

# Overlooked Implications of the Reconstruction Loss for VAE Disentanglement

IJCAI 2023 Paper Presentation

Nathan Michlo, Richard Klein, Steven James

22 Aug 2023



&



# Overview

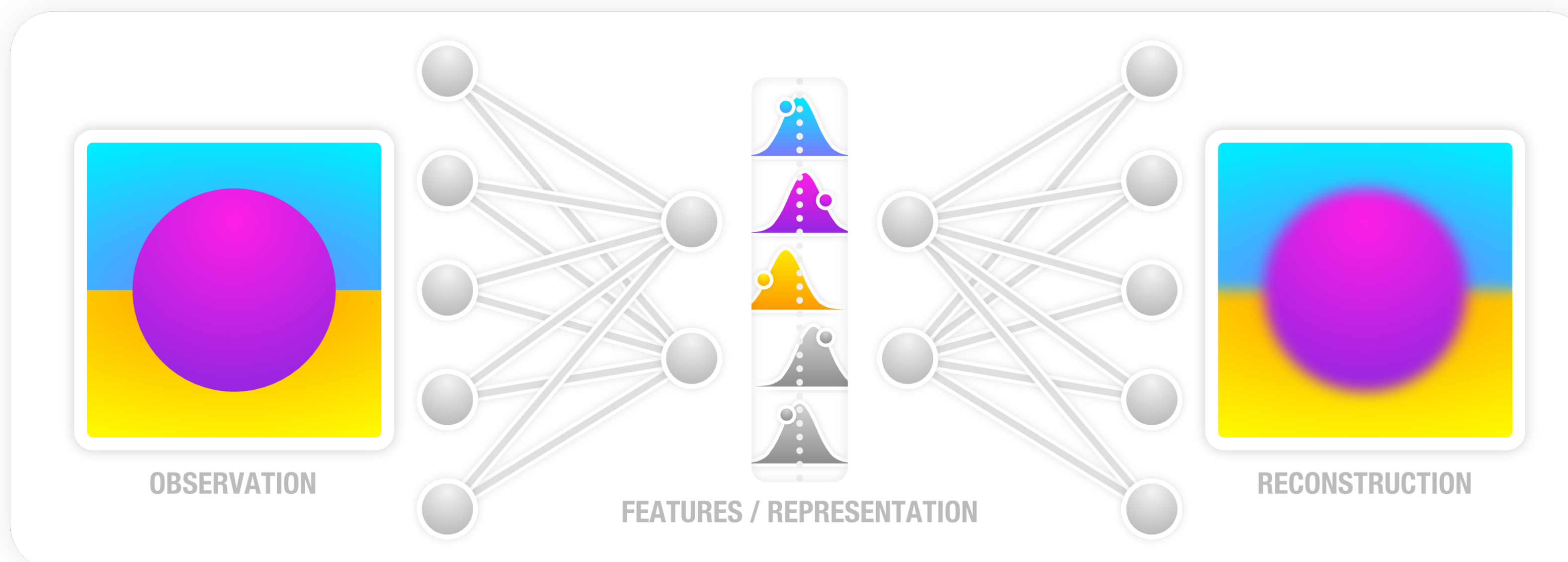
## Paper outline and contributions

- **Visualisation approaches** for VAE disentanglement benchmark datasets in terms of the reconstruction loss.
- We show VAE **perceived distances** in benchmark datasets unintentionally **align with their generating factors**.
- We give an **example of adversarial data** and show that SOTA unsupervised disentanglement methods fail.
- We **provide an example solution** to the adversarial data problem which changes the reconstruction loss., re-enabling disentanglement.
- We contribute **disent**, a general PyTorch framework for disentanglement with common models metrics and datasets.

# 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.



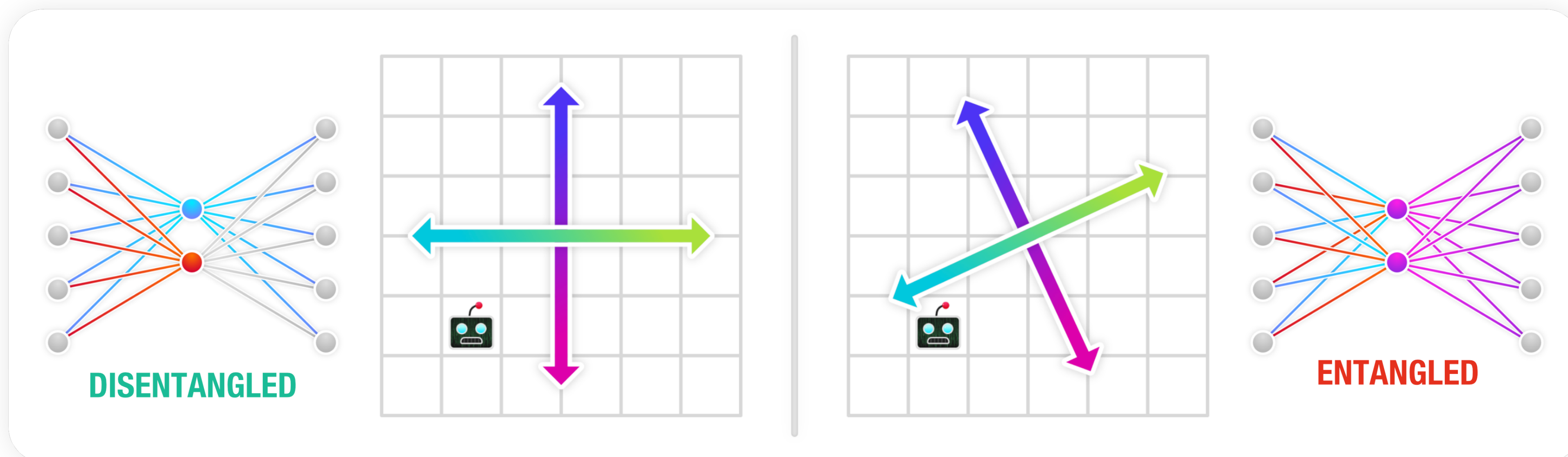
$$L_{\beta\text{VAE}} = L_{\text{rec}} + L_{\text{reg}}$$

← Regularisation is usually the focus of disentanglement research

# Background

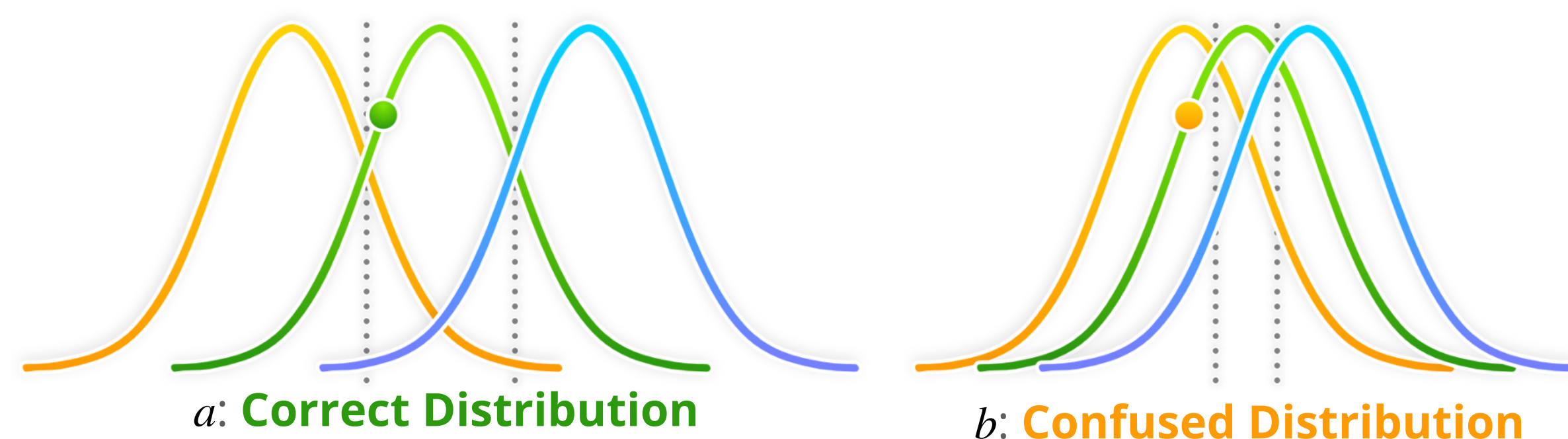
## Disentanglement (with unsupervised VAEs)

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



# Perceived Distances

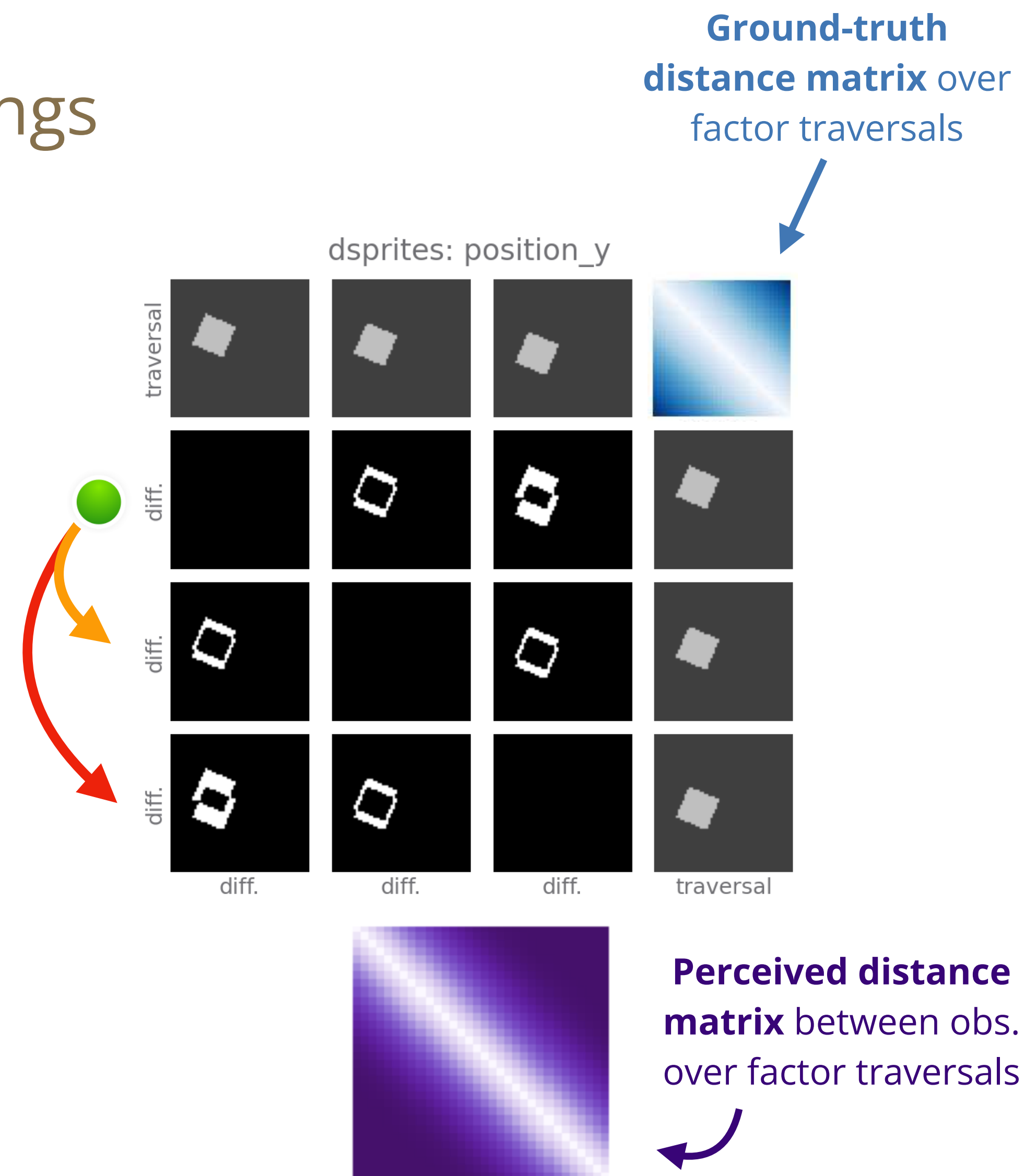
Random sampling reorganises VAE embeddings



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

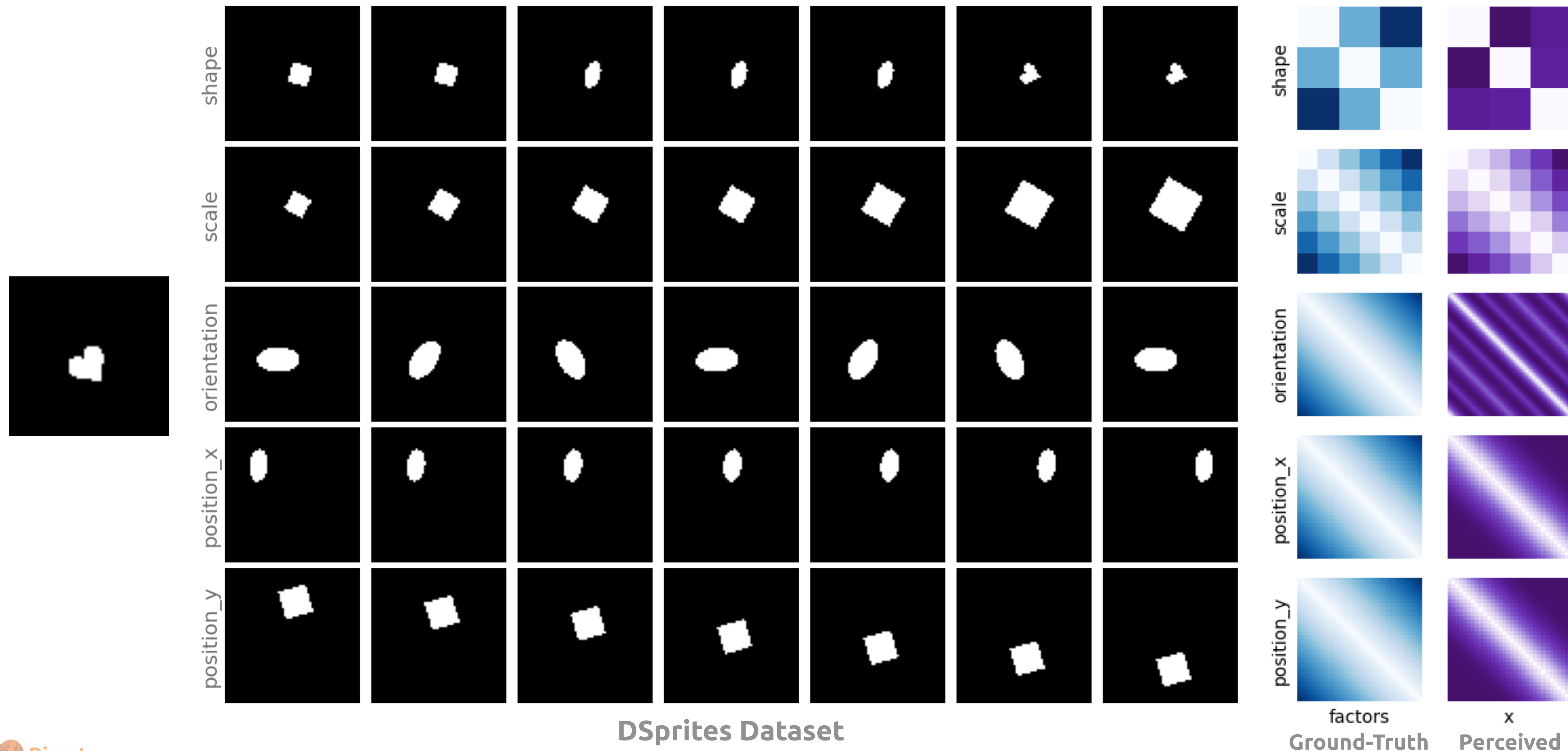
**We use this idea** to measure the similarity of observations directly using the reconstruction loss  
“**Perceived distances**”

$$d_{\text{pcv}}(\text{img}_1, \text{img}_2) = \lim_{\hat{x} \rightarrow x} \mathcal{L}_{\text{rec}}(\text{img}_1, \text{img}_2)$$



