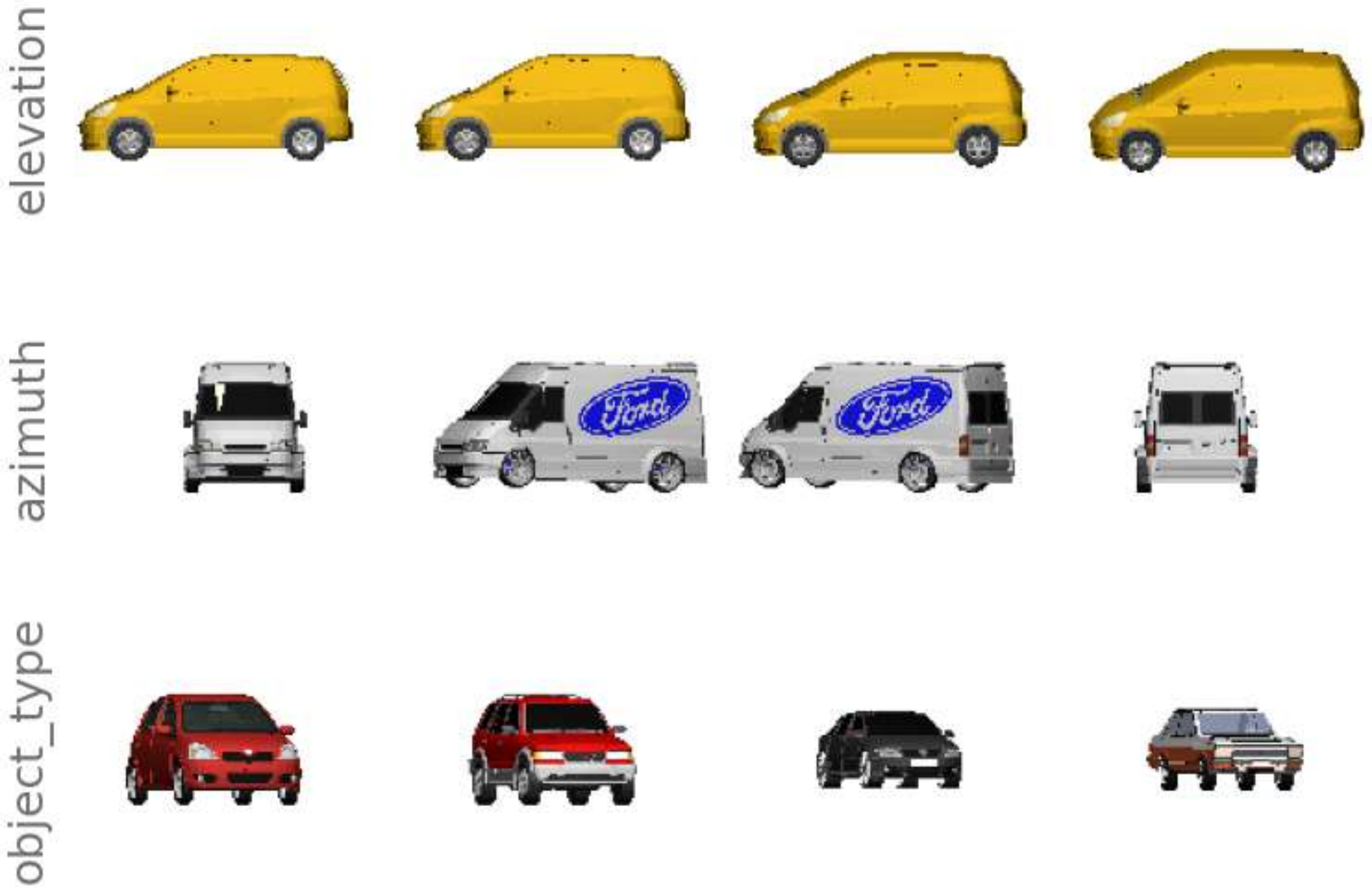


Ground-Truth Perceived

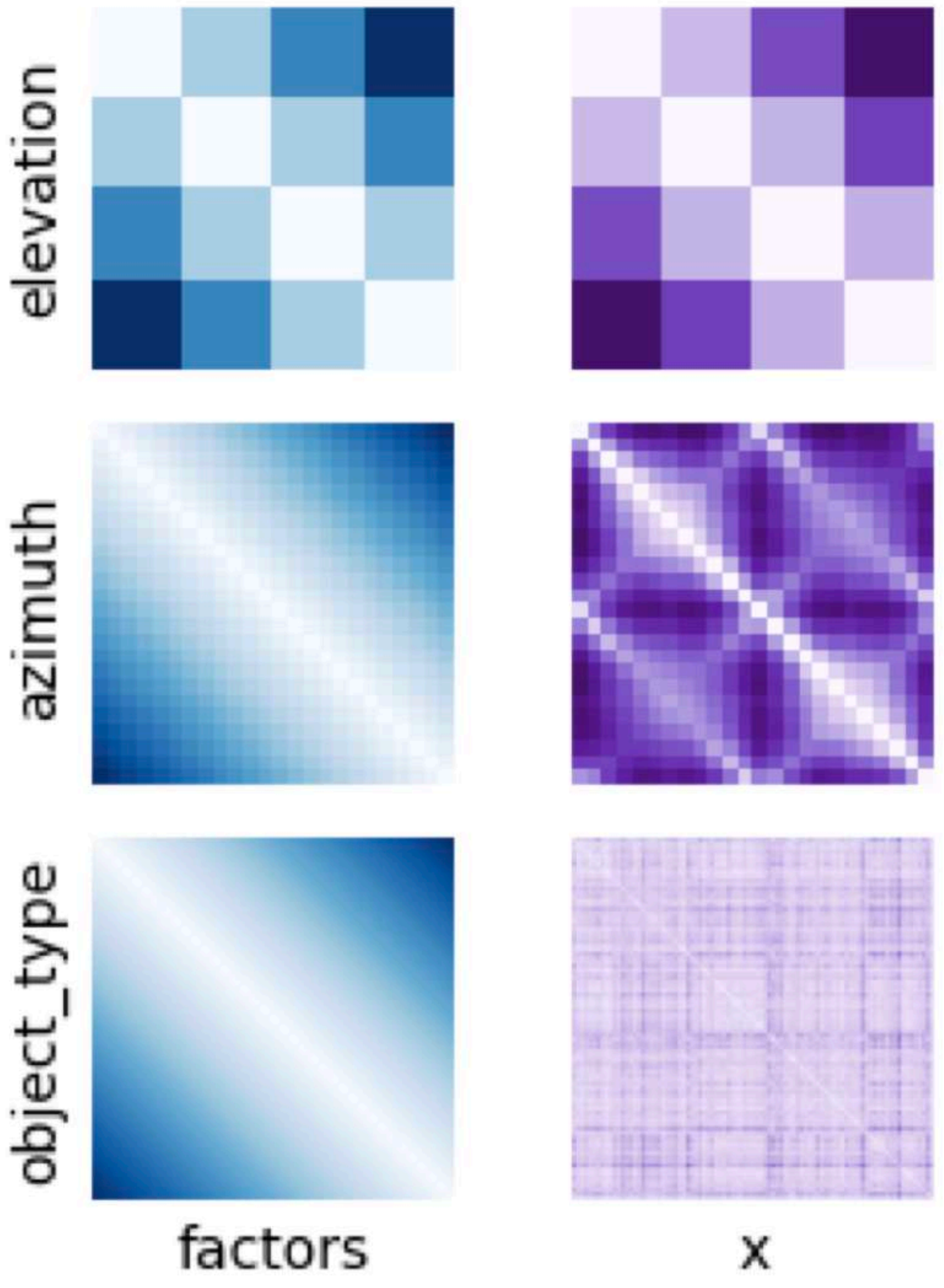
- RGB has 3 dims, unintuitive to represent in 1 factor
- Colour is circular in nature, can be learnt from any starting point