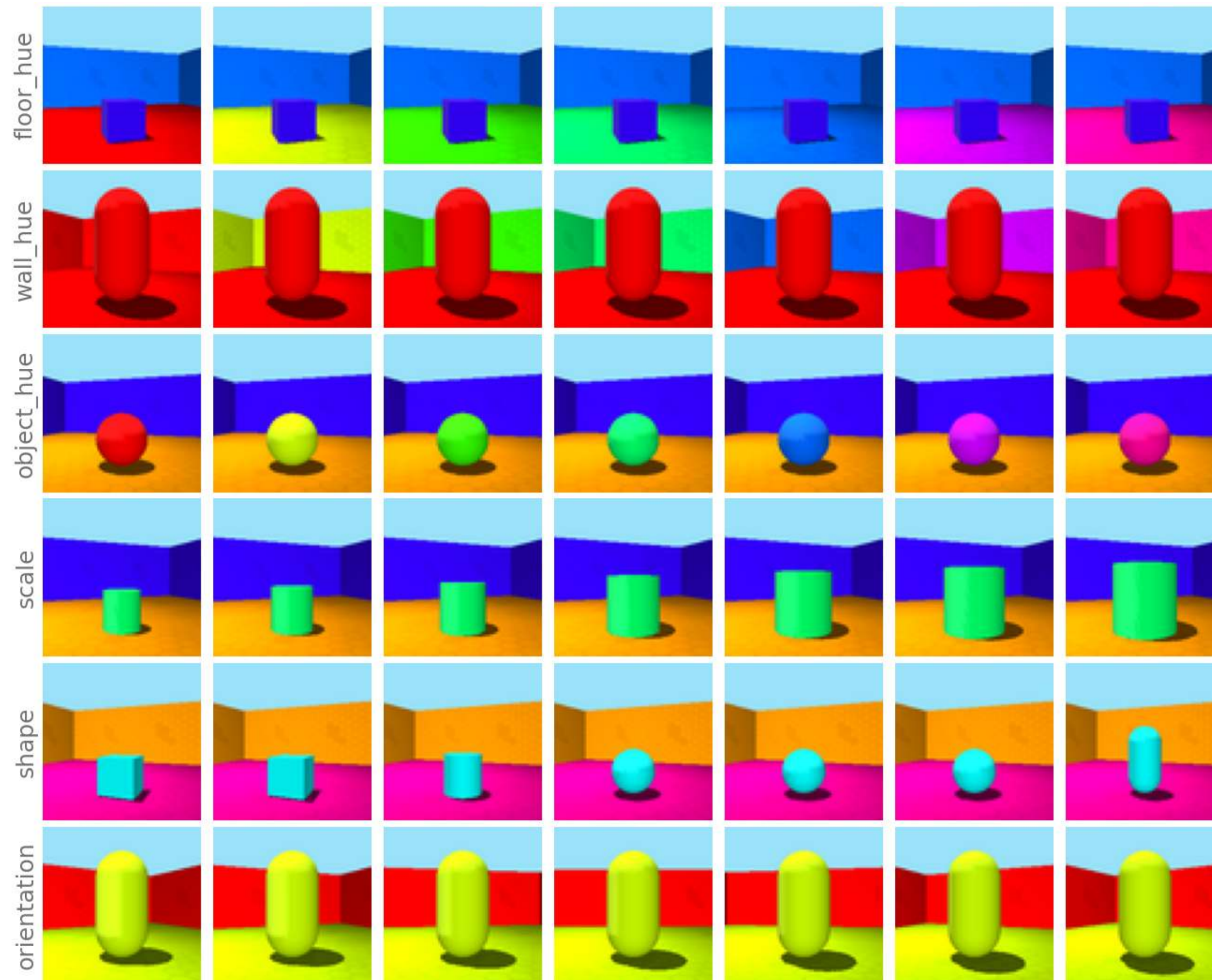
# Characterising Existing Datasets

Ground-truth distances usually correspond to VAE perceived distances

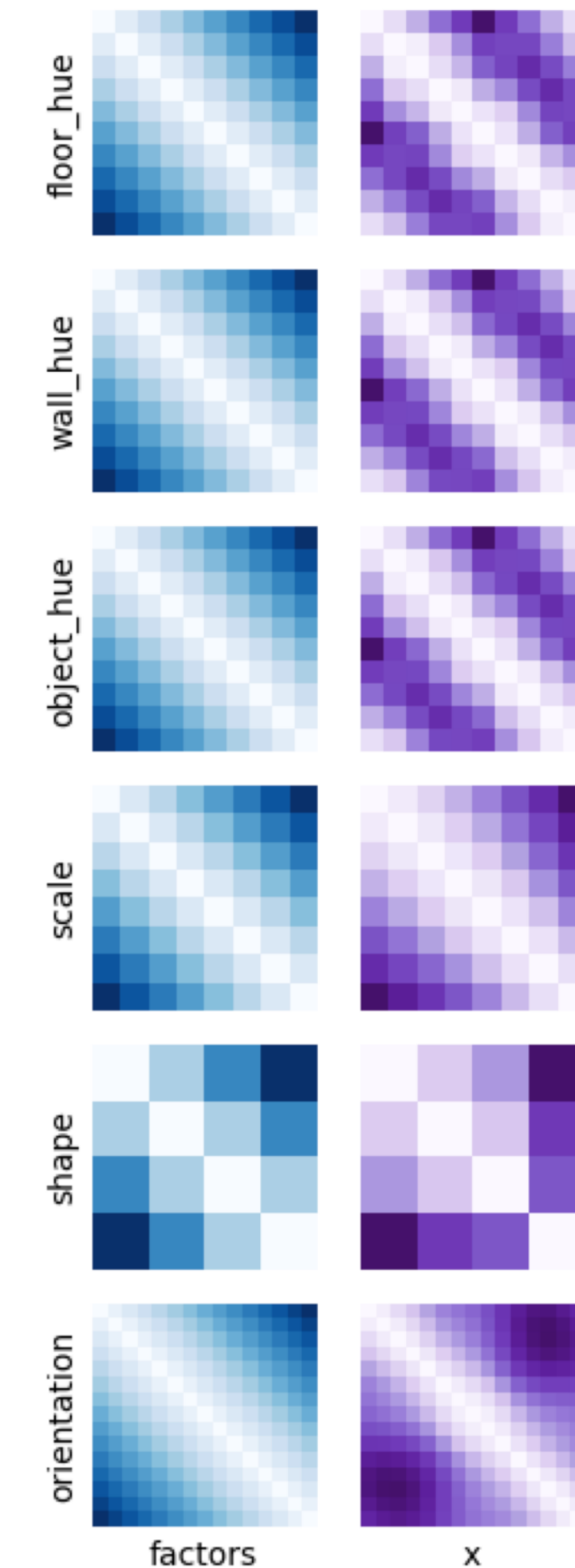


# Characterising Existing Datasets

Ground-truth distances usually correspond to VAE perceived distances



Shapes3D Dataset



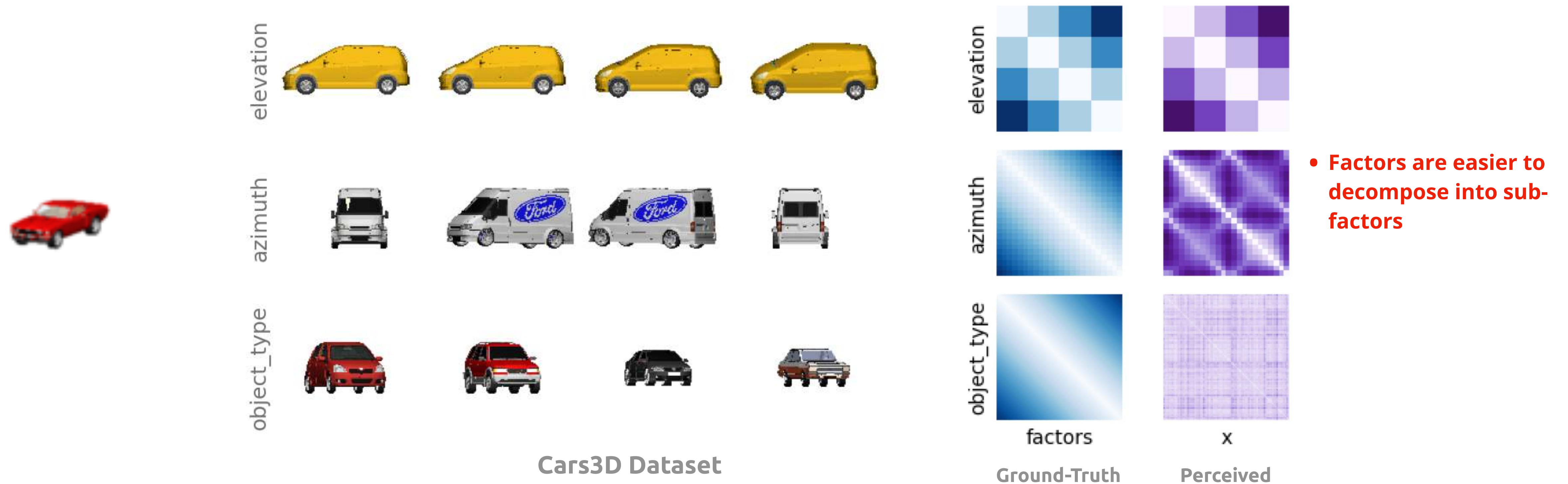
Ground-Truth

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

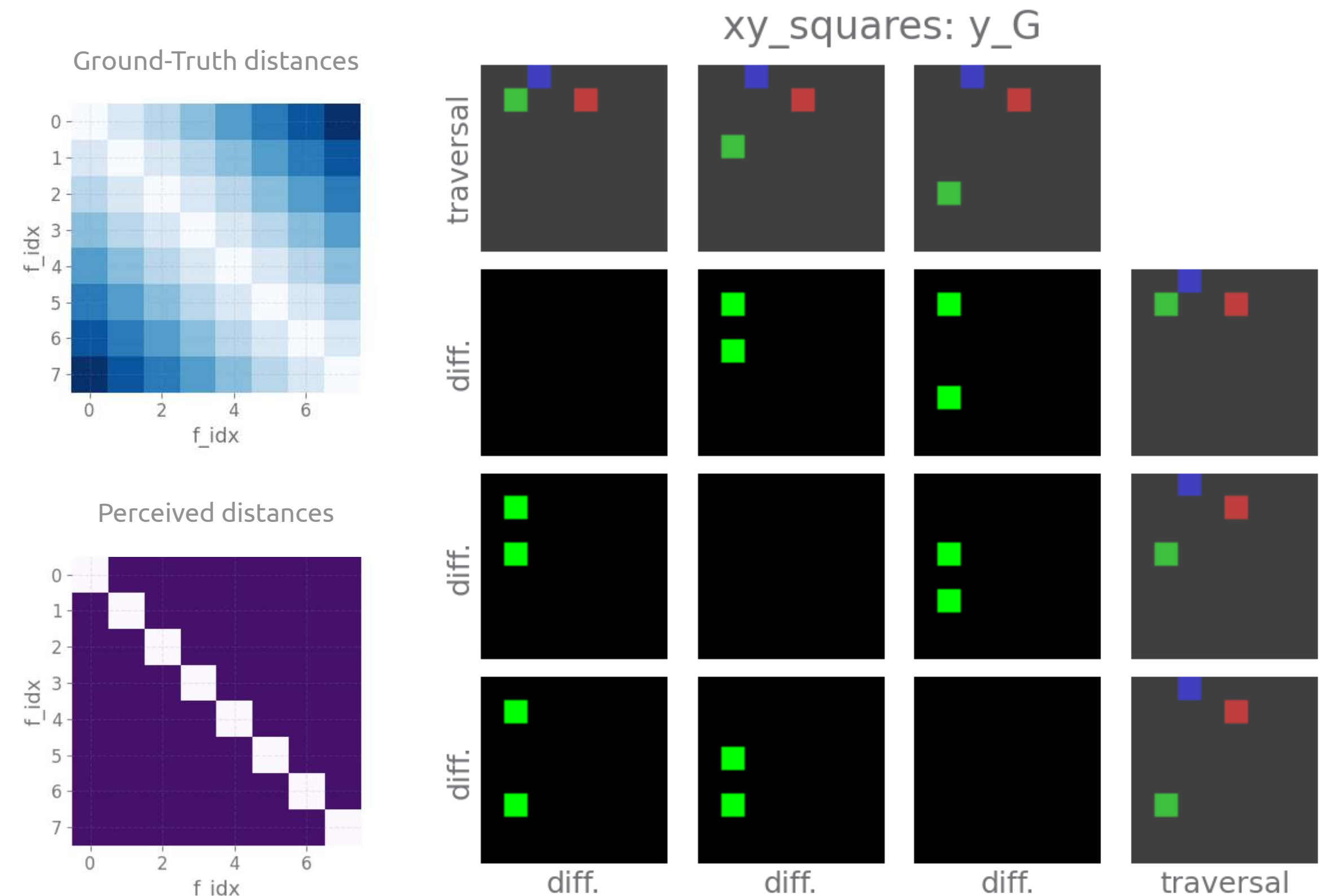
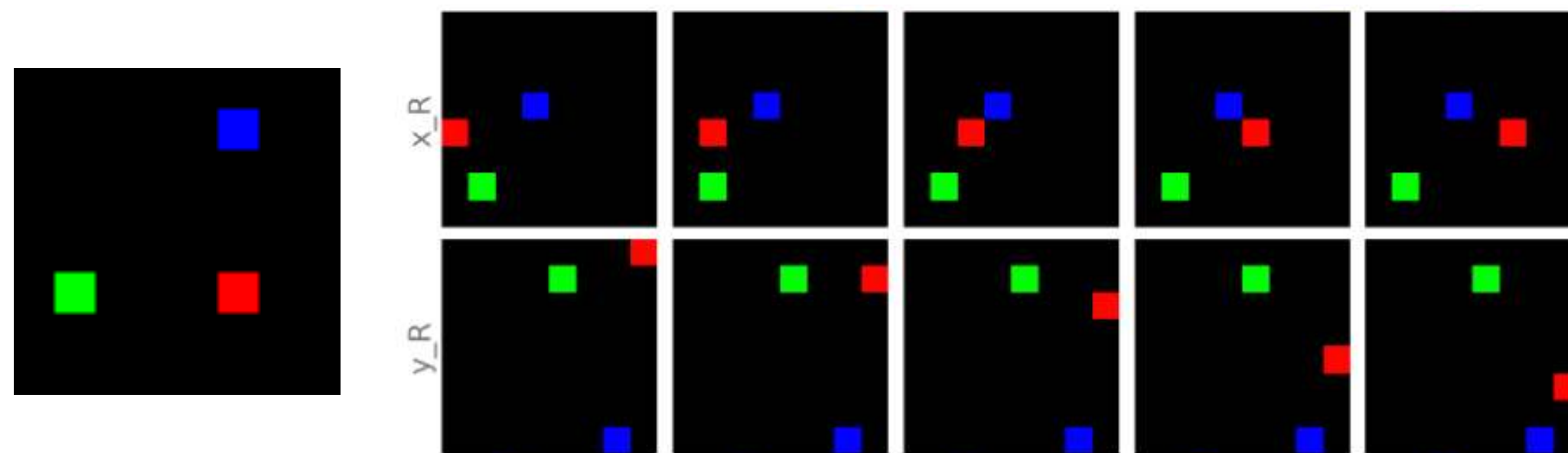
Ground-truth distances usually correspond to VAE perceived distances



# Adversarial Dataset

## XYSquares Example - Adversarial for pixel-wise reconstruction losses

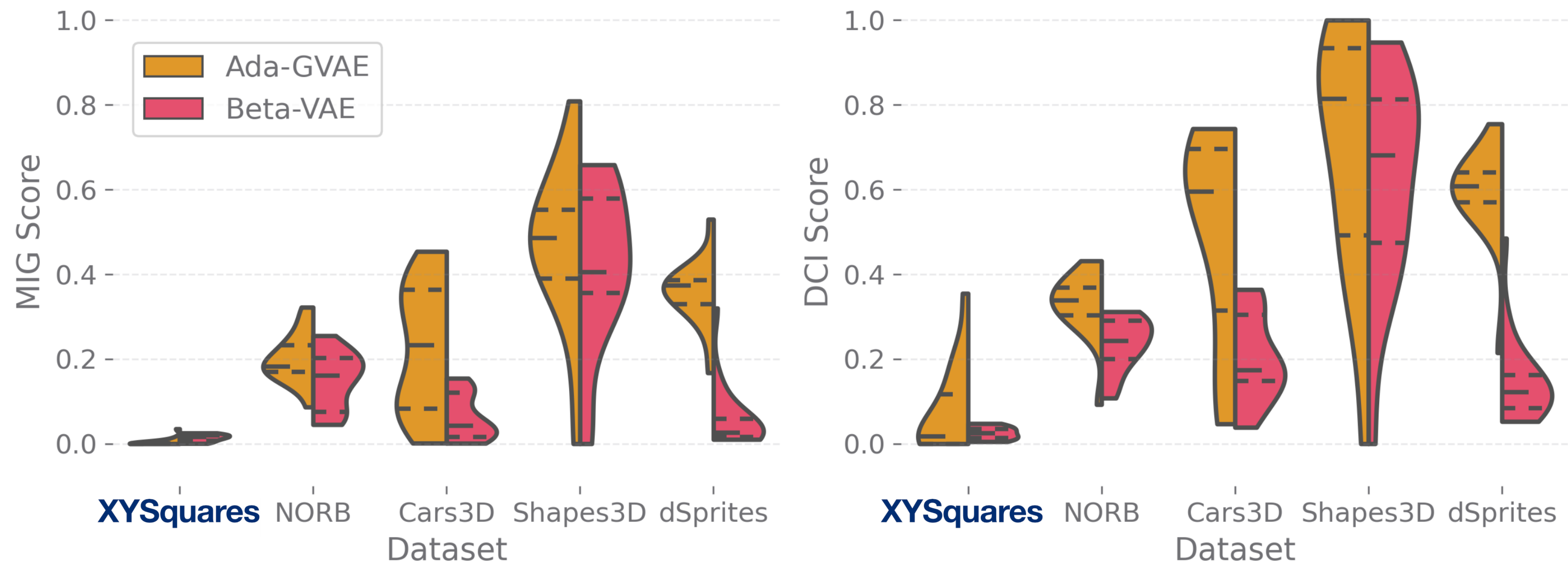
- Design example dataset with **constant overlap** along factor traversals
  - Cannot minimise recon. error due to sampling, **no ordering**
- 8x8 grid with **x** & **y** positional factors



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

# Adversarial Dataset

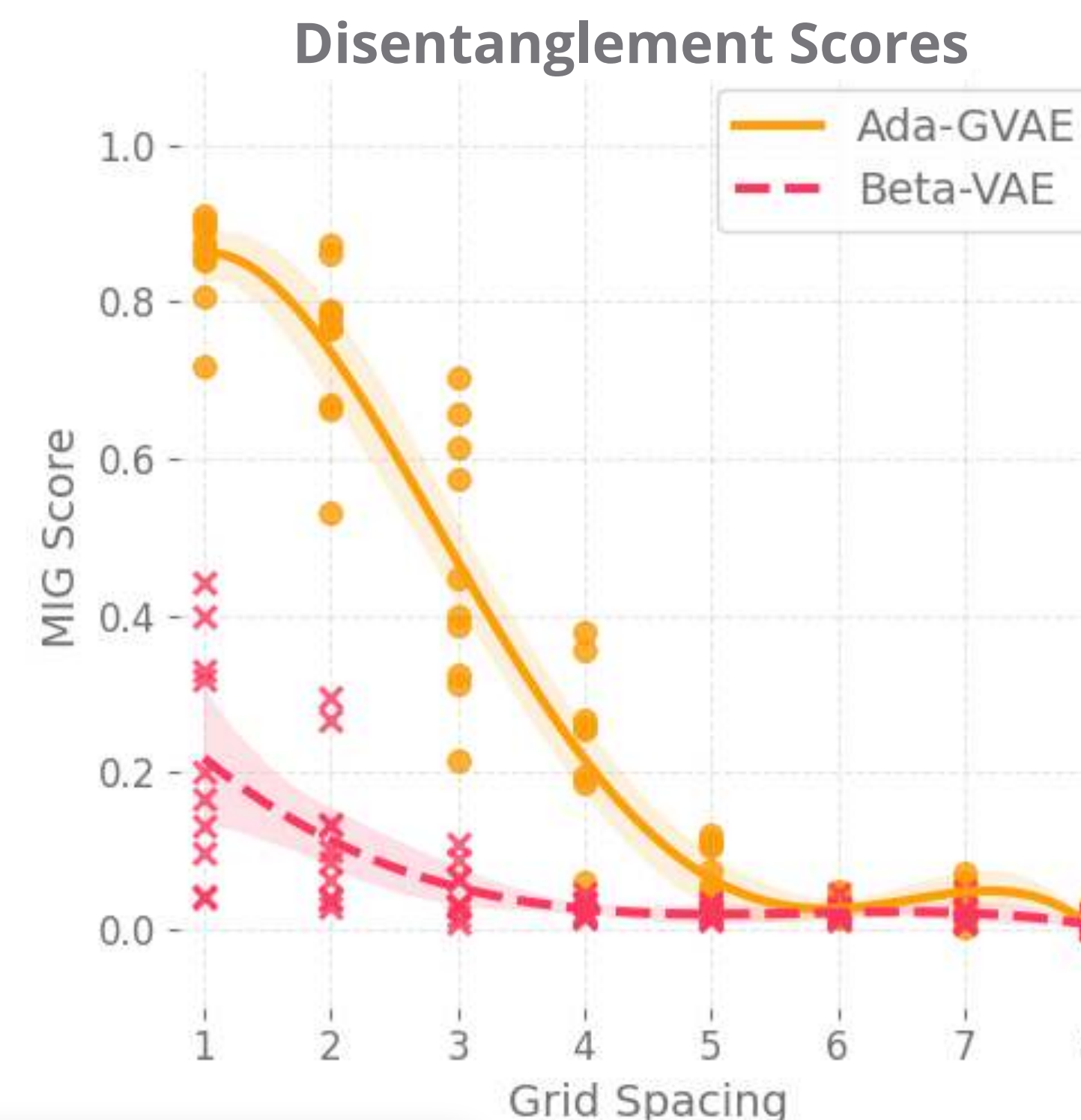
XYSquares Example - Disentanglement compared to benchmark datasets



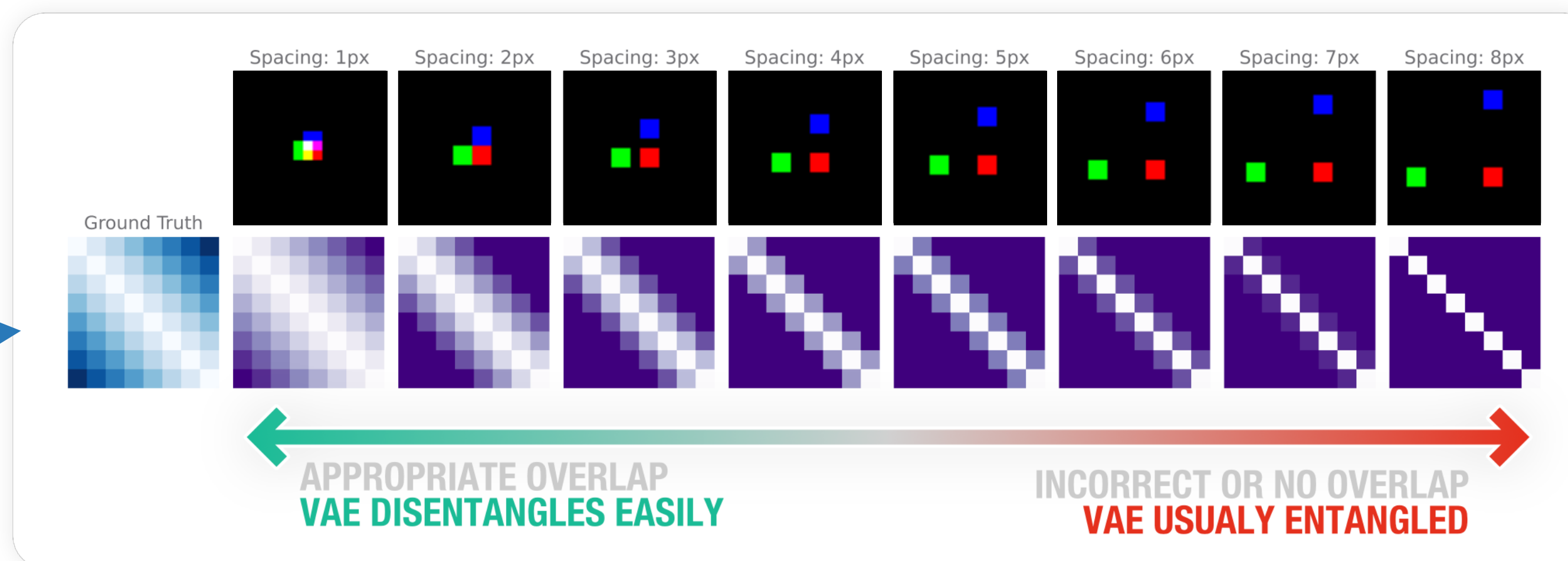
# Re-enabling Disentanglement

## 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

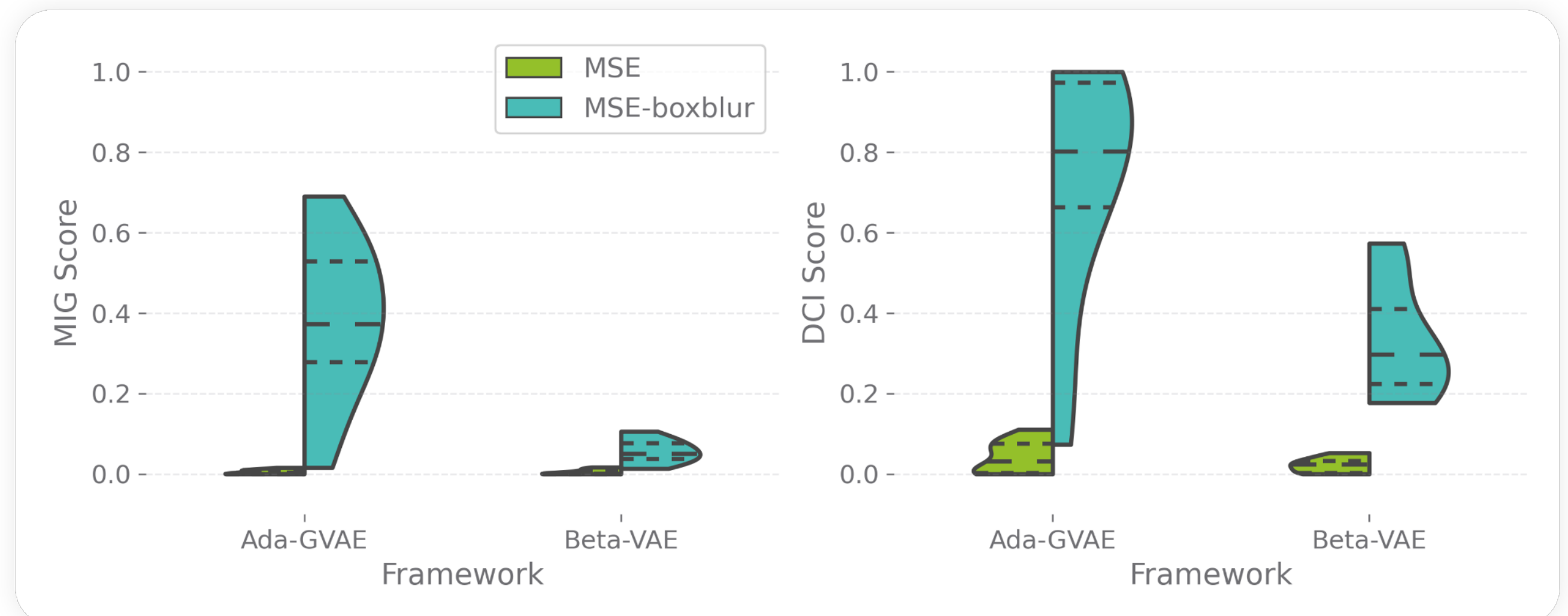
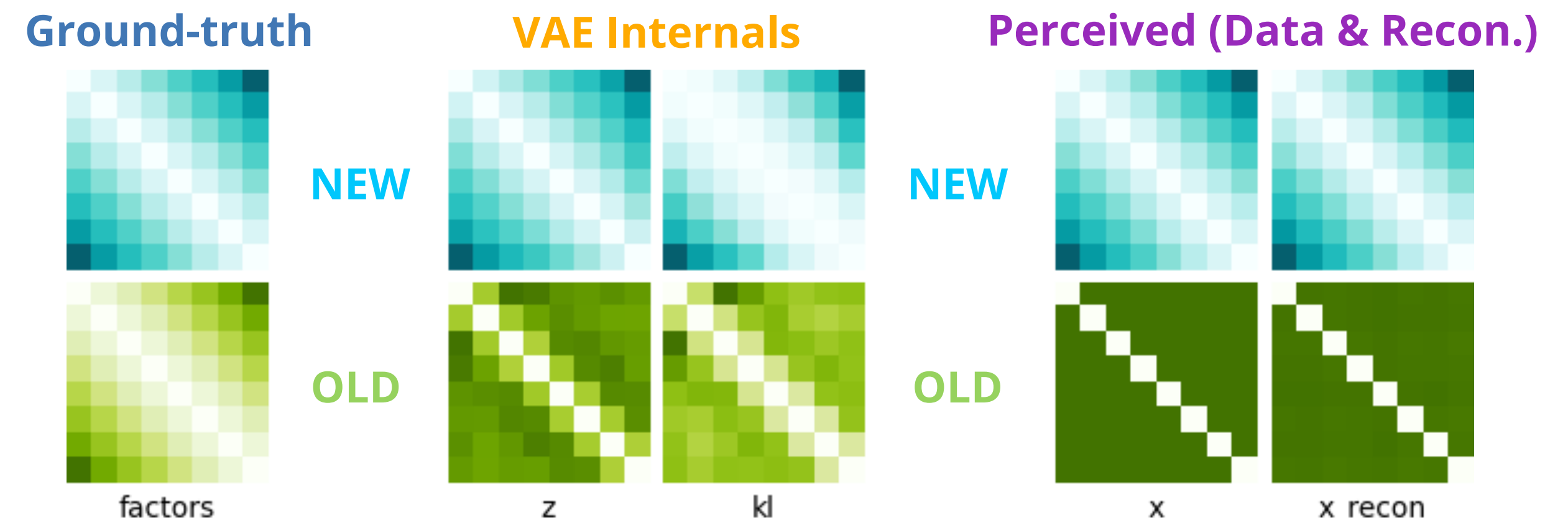