Characterising Existing Datasets

Ground-truth distances usually correspond to VAE perceived distances



Cars3D Dataset



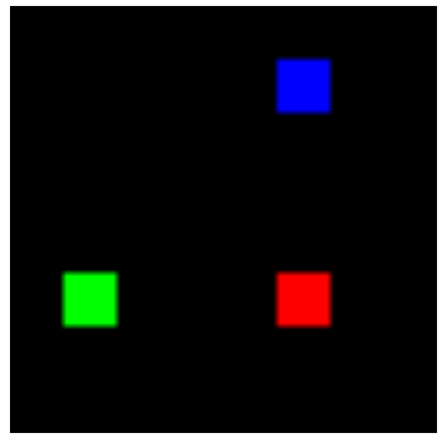
Ground-Truth   Perceived

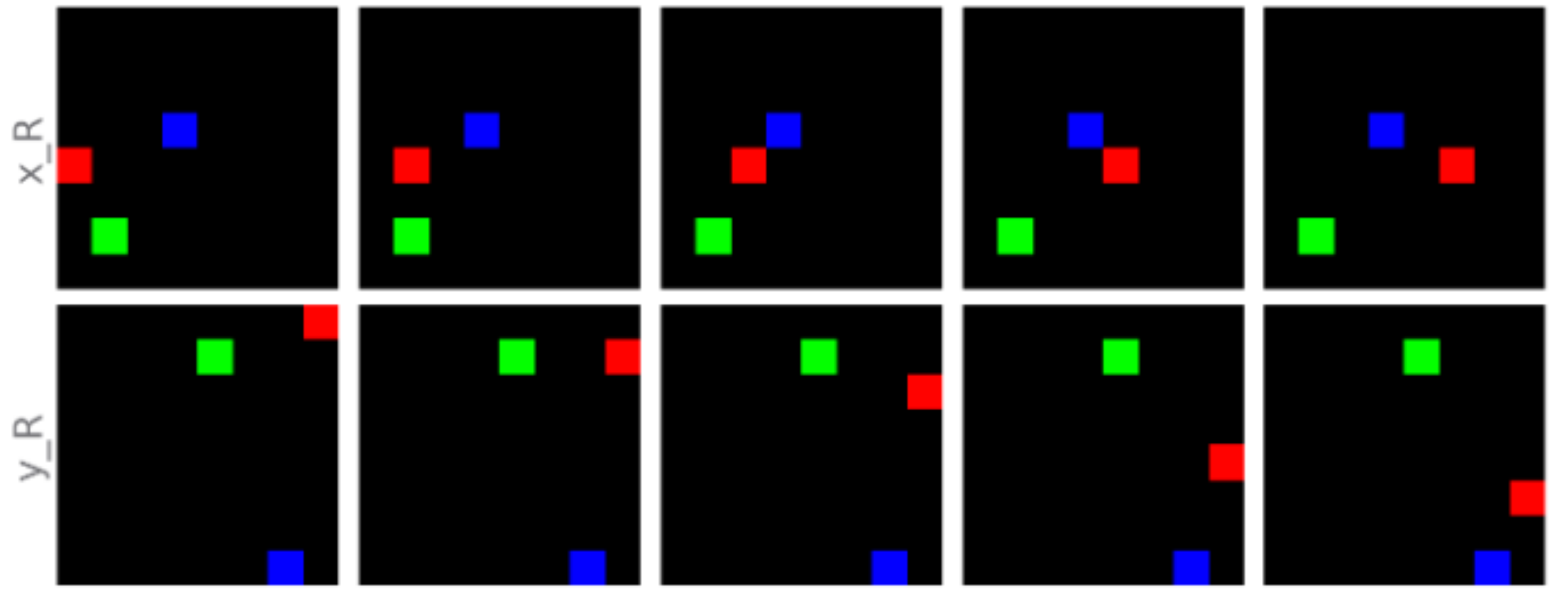
- **Factors are easier to decompose into sub-factors**

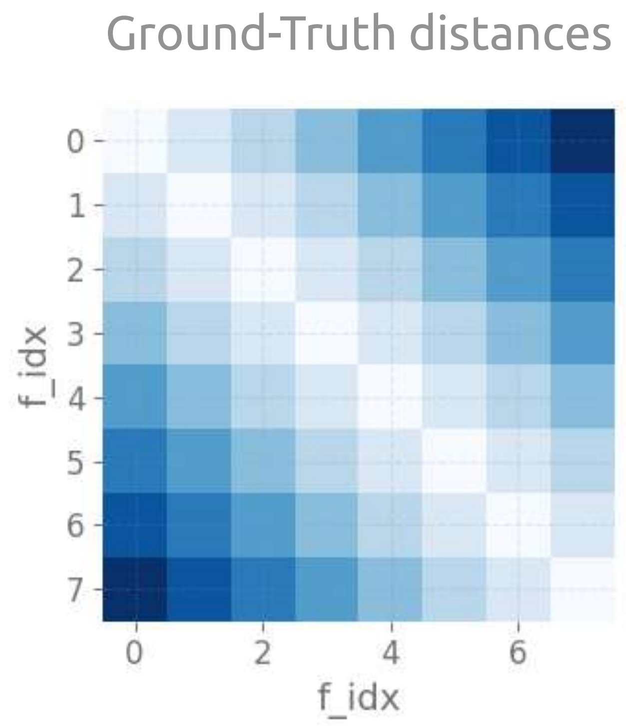
Adversarial Dataset

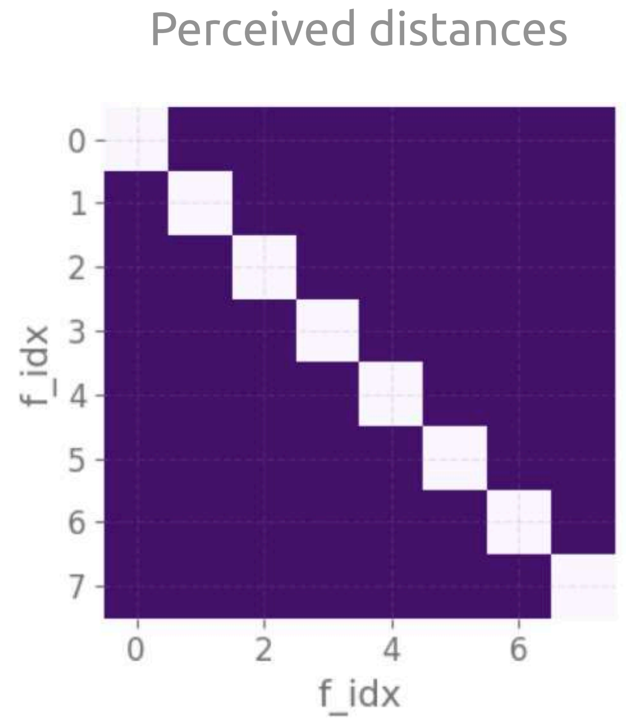
XYSquares Example - Adversarial for pixel-wise reconstruction losses

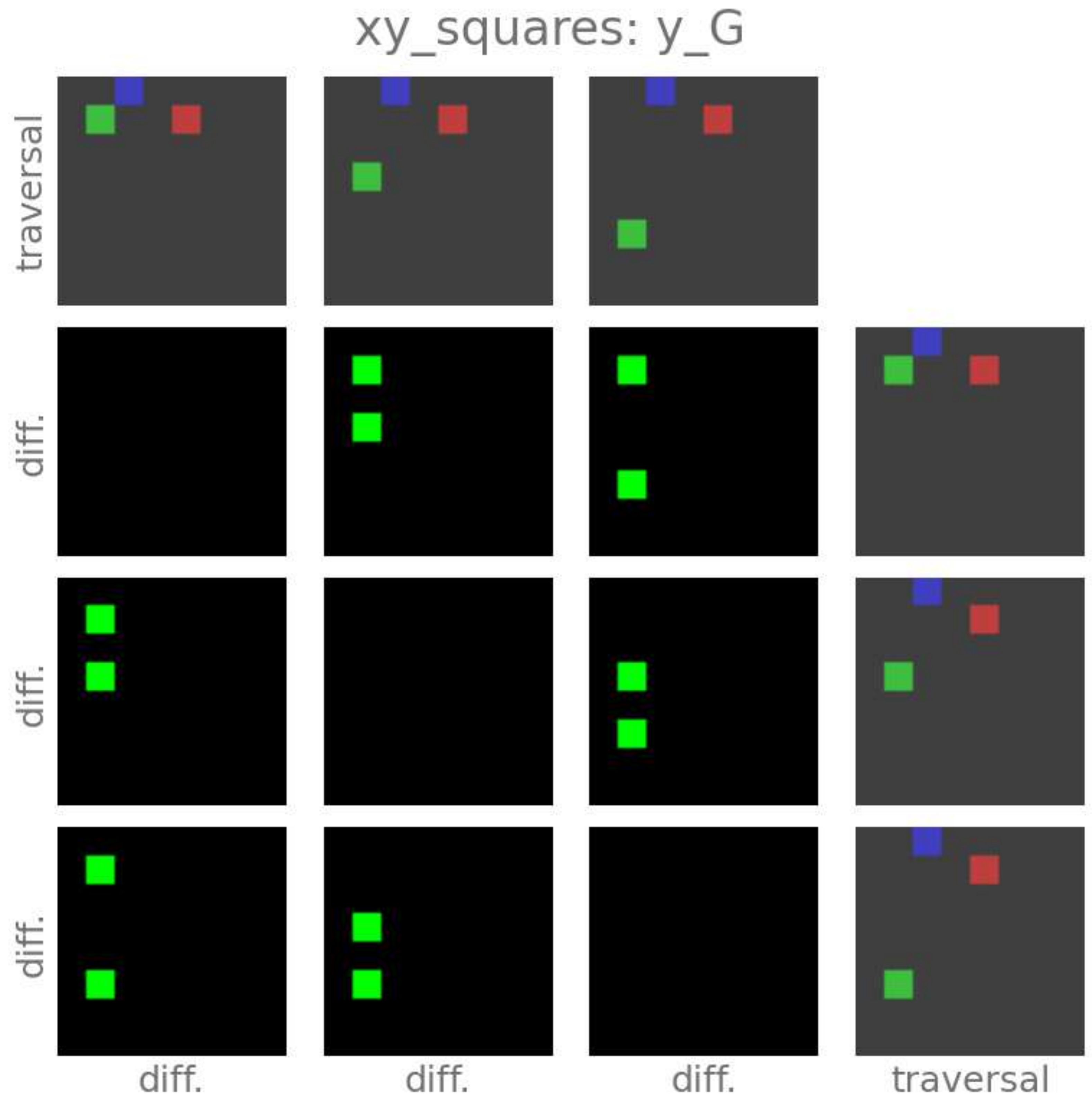
- Design example dataset with **constant overlap** along factor traversals
  - Cannot minimise recon. error due to sampling, **no ordering**
- 8x8 grid with **x** & **y** positional factors