Perceived  
distance matrices  
along factors

# Re-enabling Disentanglement

XYSquares example - Adjusting the **loss**

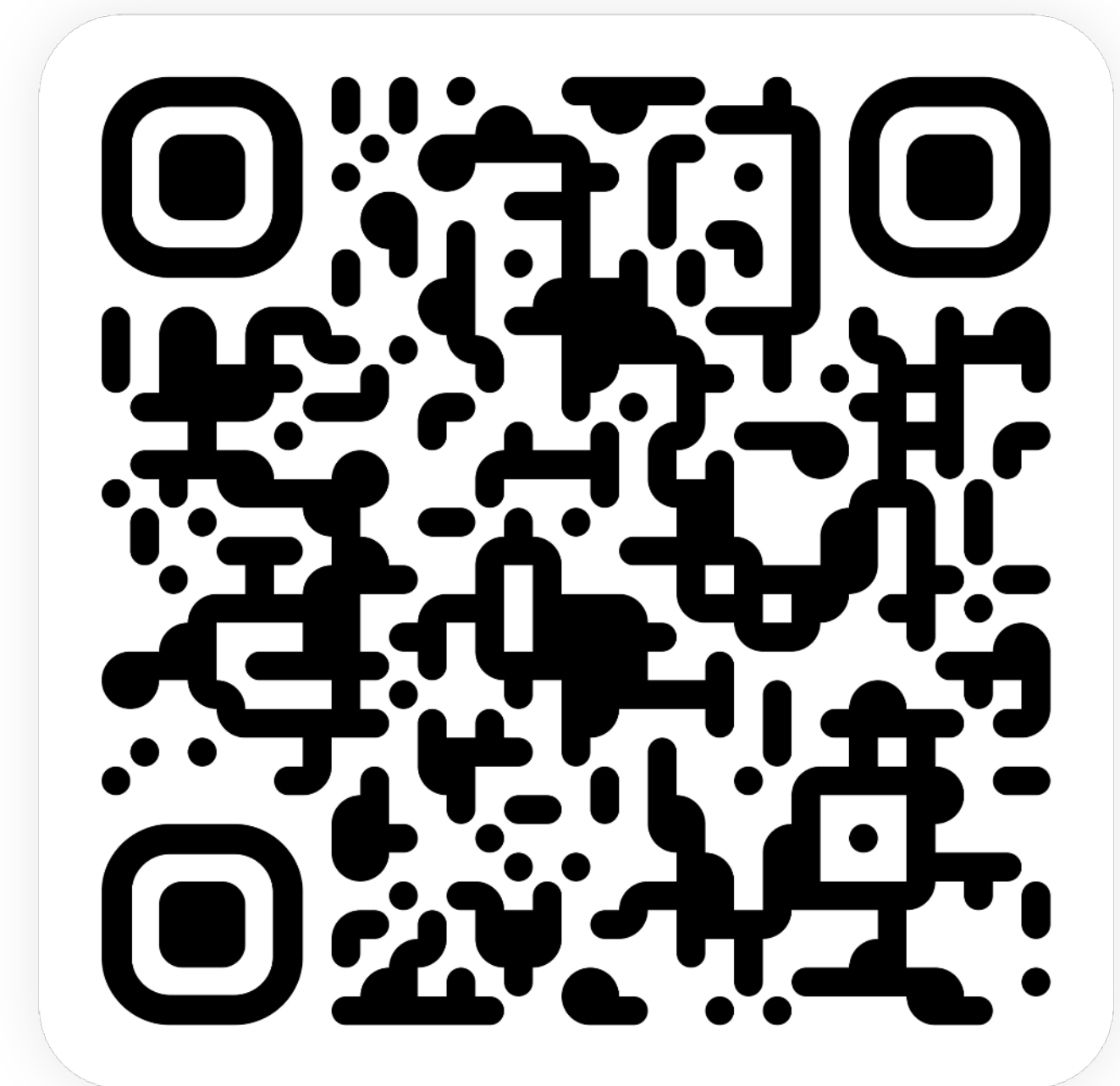
- If perceived distances once again correspond to ground-truth distances, disentanglement can occur.
- **Adjust the reconstruction loss to re-enable disentanglement.** An example that is appropriate for XYSquares is adding a box blur augment to data.



# 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 only 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 is still ultimately required in the current state.



**SCAN FOR PAPER & RESOURCES**

or visit  [github.com/nmichlo](https://github.com/nmichlo)

∞. **The End**

# -1. Background: **Representations**

# Representation Learning

## Definition

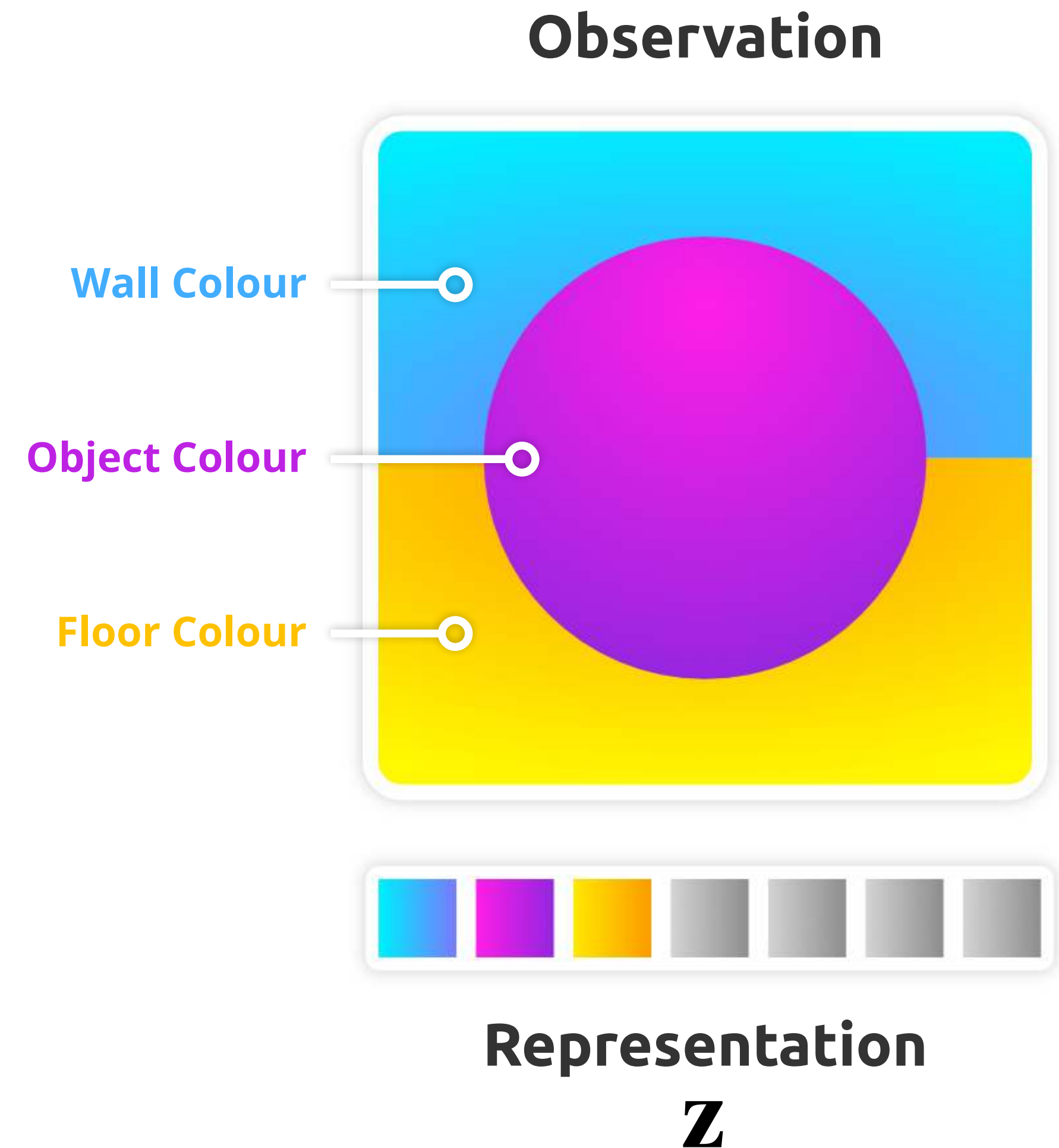
*“Learning representations of data that make it easier to extract useful information when building classifiers or other predictors”<sup>[1]</sup>*

### Example:

- Dataset has with **floor**, **object** and **wall** colours, that are the ground-truth factors.
- Learn **low dimensional vectors** that still capture these factors.

### Q: How do we learn these representations?

- Unsupervised<sup>[2]</sup> or Weakly-Supervised.
- PCA, ICA, VAEs, GANs, Ranking Loss, etc...



[1] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

[2] Locatello, Francesco, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations." *arXiv preprint arXiv:1811.12359* (2018).

# Disentangled Representations

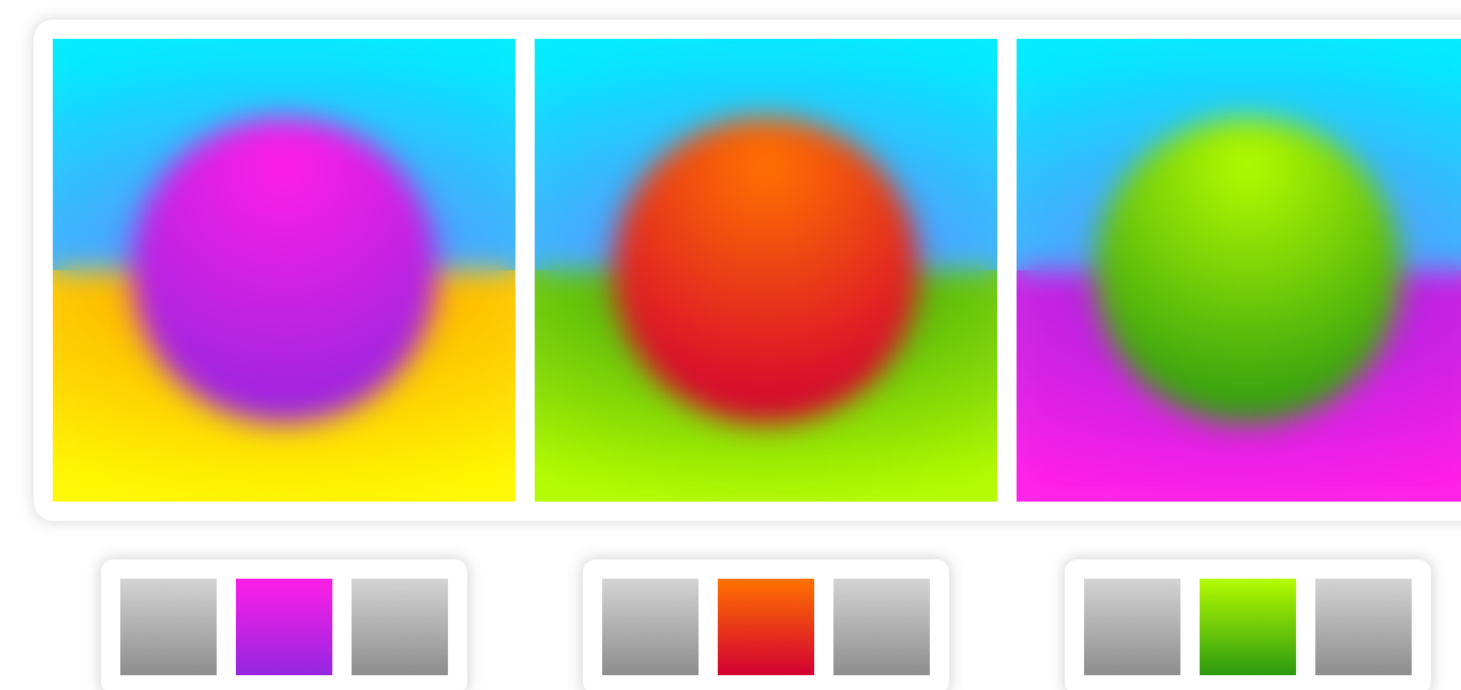
What Makes a Good Representation? <sup>[1]</sup>

- **Disentangled**

- Representations **align with ground-truth** factors
- **Sparse**: Only a subset of factors are relevant
- **Shared**: Differences attributed to a subset of factors
- Hierarchical components
- Interpretable & Fair

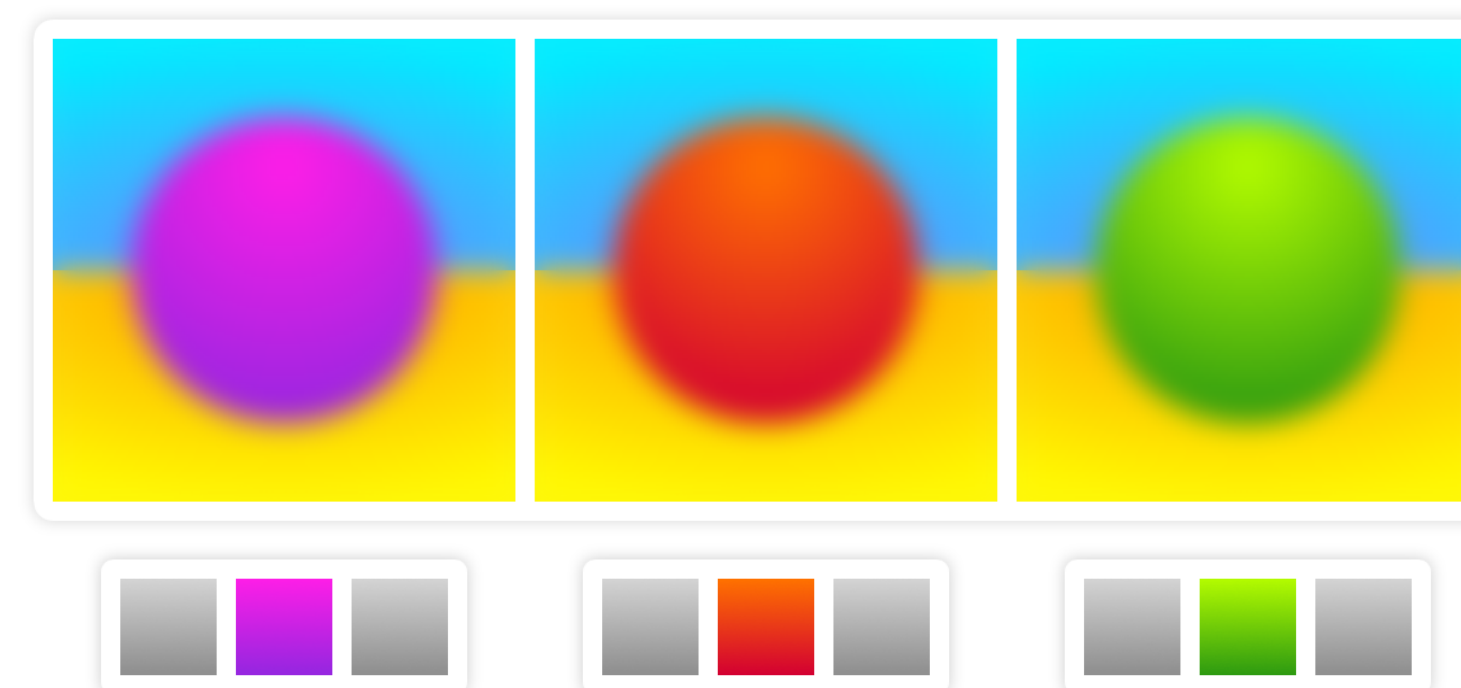
- **Other Properties**

- **Smooth**, without discontinuities
- Temporal and spatial coherence
- Natural clustering



## Entangled Representations

**Unintuitive** - Latent variables affect multiple factors.



## Disentangled Representations

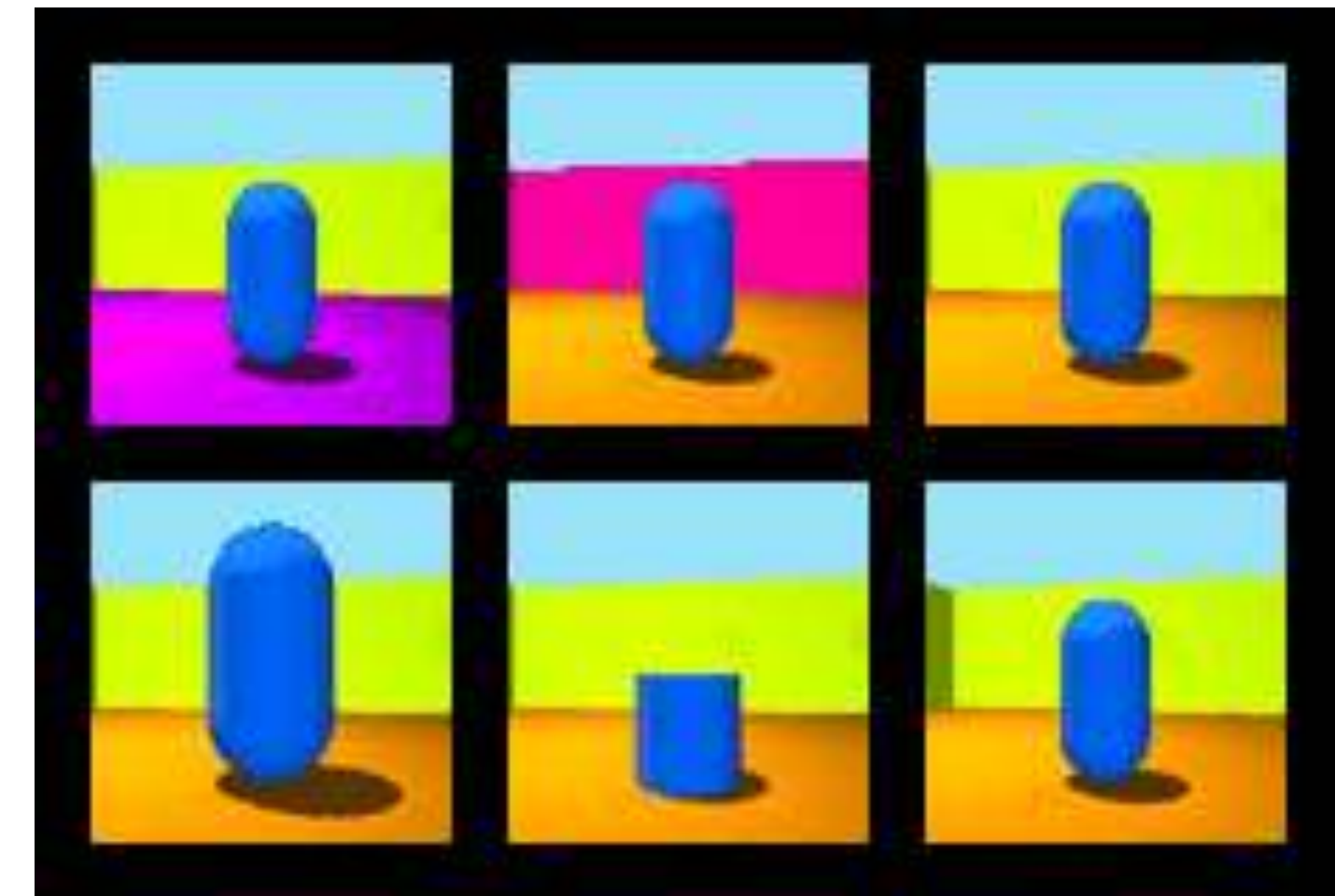
**Axis-aligned** - Latent variables affect distinct factors.

[1] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1798-1828.

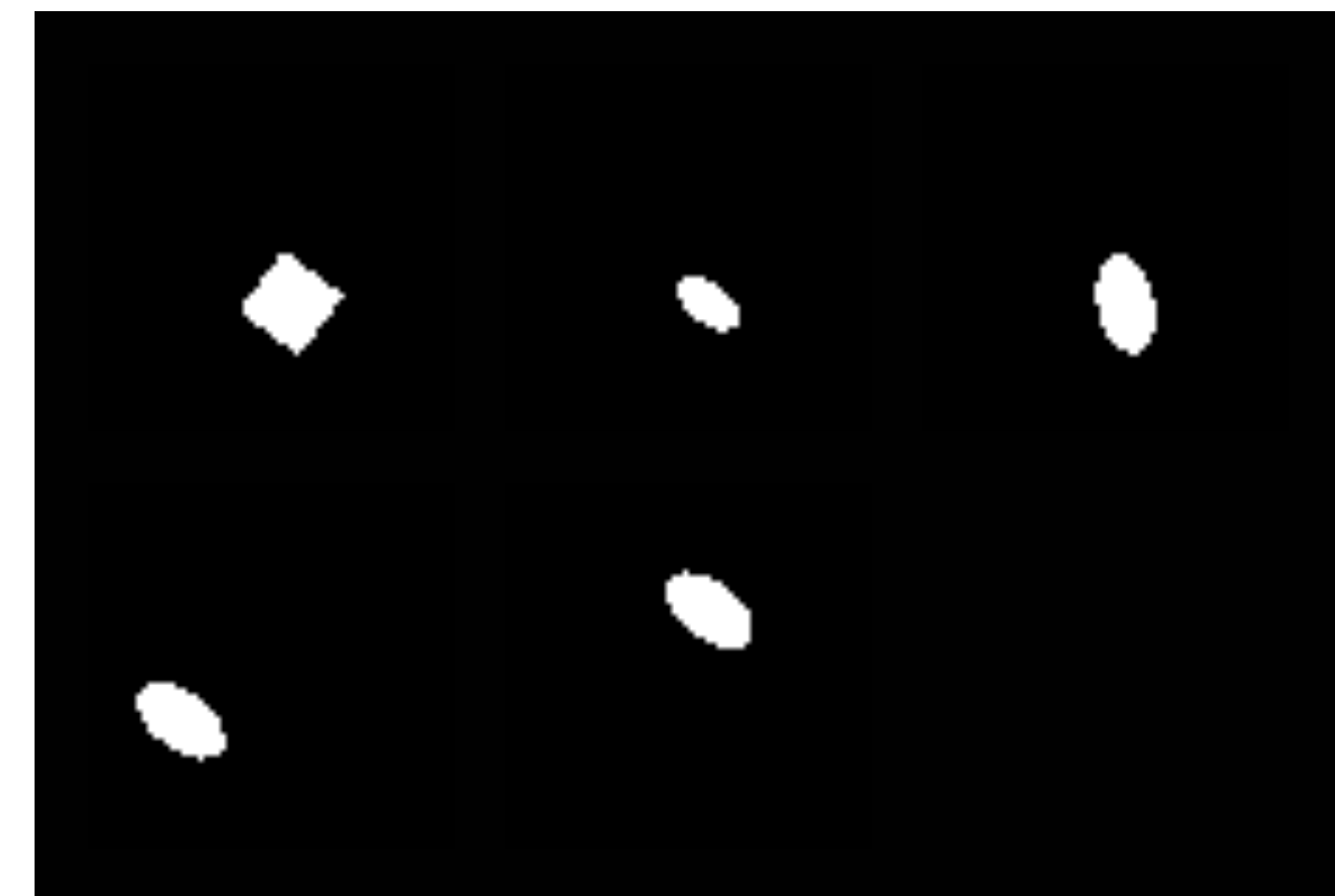
# How Do We Evaluate Representations?

Use Synthetic Datasets With Known Factors

- Use datasets with **known ground-truth factors**
  - 3DShapes, DSprites, MPI3D, Cars3D, Small NORB
- Using disentanglement metrics
  - **Compare learnt representations** and the **ground-truth factors** of each observation
  - $\beta$ -VAE Score, FactorVAE Score, DCI, MIG, SAP Score, Modularity
  - Each score has assumptions & biases, use multiple



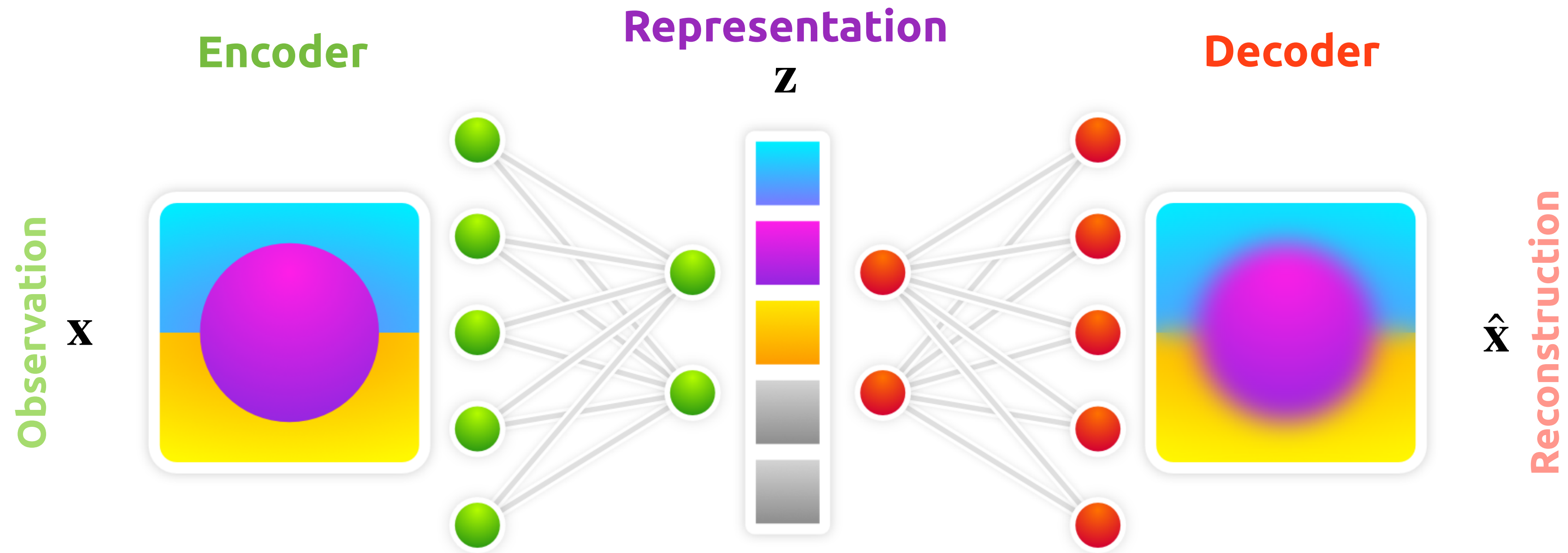
3DShapes  
Ground-Truth Factors



DSprites  
Ground-Truth Factors

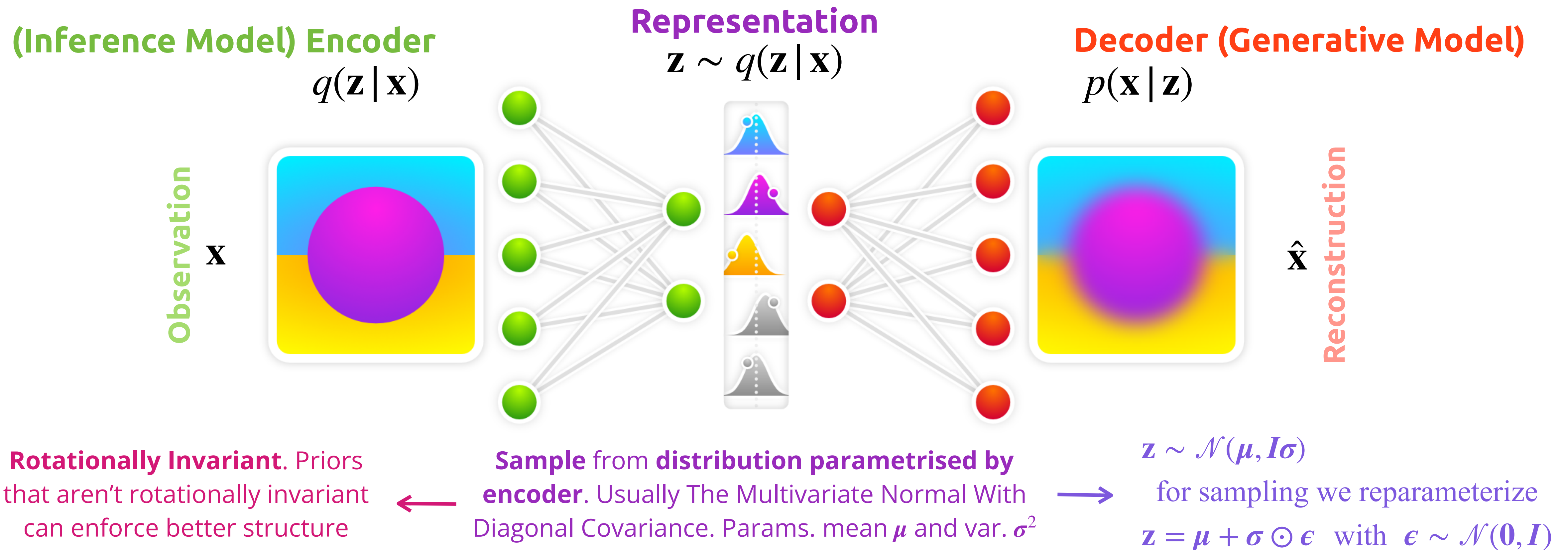
# Auto-Encoders

Representation Learning With Auto-Encoders



# Variational Auto-Encoders

Representation Learning With Variational Auto-Encoders (VAEs) <sup>[1]</sup>



[1] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

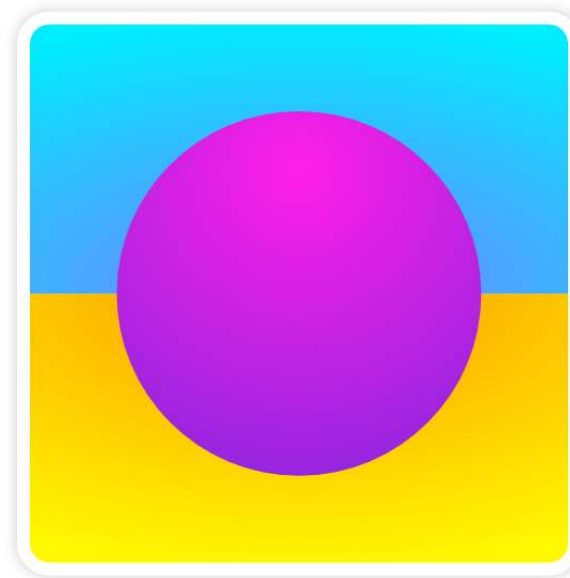
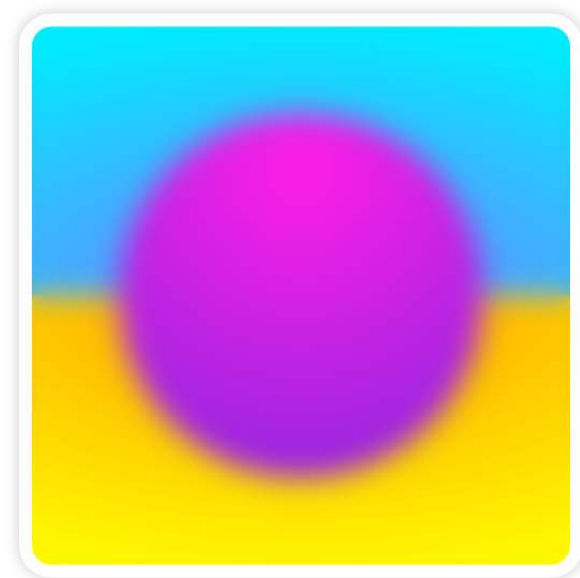
# Variational Auto-Encoders

$\beta$ -VAE Objective <sup>[1]</sup>

Coefficient adjusts strength of regularisation. <sup>[1]</sup> Higher values encourage better disentanglement but worse reconstruction quality. <sup>[2]</sup>

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

Reconstruction Term (Decoder)



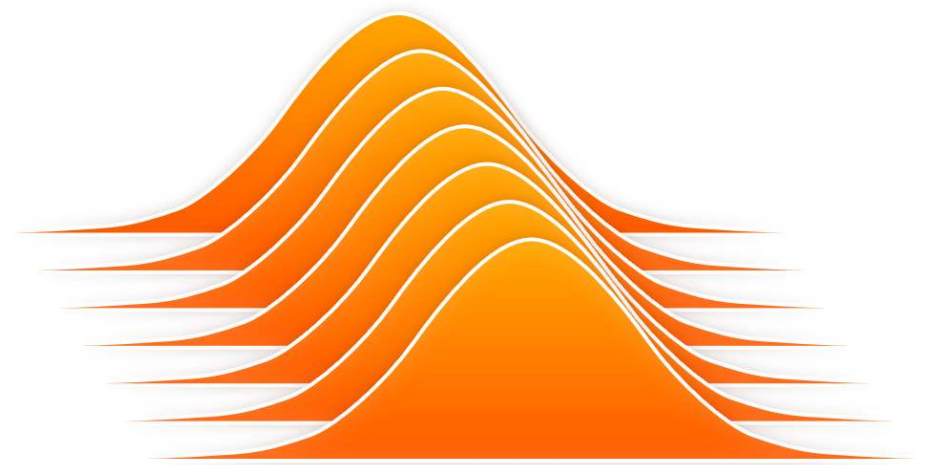
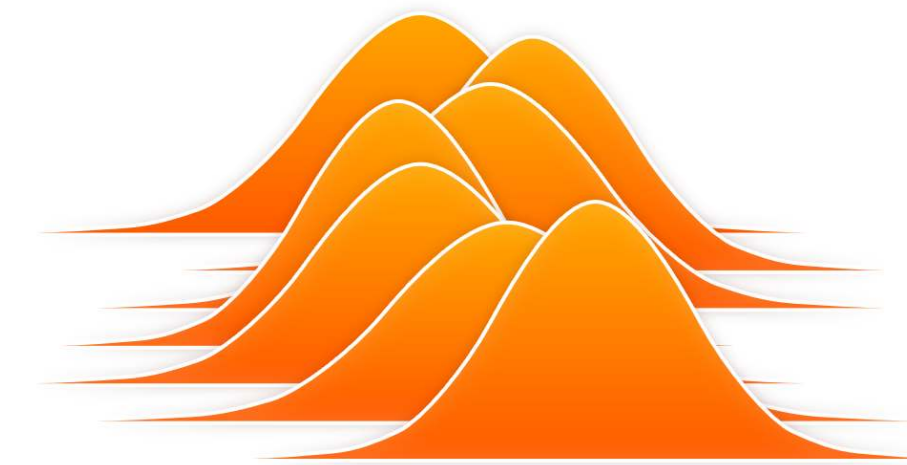
Reconstructions

**Improves** the **decoder** by encouraging reconstructions to match observations.

Observations

**Stronger regularisation**, sends less information to the decoder, resulting in **more blurred** reconstructions.

Regularising Term (Encoder)



Approx. Posterior

**Improves** the **encoder** by encouraging the approximate posterior to match the prior by minimising the KL Divergence.

Assumed Prior

The prior is usually assumed to be **standard normal distributions**, with a mean of 0 and variance of 1.

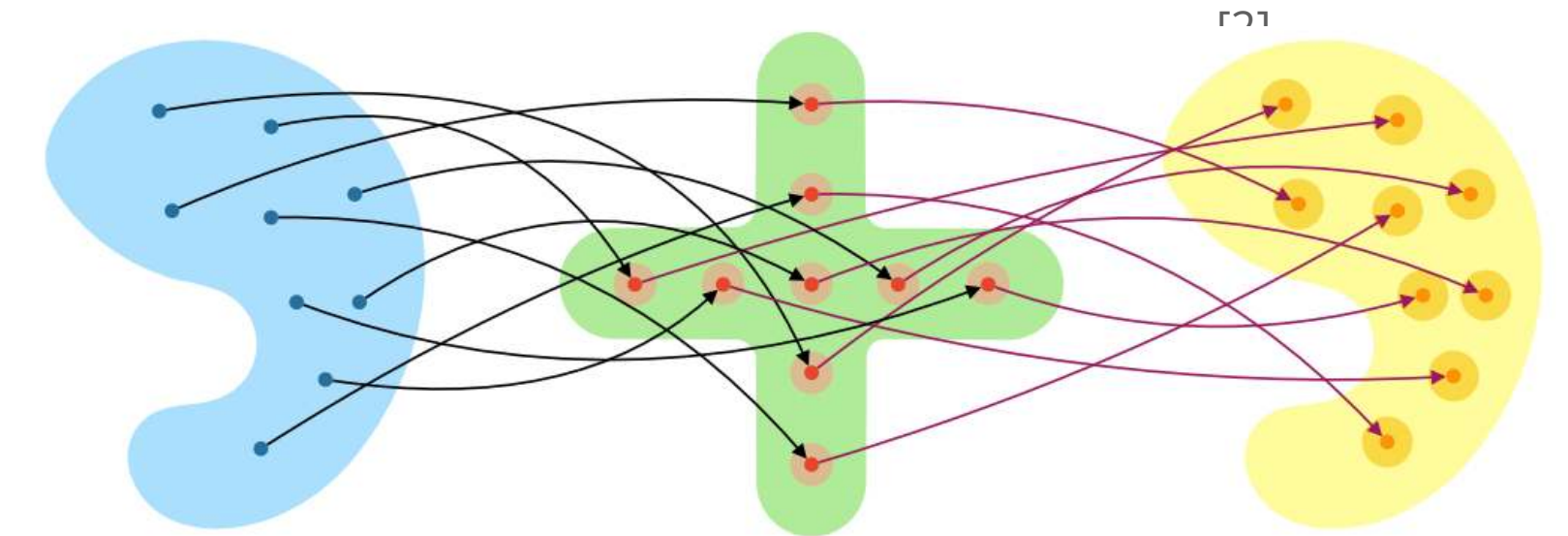
[1] Higgins, Irina, et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework." ICLR 2.5 (2017): 6.

[2] Burgess, Christopher P., et al. "Understanding disentangling in  $\beta$ -VAE." arXiv preprint arXiv:1804.03599 (2018).

# Disentanglement In Beta-VAEs

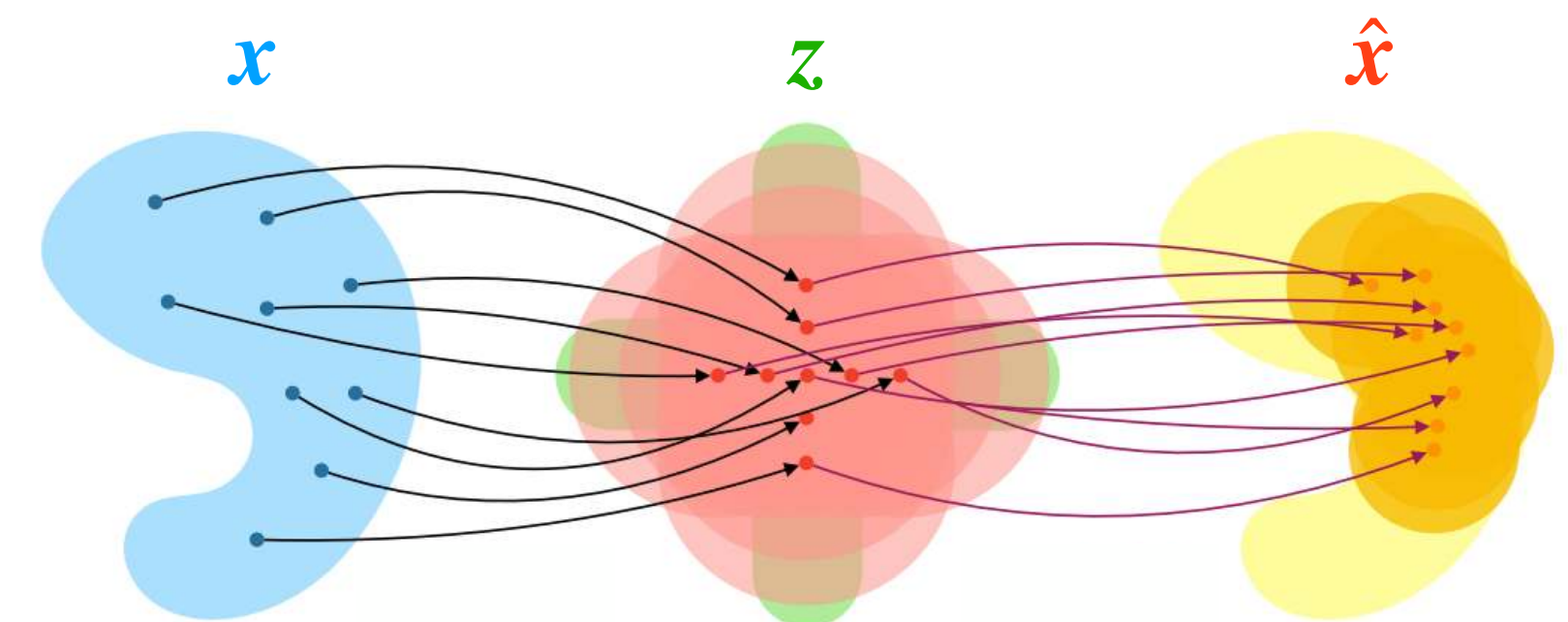
$\beta$ -VAEs Control Latent Overlap<sup>[2]</sup> And Sampling Errors<sup>[1]</sup>

- **Regulariser encourages** the latent distributions to **overlap**
  - Good for **smoothness & disentanglement**
  - Decoder can confuse sampled representations
- **Disentanglement?**
  - **Reorders latent space** to minimise reconstruction error because of sampling<sup>[1]</sup>
  - **Similar observations** in terms of loss **are placed together** in the latent space.