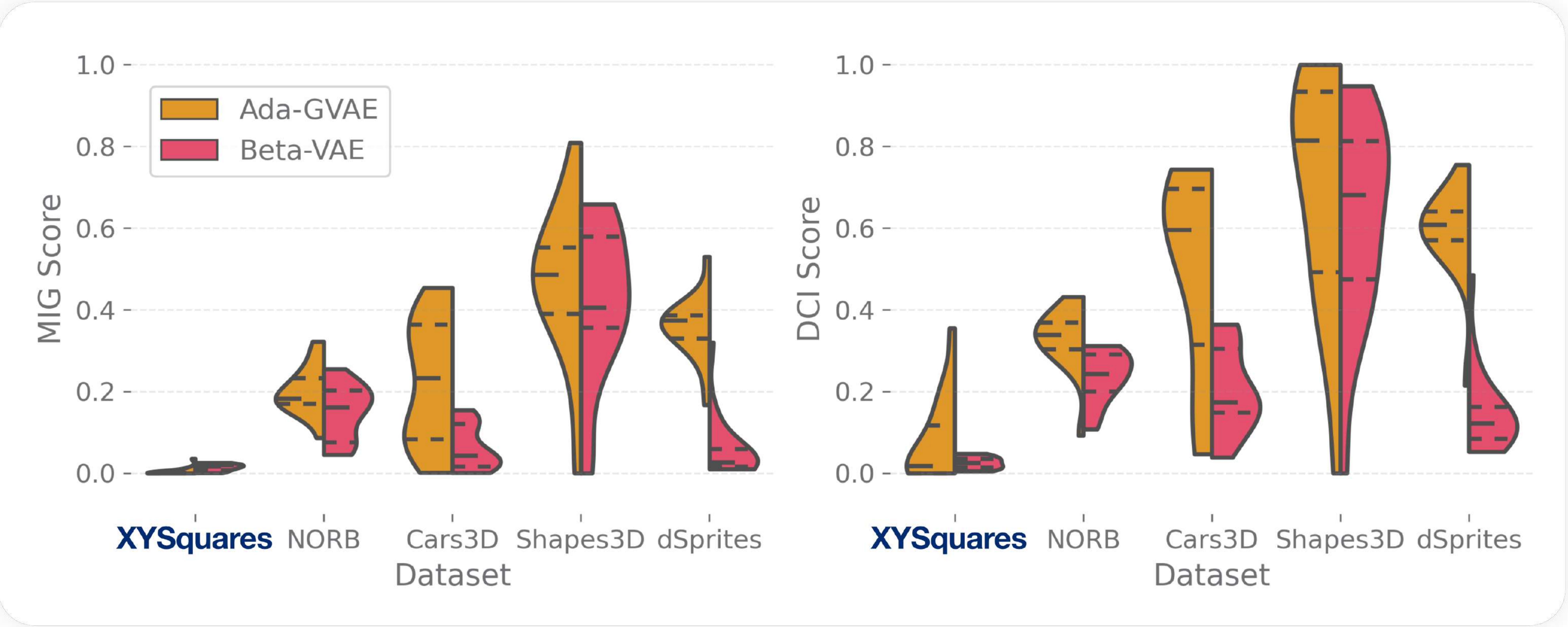




Same amount of overlap between each obs. in a factor traversal

Adversarial Dataset