**Weak Beta:**  $\beta \rightarrow 0$

Acts like lookup table + Less sampling errors + Worse disentanglement



**Strong Beta:**  $\beta \rightarrow \infty$

Carries little information + More sampling errors + Better Disentanglement

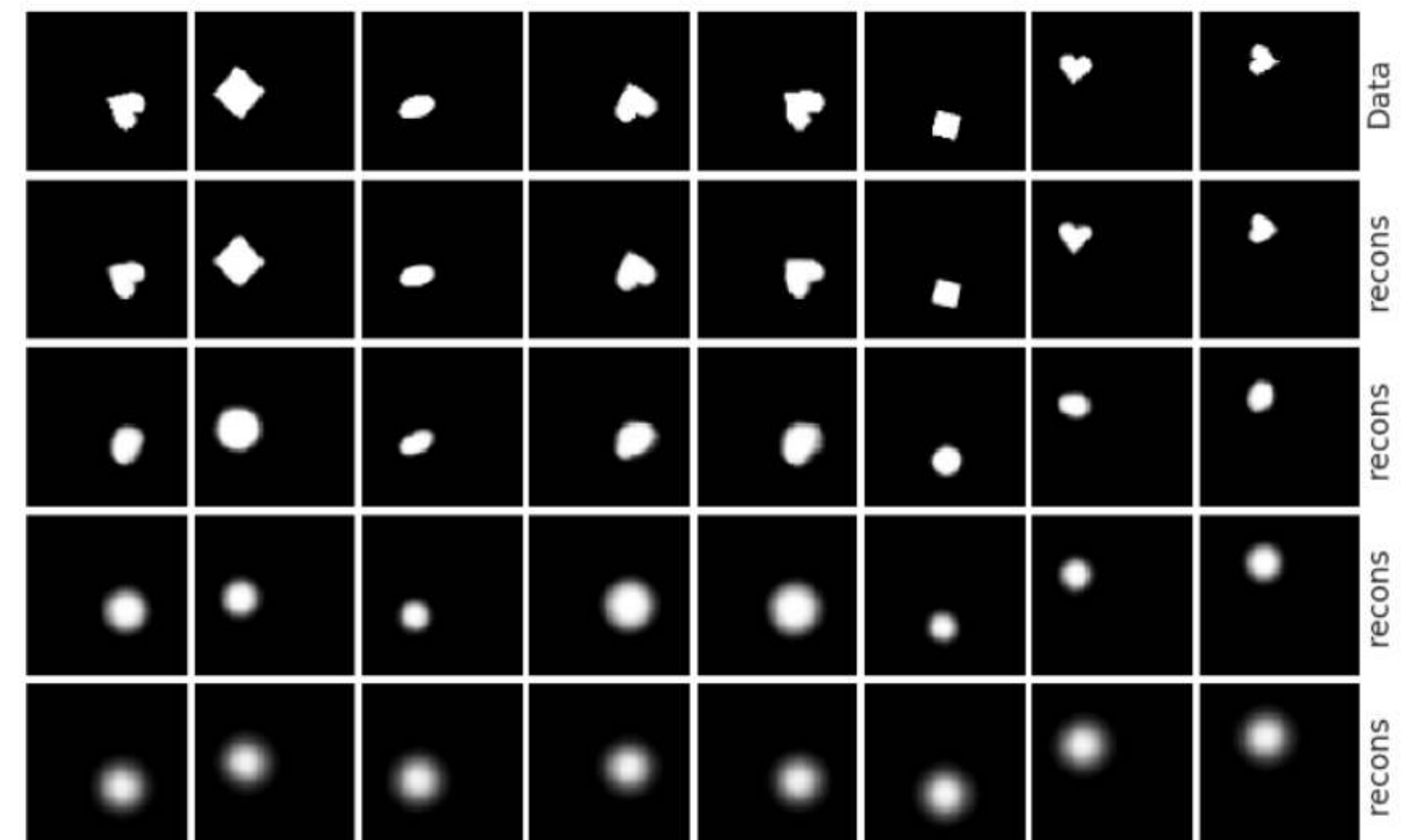
[1] Burgess, Christopher P., et al. "Understanding disentangling in  $\beta$ -VAE." arXiv preprint arXiv:1804.03599 (2018).

[2] Mathieu, Emile, Tom Rainforth, N. Siddharth, and Yee Whye Teh. 2018. "Disentangling Disentanglement In Variational Autoencoders". Arxiv.Org. <https://arxiv.org/abs/1812.02833>.

# Information Bottleneck

## $\beta$ -VAEs From The Information Bottleneck Perspective <sup>[1]</sup>

- Information Bottleneck
  - **Tradeoff** between accuracy and compression.
  - As  $\beta$  is increased, **less information can be encoded** by the representation
- Since this bottleneck exists
  - The **factors** that minimise the reconstruction loss are prioritised.
  - Sampling **errors reorder latent space**

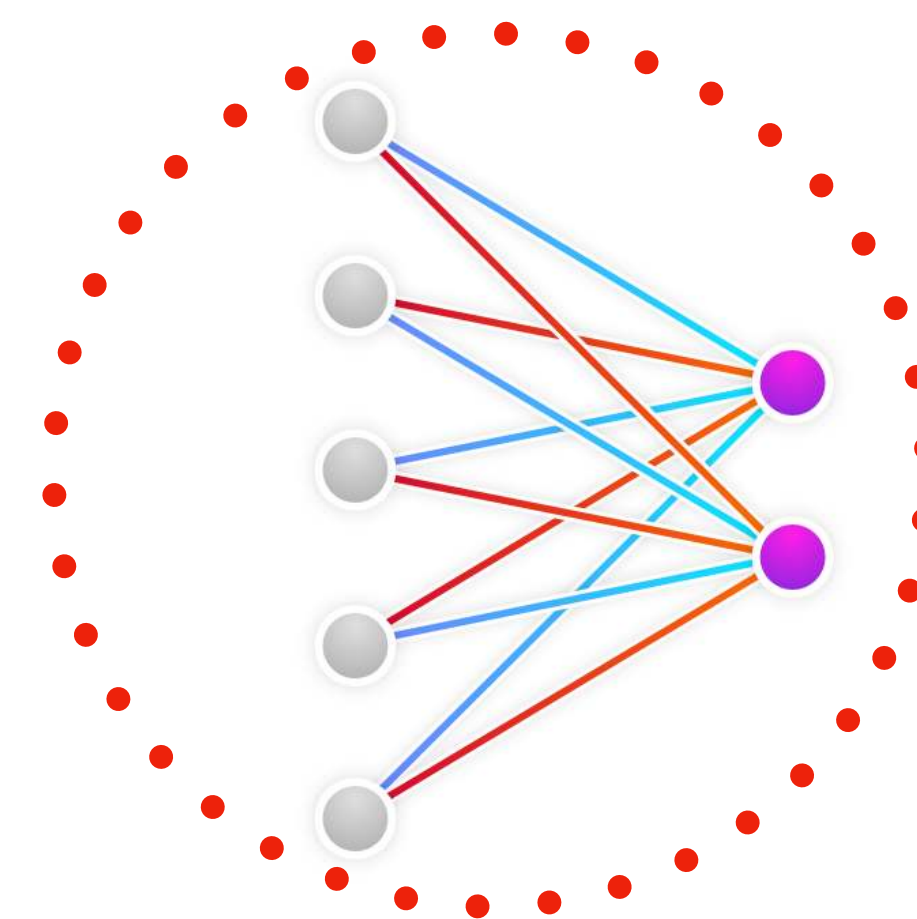
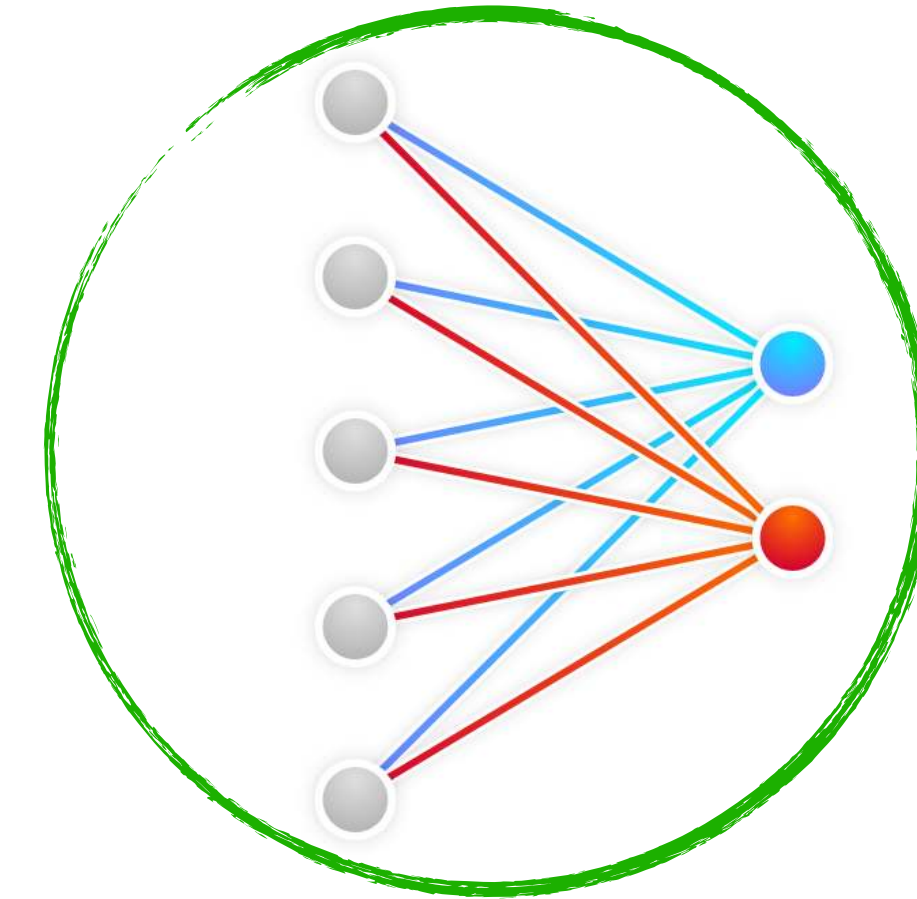


Q: Where does the axis-alignment come from?  
Bottleneck doesn't explain this!

# VAEs In Practice - Impossibility Result

Non-Identifiability of Unsupervised Models <sup>[1]</sup>

- Fundamentally **impossible** to reliably learn disentangled representations **without inductive biases**.
  - Performance of **unsupervised** methods largely **depends on** the initial **random seeds**.
  - We can use **model selection**, but this is impractical and still **requires** knowledge of the domain and **supervision**.



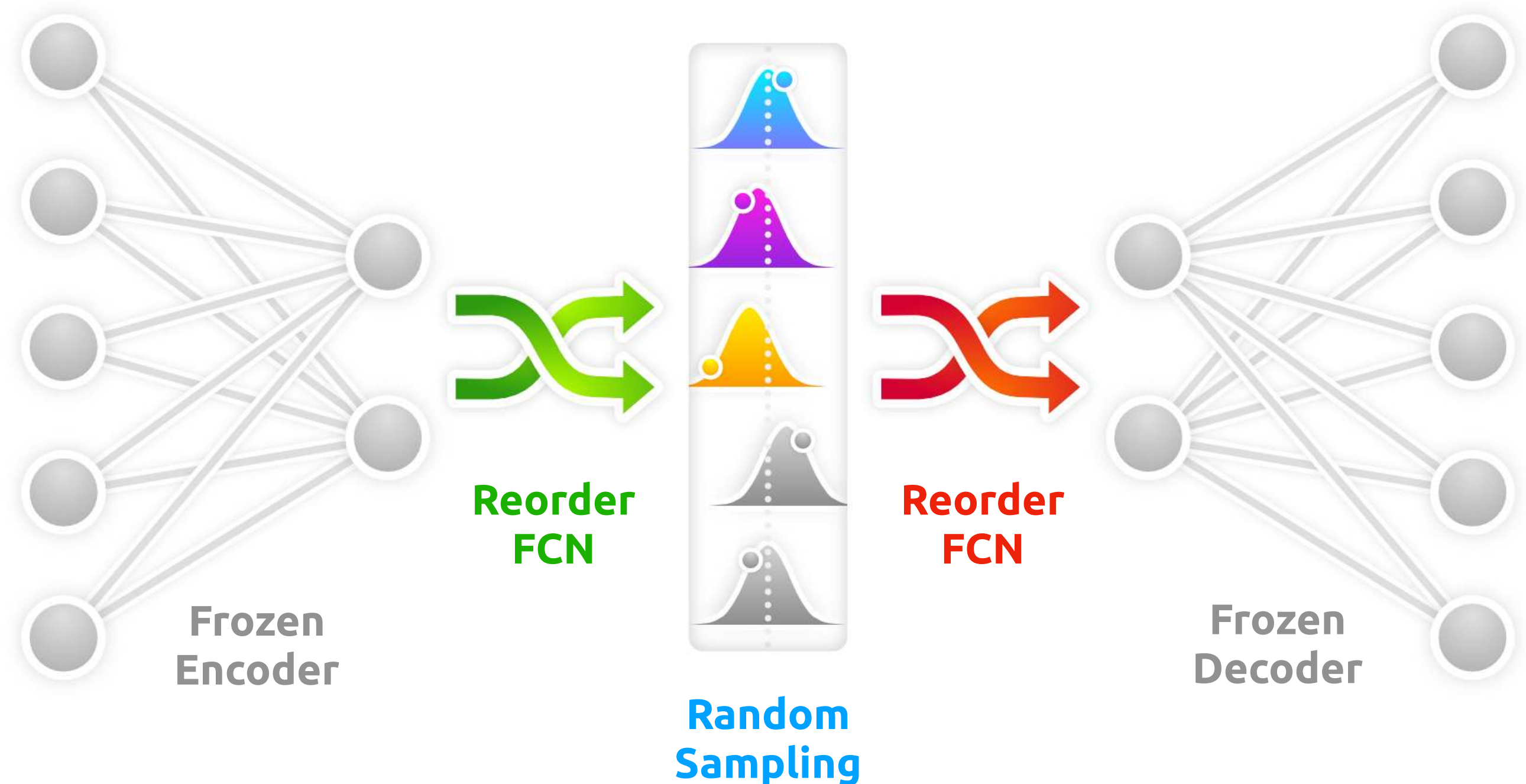
[1] Locatello, Francesco, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations." arXiv preprint arXiv:1811.12359 (2018).

## **-2. Overlap In Existing Datasets**

# Minimise Sampling Error To Disentangle

A Simplified Example Of How VAEs Disentangle

1. **Consider** pre-trained **auto-encoder** with **perfect reconstructions**
2. **Freeze** auto-encoder **weights**
3. **Convert into VAE** by adding fully connected layers
4. **Train as VAE**
  - Training **reorders** the **latent space** to **minimise** reconstruction loss **error** due to **random sampling**

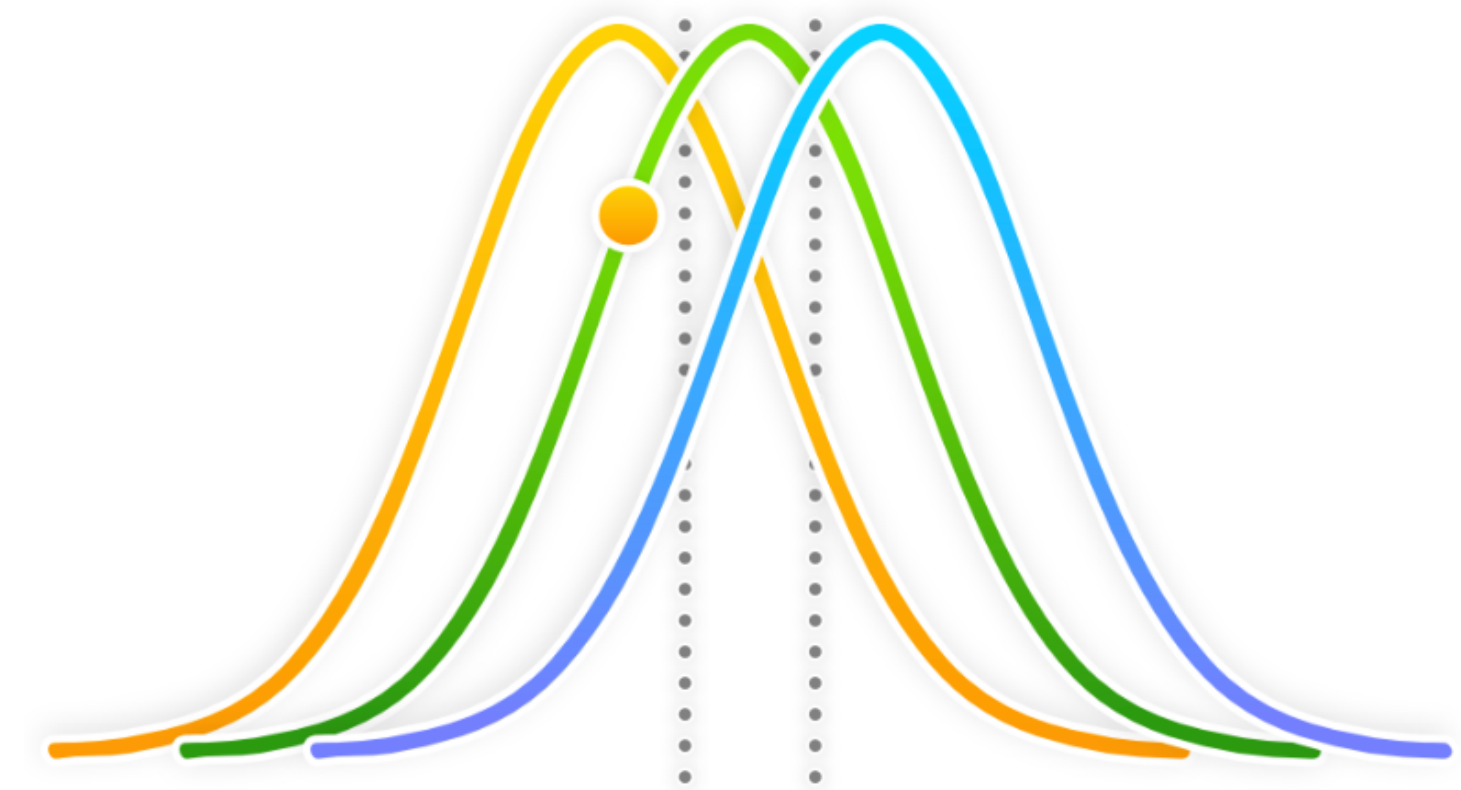


# Visual Distance (Visual Overlap)

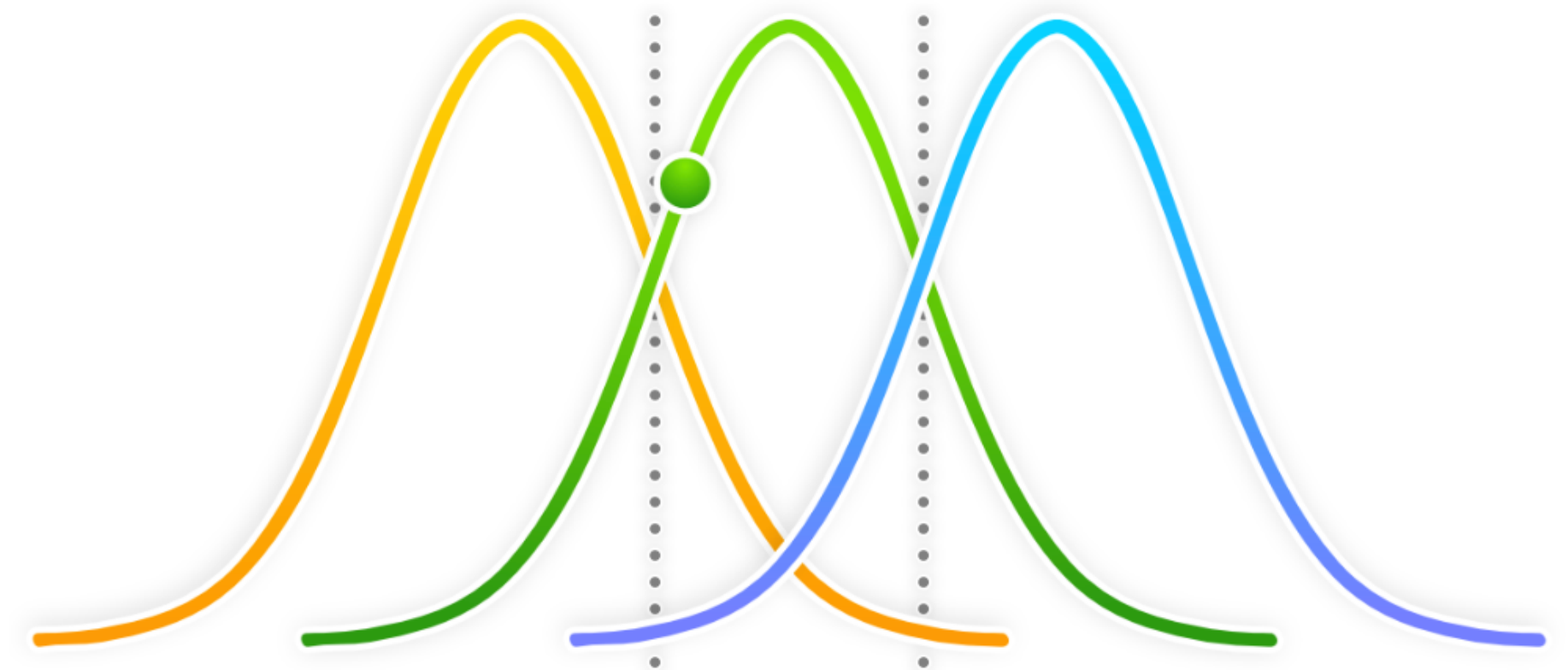
A Way To Examine Existing Datasets

- Use reconstruction loss as a distance metric
  - Compare error between true reconstruction and sampled reconstruction
  - Assume perfect reconstructions we are just comparing elements from the dataset itself
  - MSE is just square euclidean distance

$$\begin{aligned}d_{\text{vis}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) &= \lim_{\hat{\mathbf{x}} \rightarrow \mathbf{x}} \mathcal{L}_{\text{rec}}(\hat{\mathbf{x}}^{(a)}, \hat{\mathbf{x}}^{(b)}) \\ &= \mathcal{L}_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})\end{aligned}$$



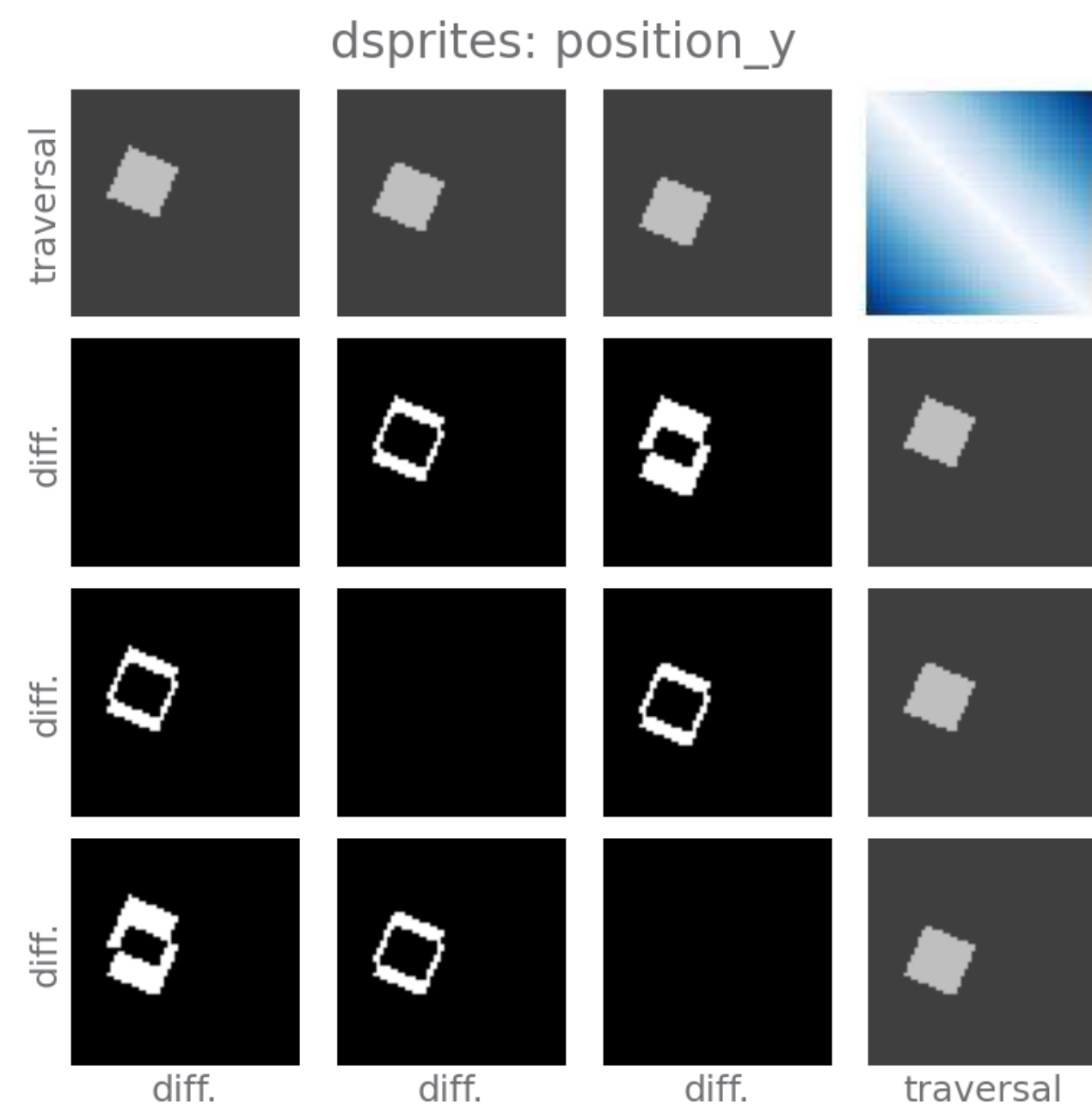
*b*: Confused Distribution



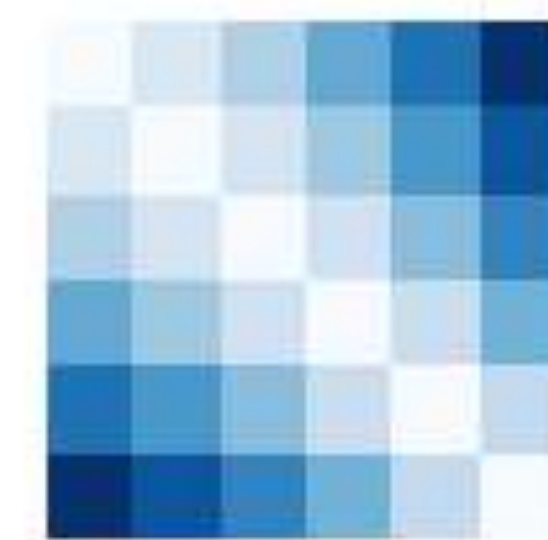
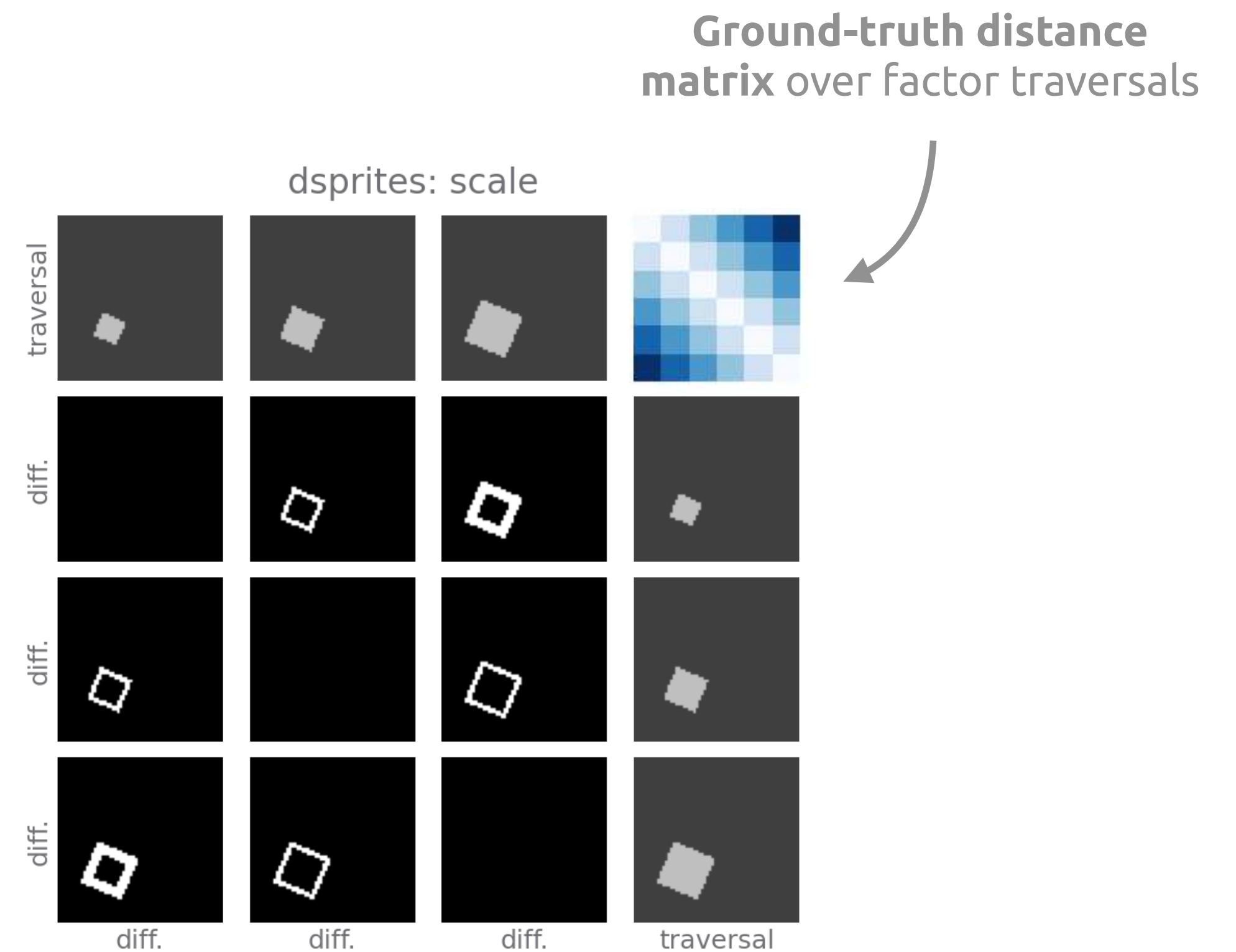
*a*: Correct Distribution

# Distance Matrices

A Way To Examine Existing Datasets

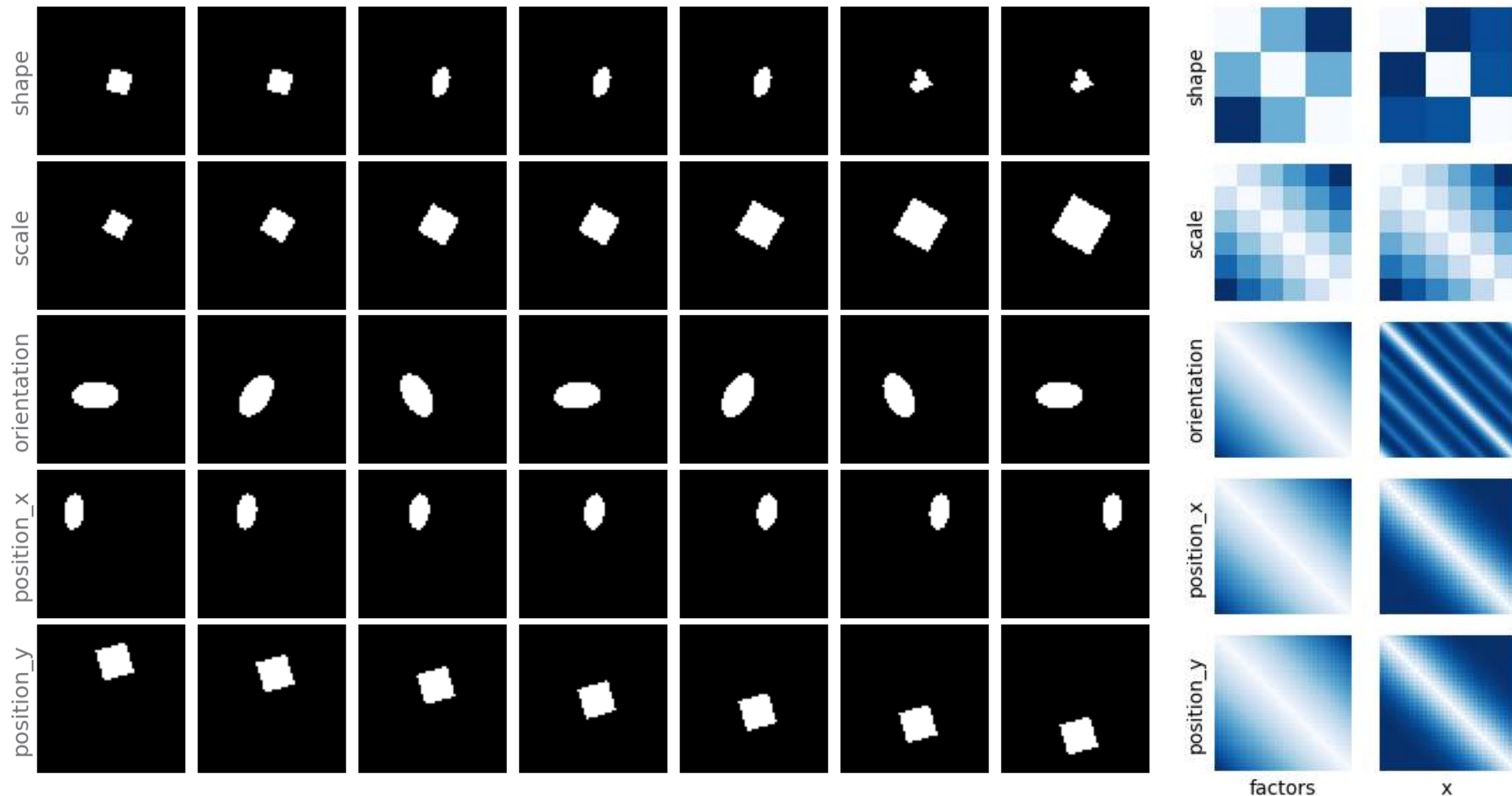


Visual distance matrix  
between observations  
over factor traversals



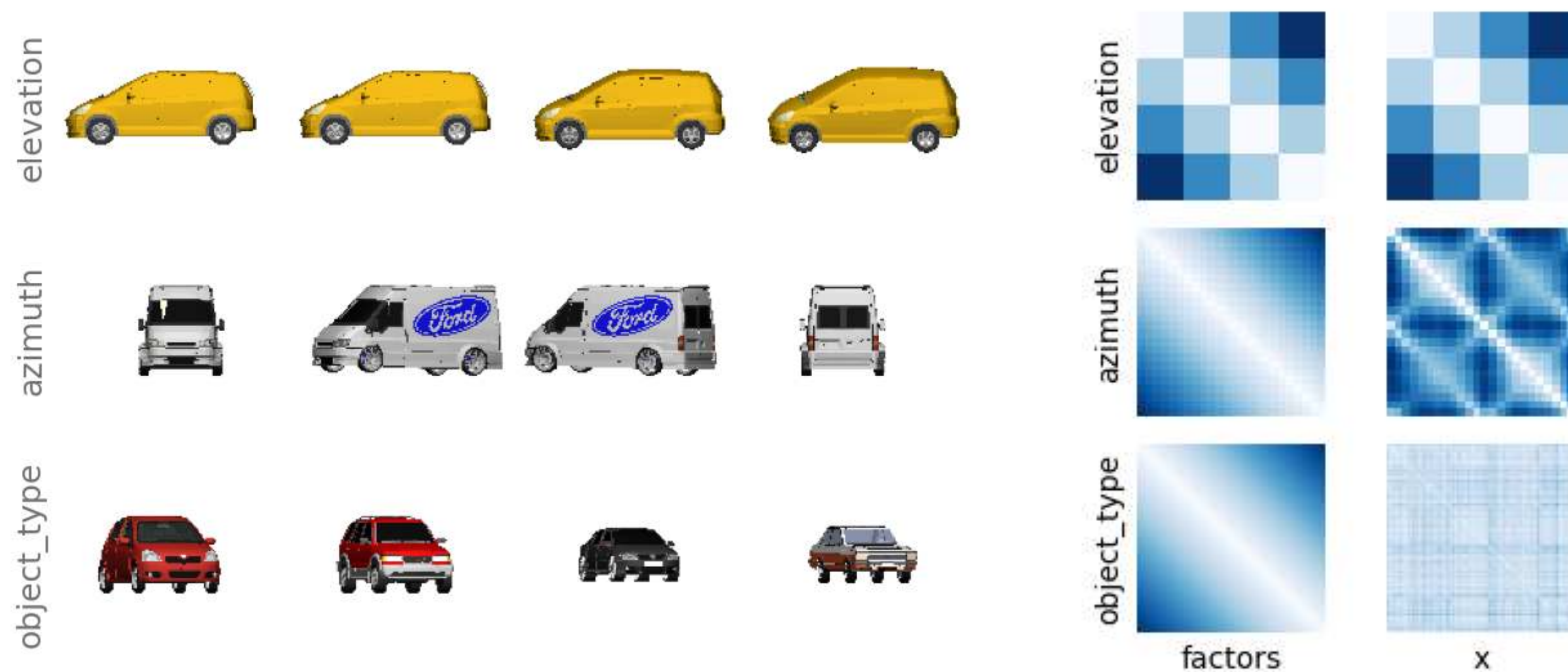
# Existing Ground-Truth Datasets

DSprites: Ground-Truth Factors Correspond To Observation Distances



# Existing Ground-Truth Datasets

Cars3D: Ground-Truth Factors Sometimes Correspond To Observations



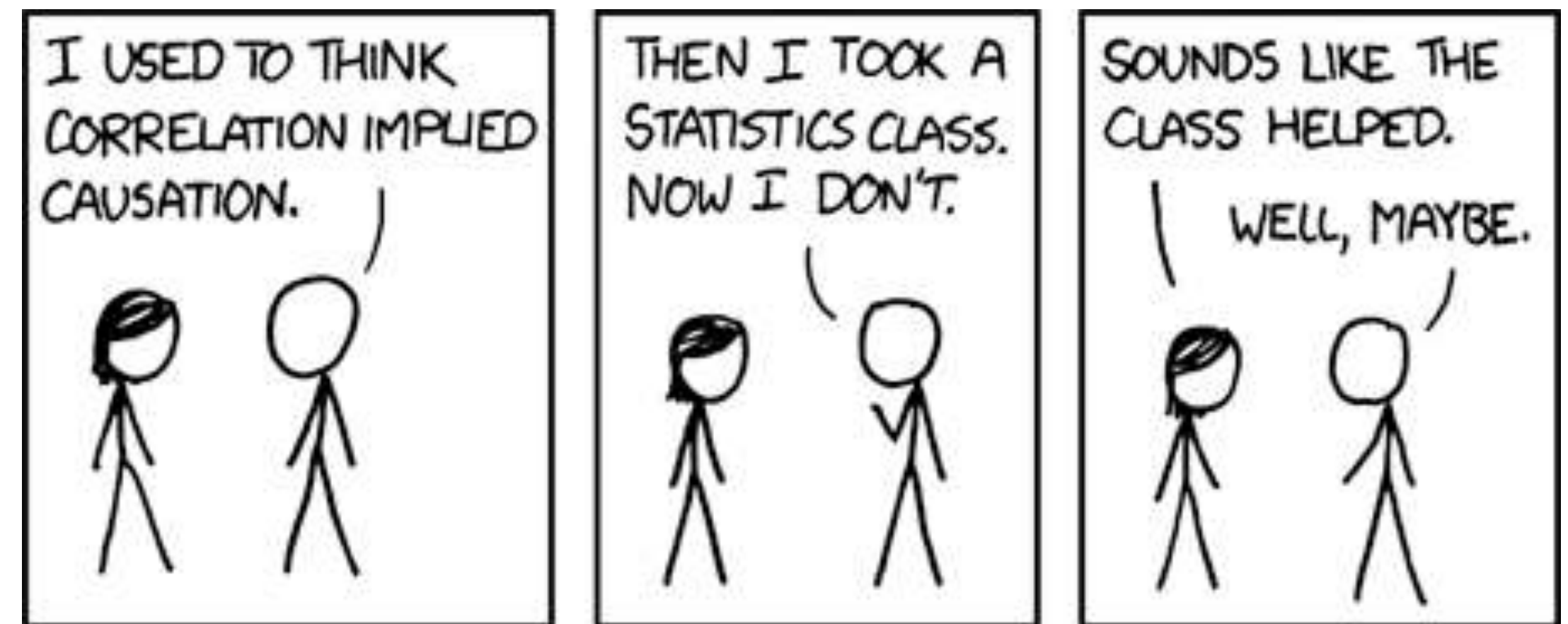
**Same factors are  
easier to decompose  
into sub-factors**

# Something Seems Wrong

Are We Disentangling By Chance?

- Distances between ground-truth **factors naturally align with** distances between **observations** in the datasets based on the loss
  - Does this mean everything is happening by accident?
  - Can we purposely break VAEs by carefully designing the input data?
- Is this why these methods struggle on real world data?

Then... Again...



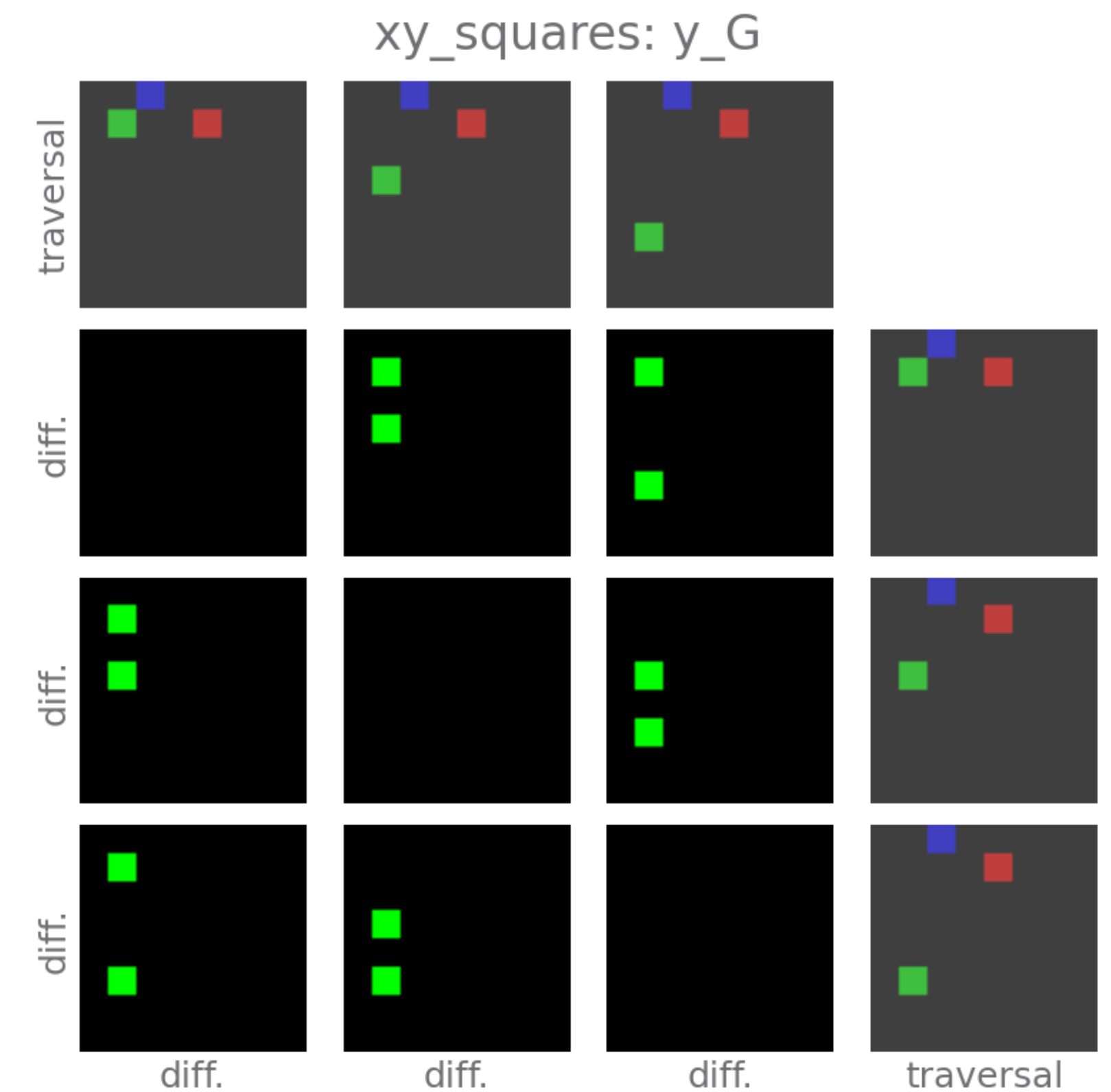
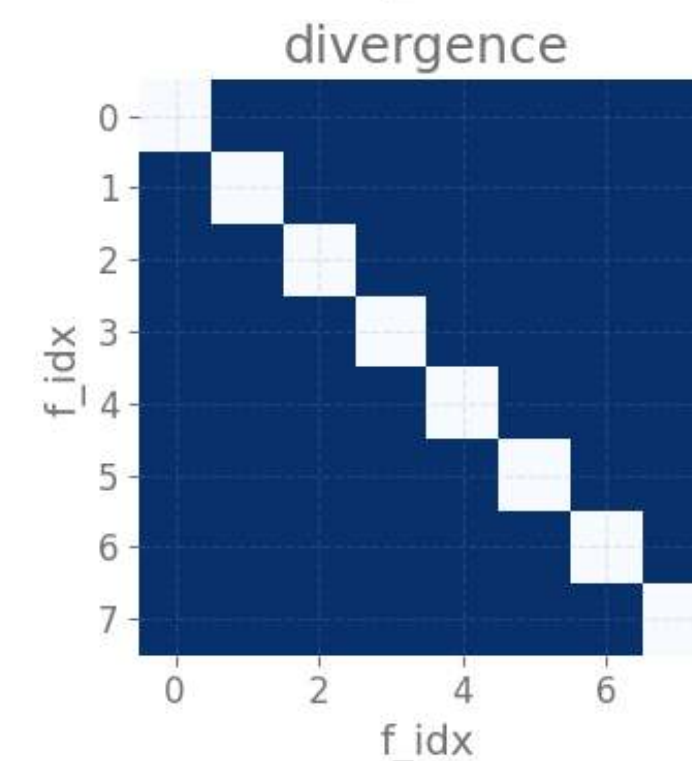
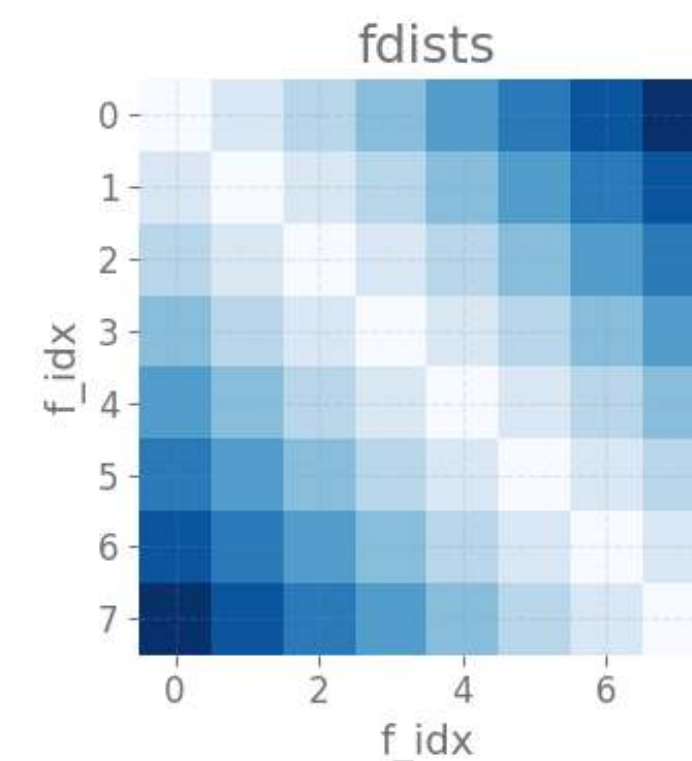
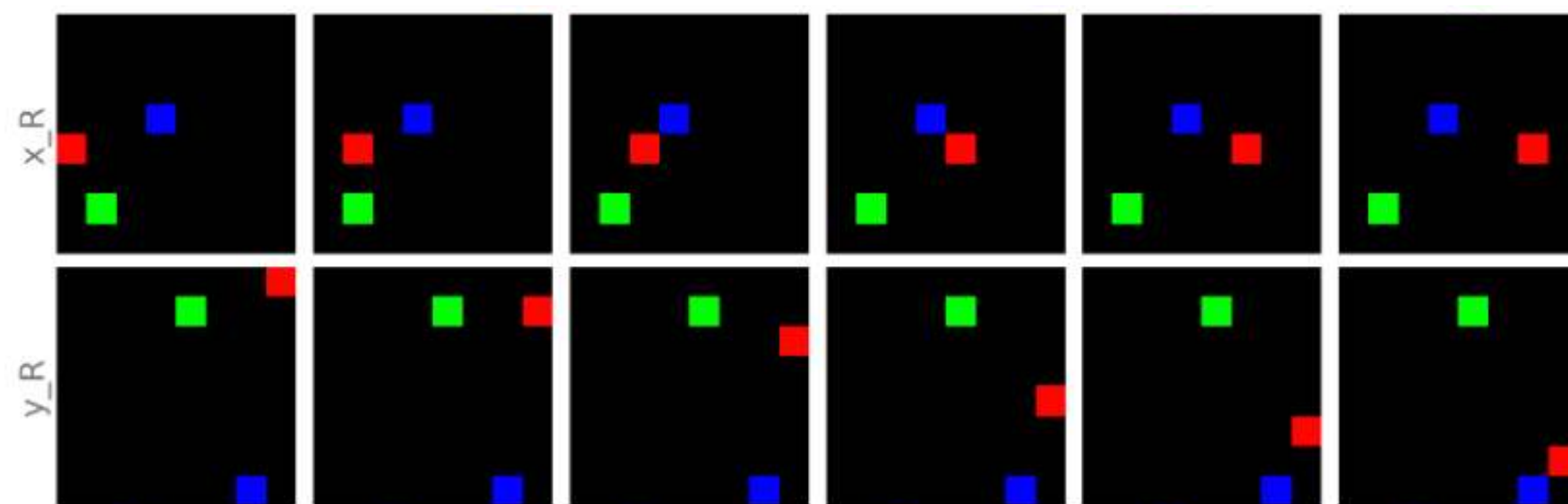
XKCD: Correlation from <https://xkcd.com/552>