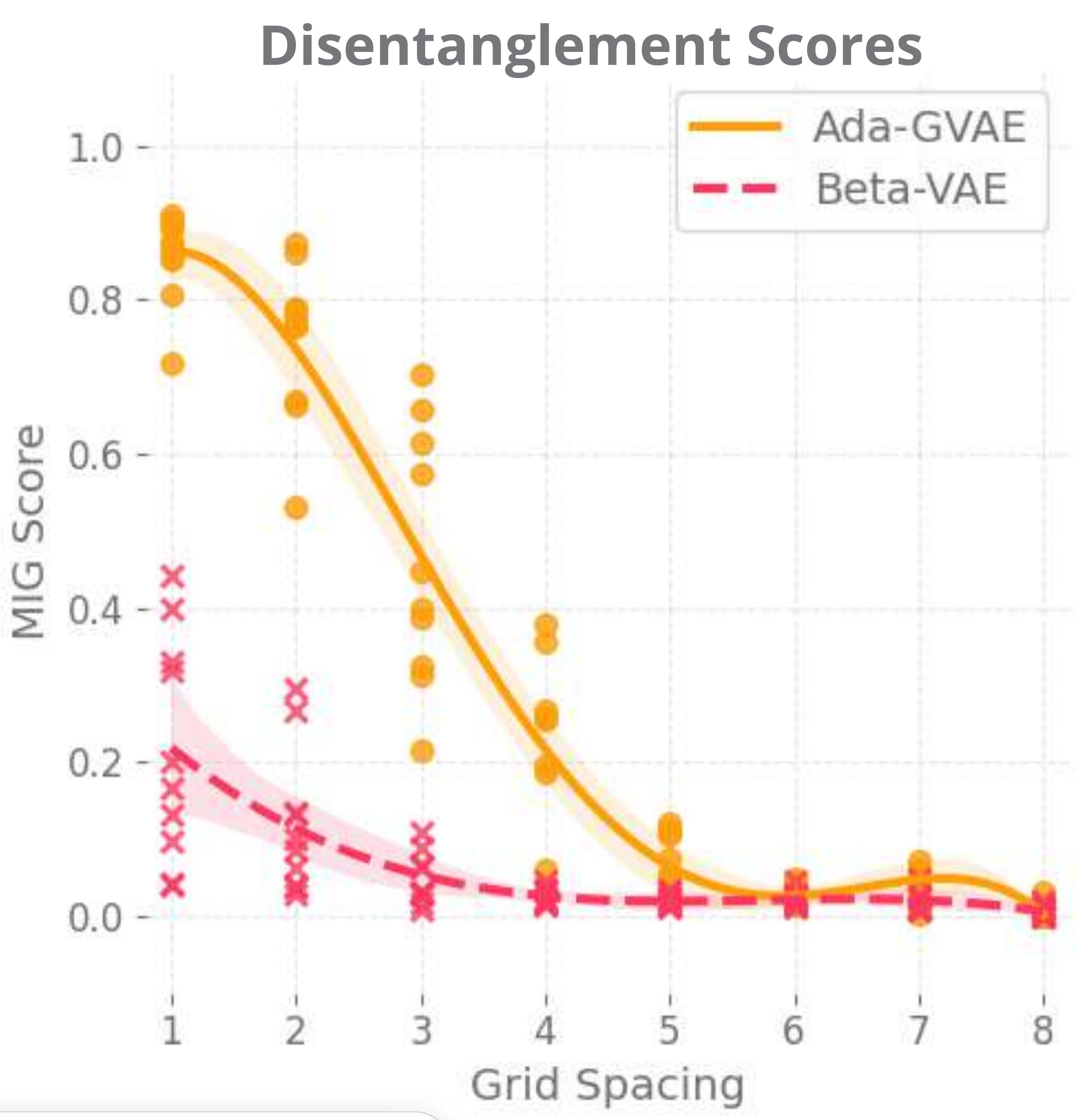
XYSquares Example - Disentanglement compared to benchmark datasets