## **-3. Adversarial Dataset**

# Constant Overlap

How Can We Fool The Frameworks?

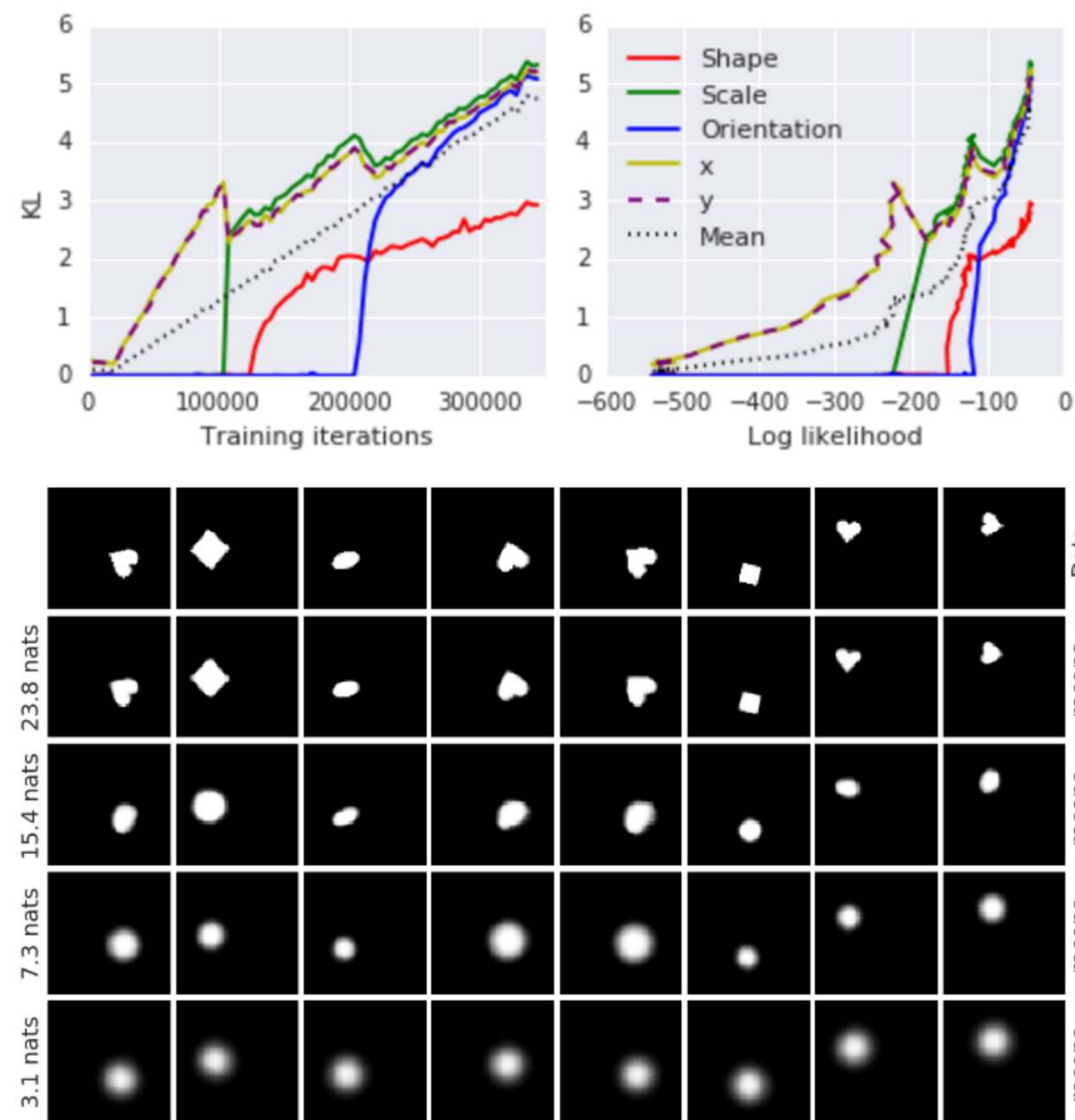
- Design dataset with **constant overlap along factor traversals**
  - Minimising reconstruction error due to sampling **cannot induce ordering**
  - “no overlap” is technically a wrong description for this dataset



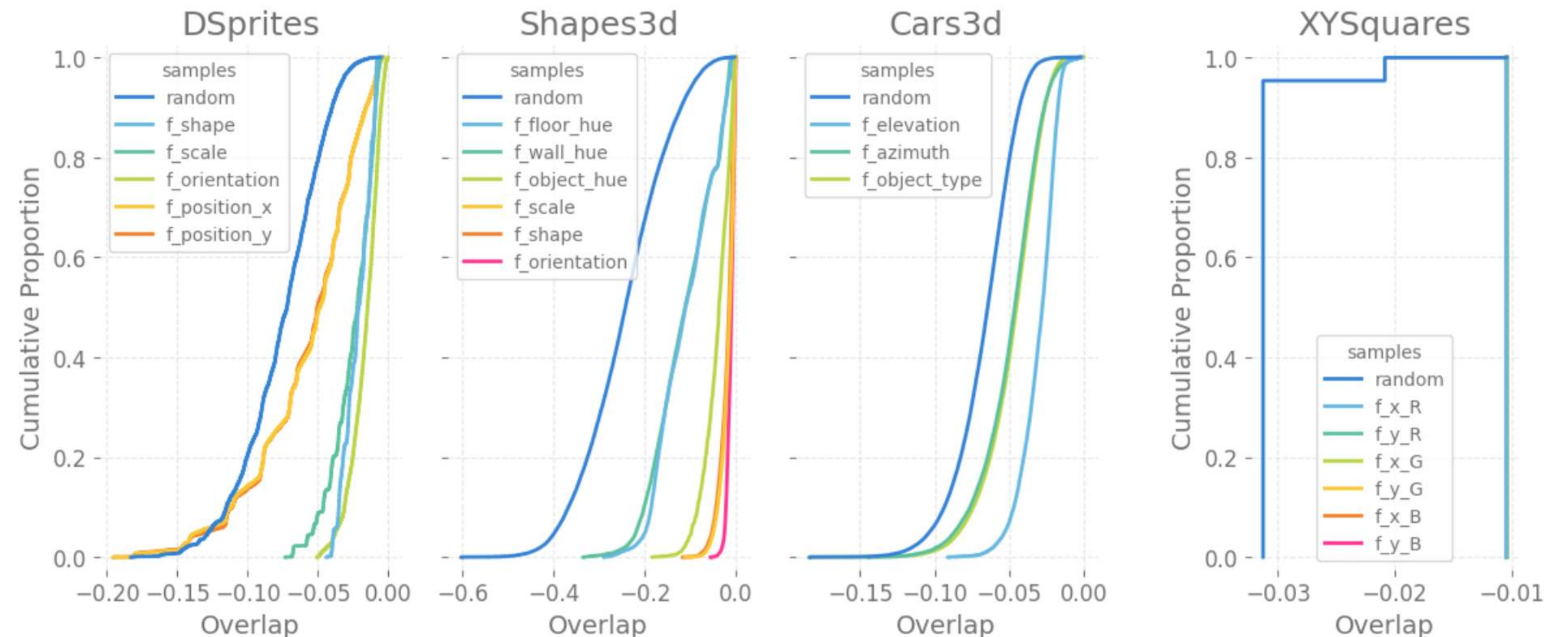
Same amount of overlap  
between each obs. in a  
factor traversal

# Overlap In Datasets

The Most Important Factors Can Be Found By Looking At Their Overlap



**From:** Understanding Disentanglement in Beta-VAEs



(a) Common Disentanglement Datasets

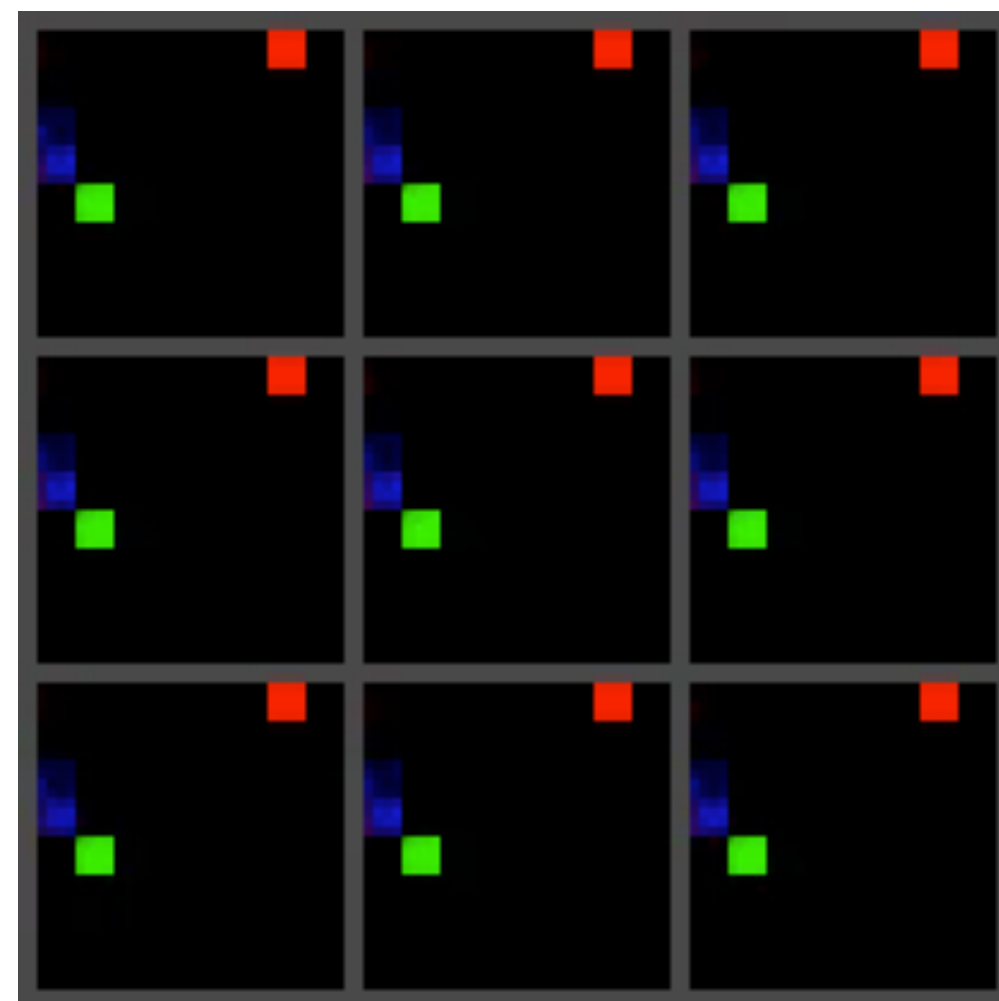
(b) Adversarial Dataset

$$\mathcal{L}_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})$$

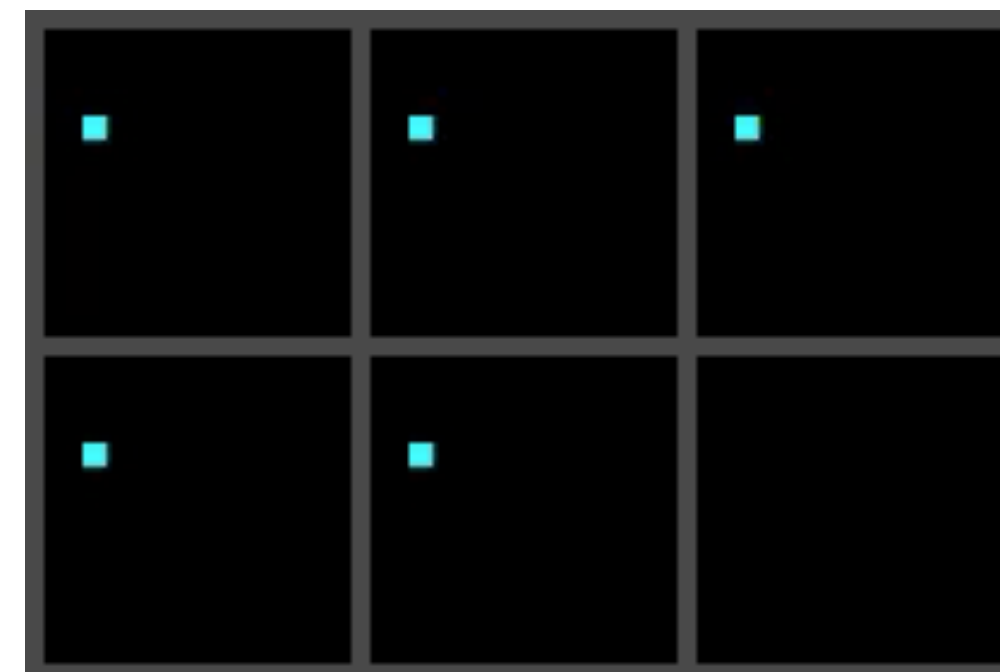
**Our Results:** Direct Pairwise Loss Between Dataset Samples

# Adversarial Results

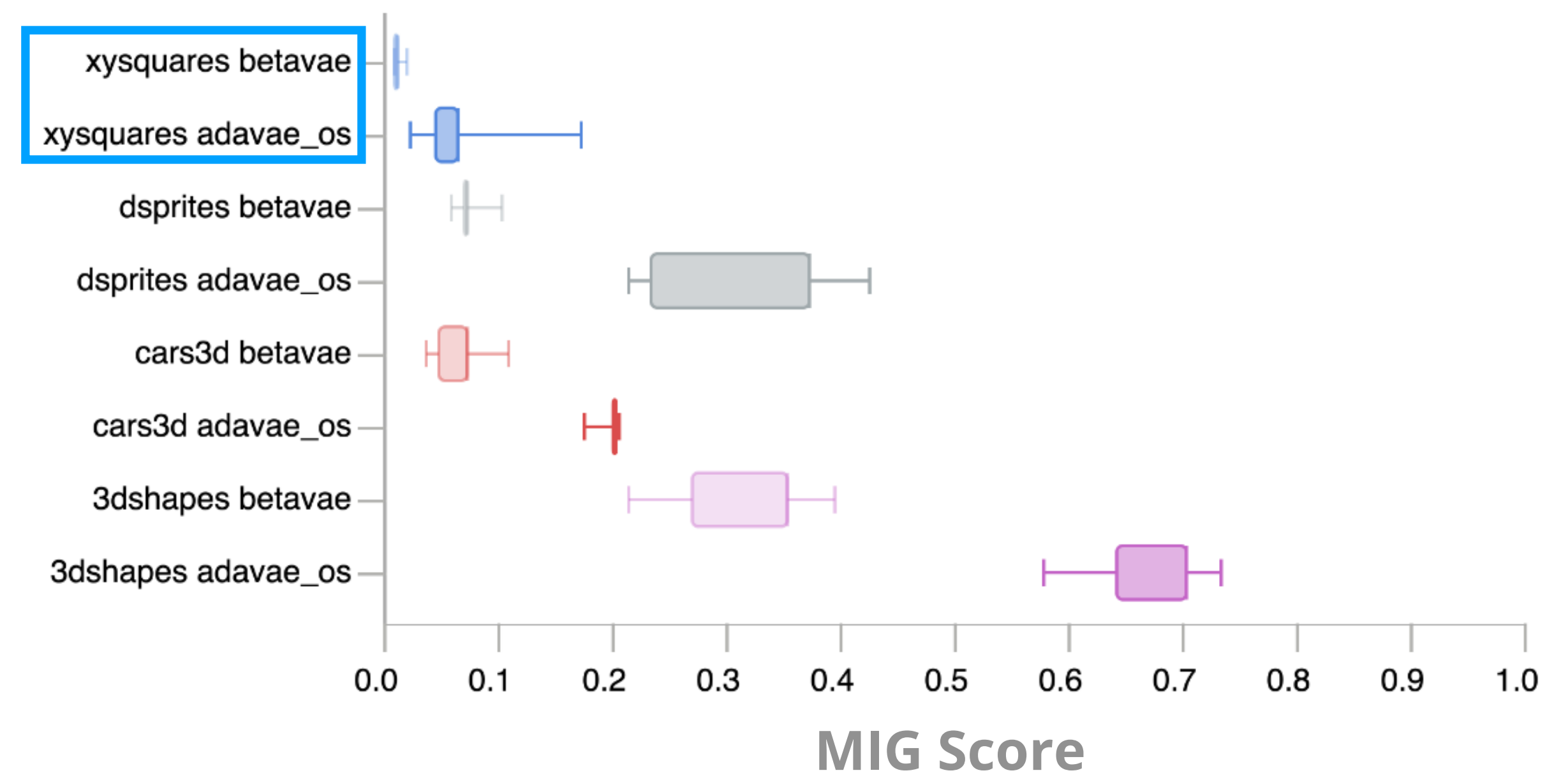
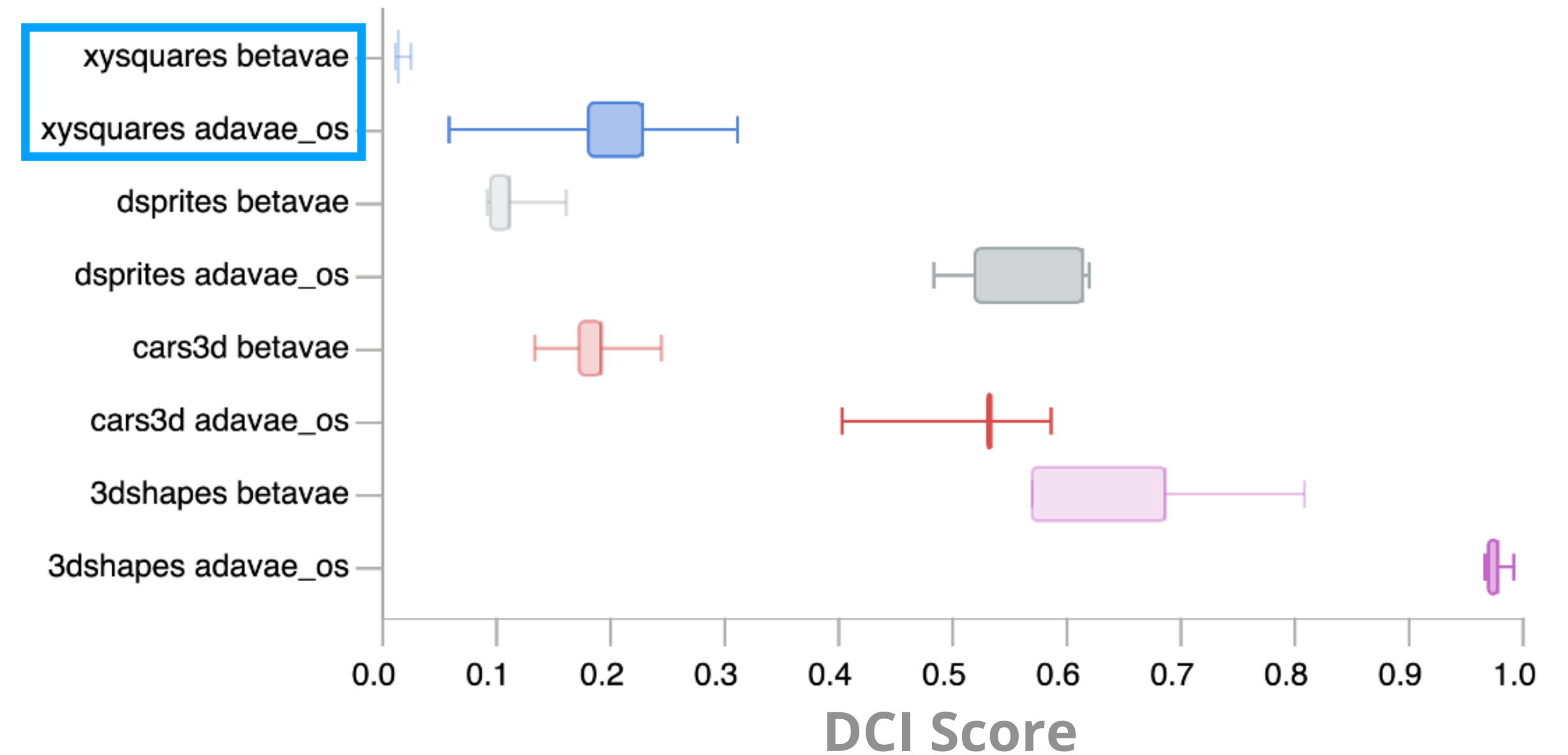
Everything Breaks!