Disent

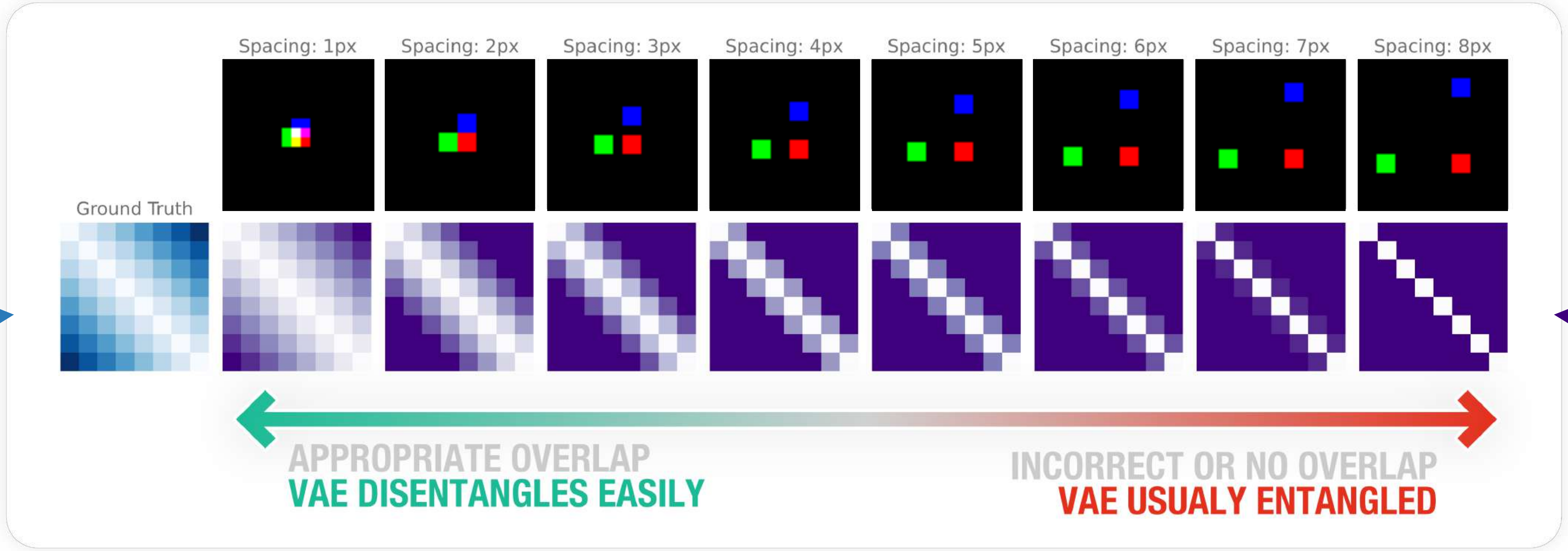
Re-enabling Disentanglement

XYSquares example - Adjusting the **data**



- Train VAEs over XYSquares data with varying spacing which **changes perceived distances**.
  - **decrease spacing**, appropriate overlap, **correlated distances** between observations, **better disentanglement**
  - **increase spacing**, no overlap, **constant distances** between observations, **worse disentanglement**

Ground-truth distance matrices along factors



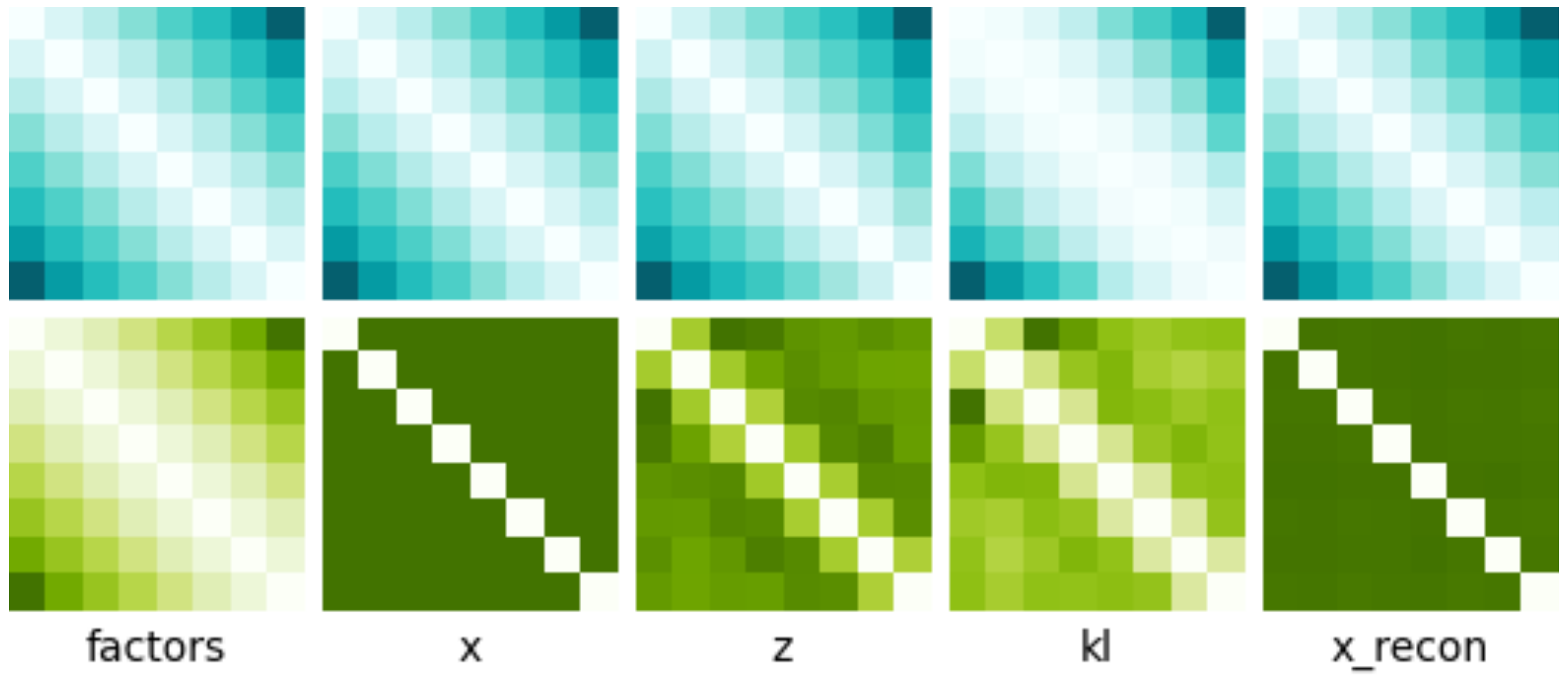
Perceived distance matrices along factors

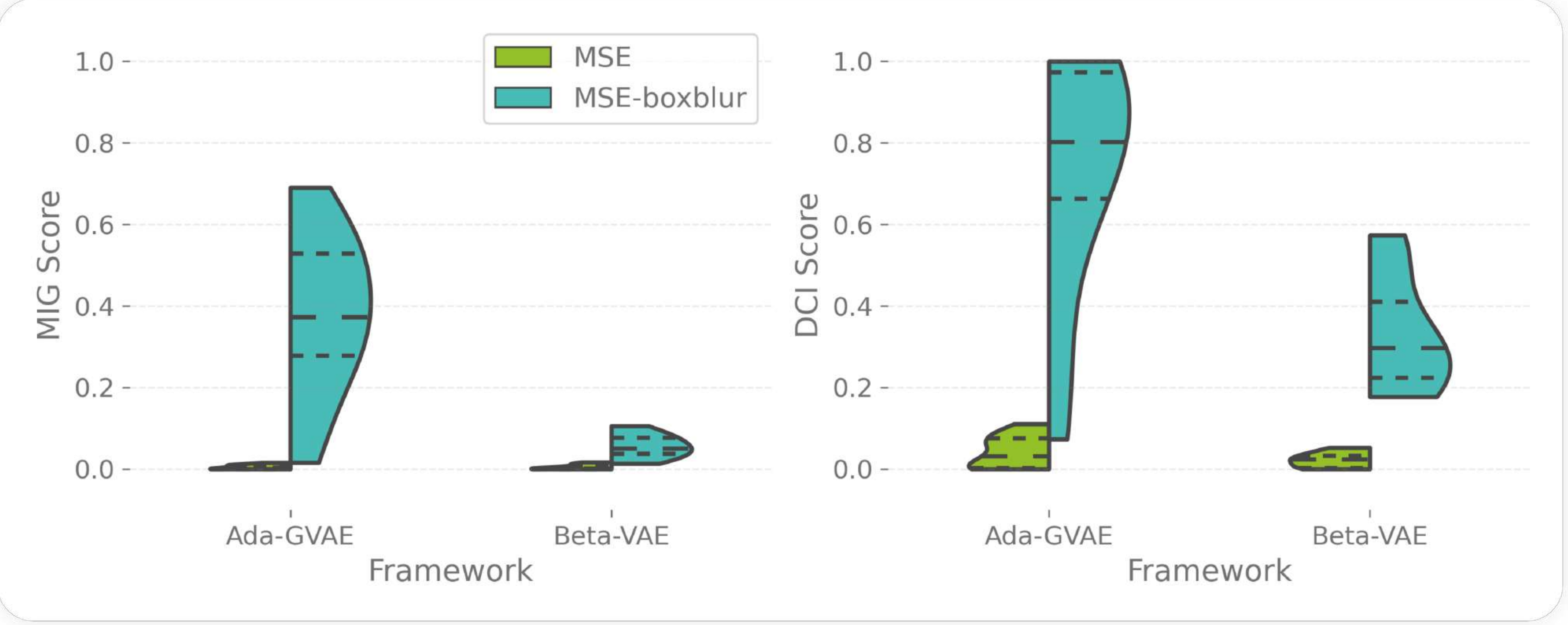
Disent

Re-enabling Disentanglement

XYSquares example - Adjusting the **loss**

- If perceived distances once again correspond to ground-truth distances, disentanglement can occur.
- **Adjust the reconstruction loss** to **re-enable disentanglement**. An example that is appropriate for XYSquares is adding a box blur augment to data.





Disent

Conclusion

And considerations for unsupervised disentanglement research

- **Disentanglement in benchmarks is largely accidental**
  - Fundamental characteristics of existing benchmark datasets encourage VAEs to learn disentangled representations.
  - New benchmark datasets are required.
- **Disentanglement depends on the data and reconstruction loss too**.
  - Unsupervised disentanglement is ultimately not from special regulariser and algorithmic choices.
- **Disentanglement is subjective**
  - e.g. RGB, HSV or categorical representations for colours, binary or continuous encodings for positions, split or combined factors.
  - There are infinitely many datasets with infinitely many choices of what constitutes their ground-truth factors.
  - Supervision ultimately required

SCAN FOR PAPER & RESOURCES

or visit github.com/nmichlo

∞. The End