**Non-Overlapping Dataset**  
Ada-GVAE Latent Traversal

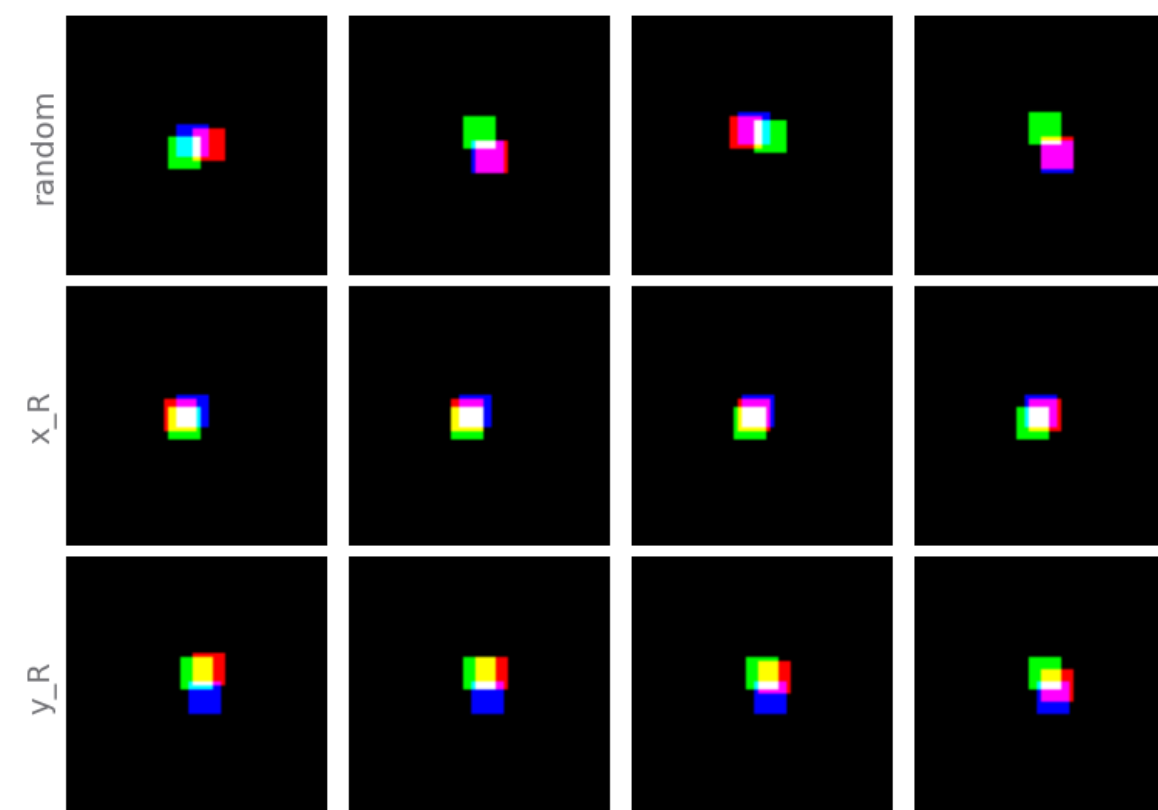


**Overlapping Dataset**  
Ada-GVAE Latent Traversal

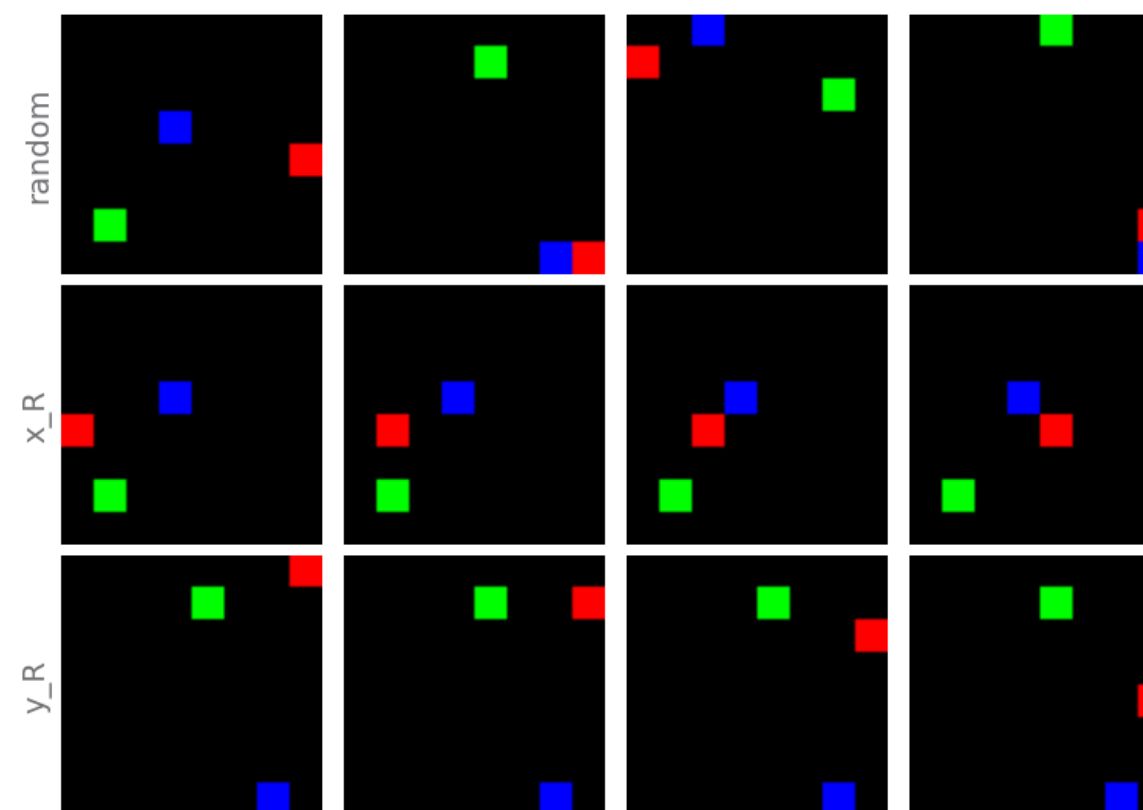


# Increasing Overlap

Changing The Data

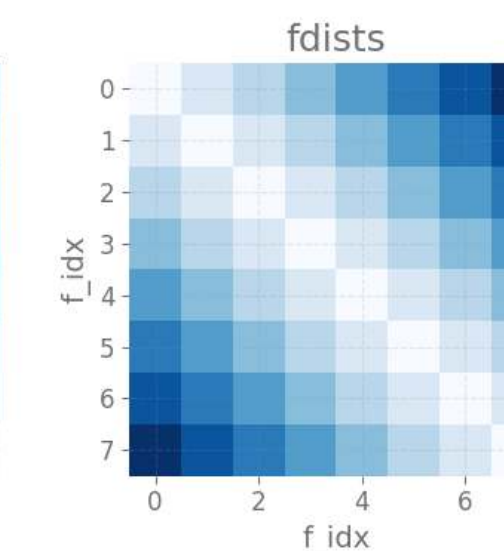
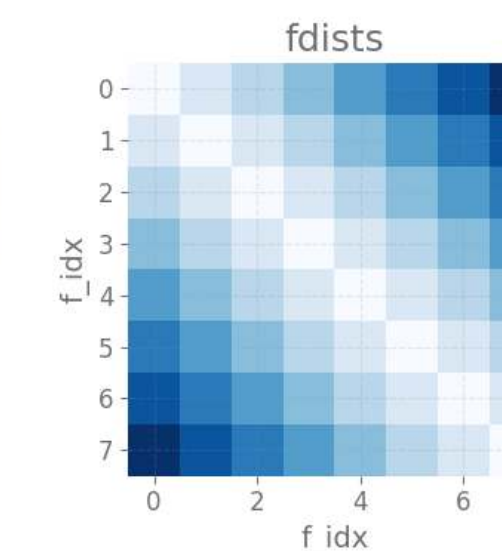
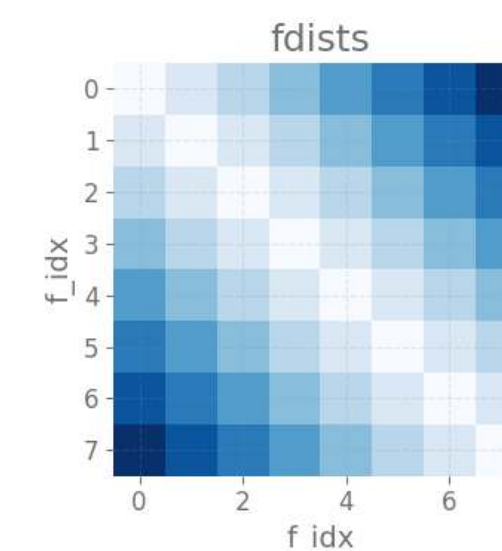
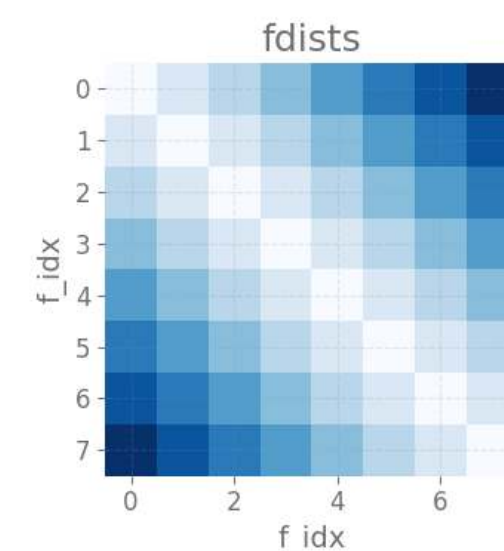
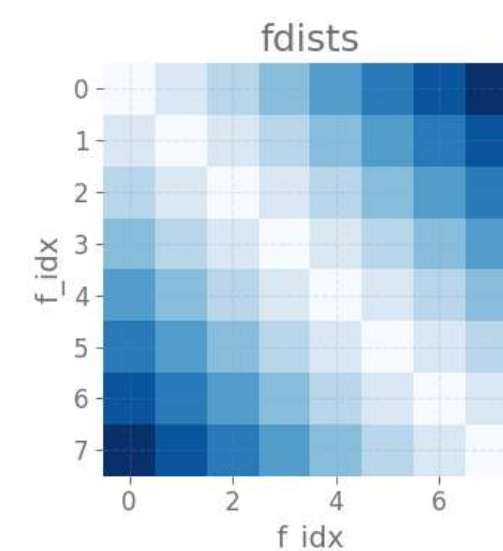
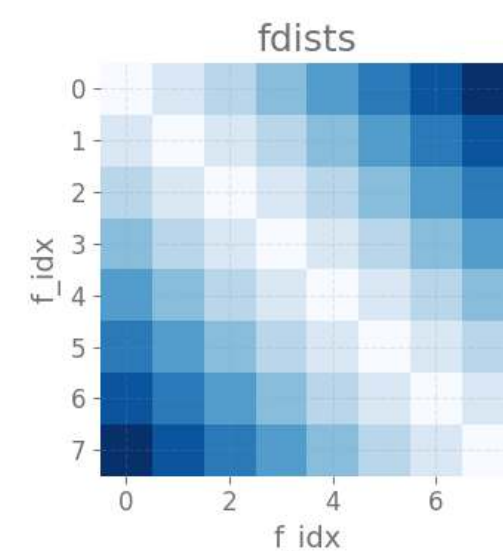
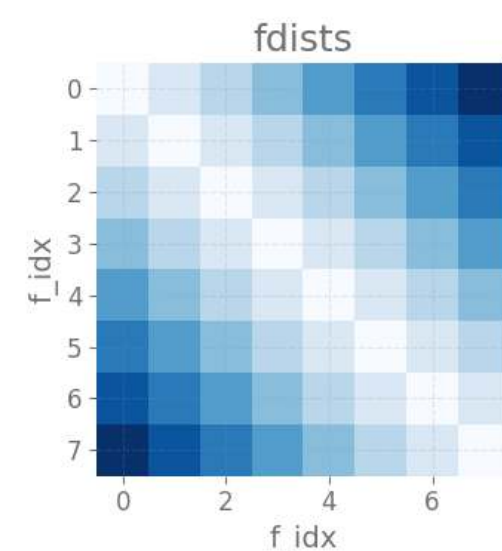
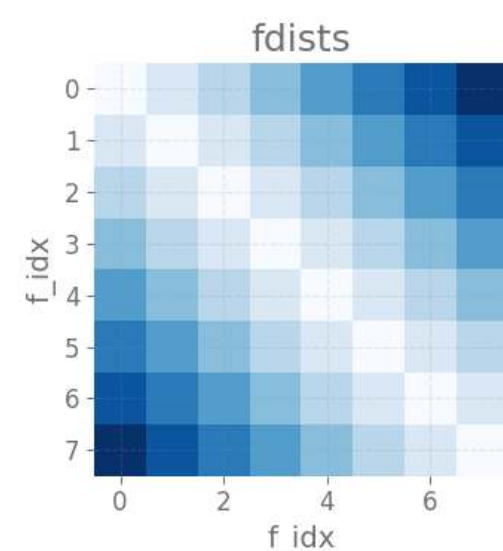


Overlapping Dataset  
7 Pixels Overlap

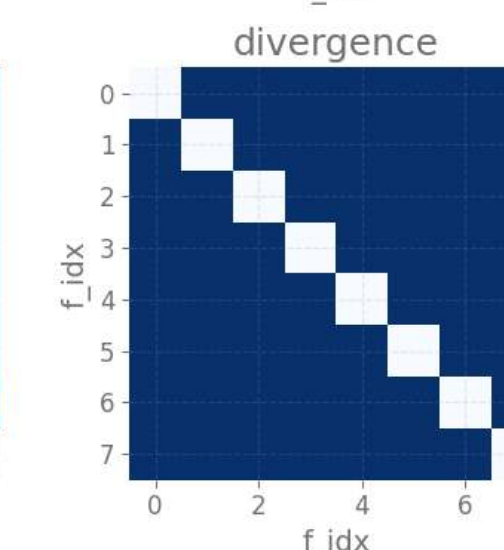
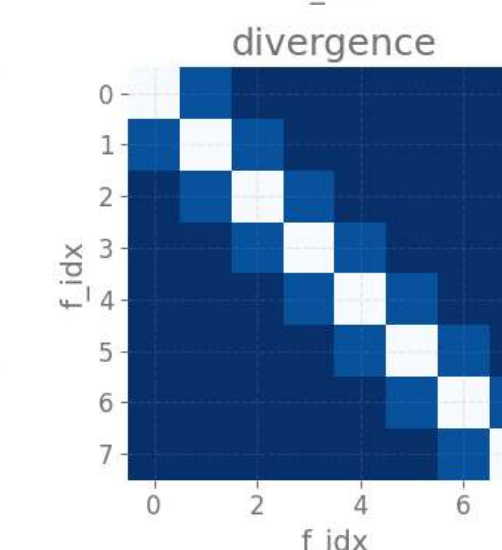
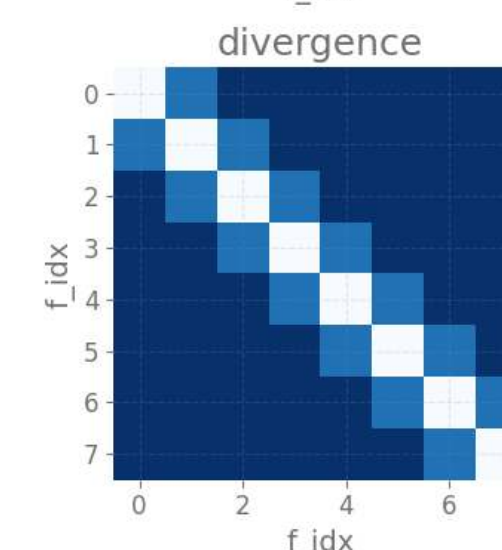
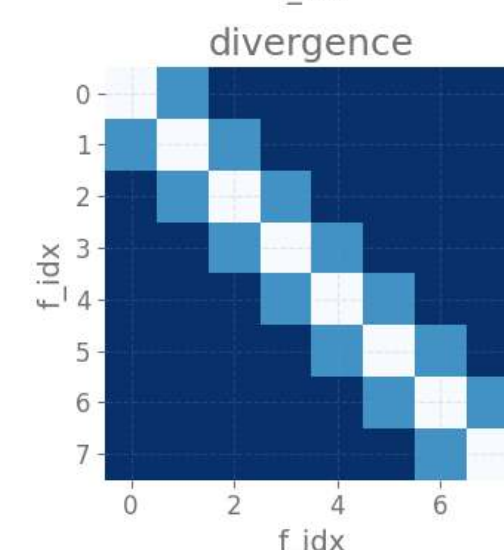
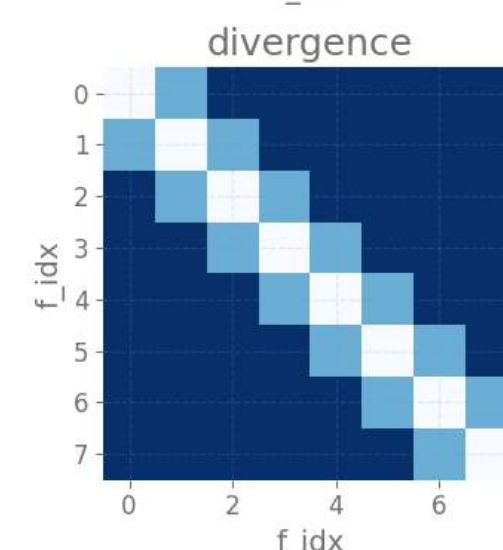
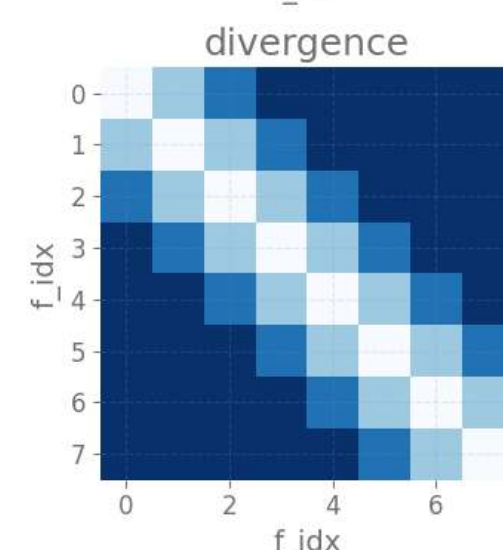
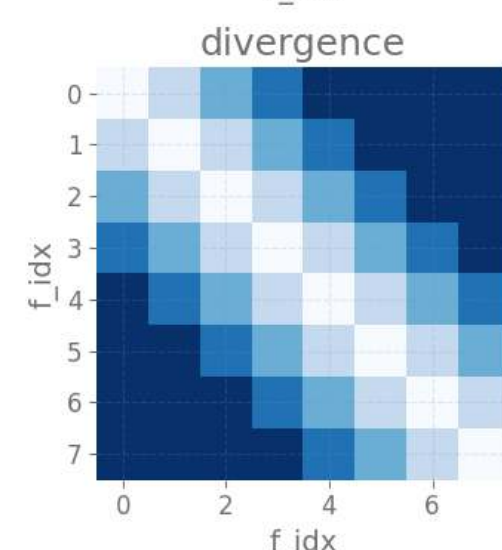
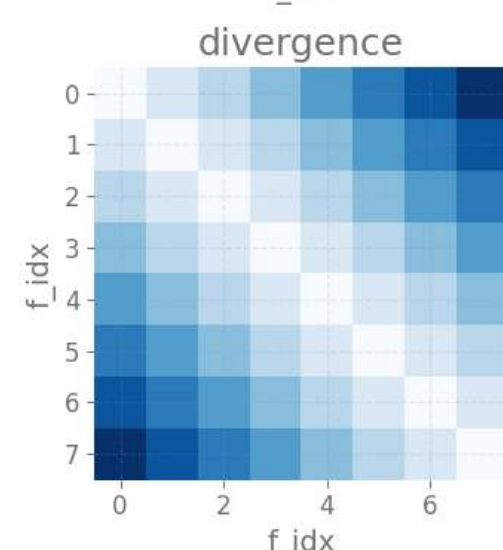


Original Adversarial Dataset  
0 Pixels Overlap

Factor Dist.  
Mat.



Overlap Dist.  
Mat.



7px overlap

6px overlap

5px overlap

4px overlap

3px overlap

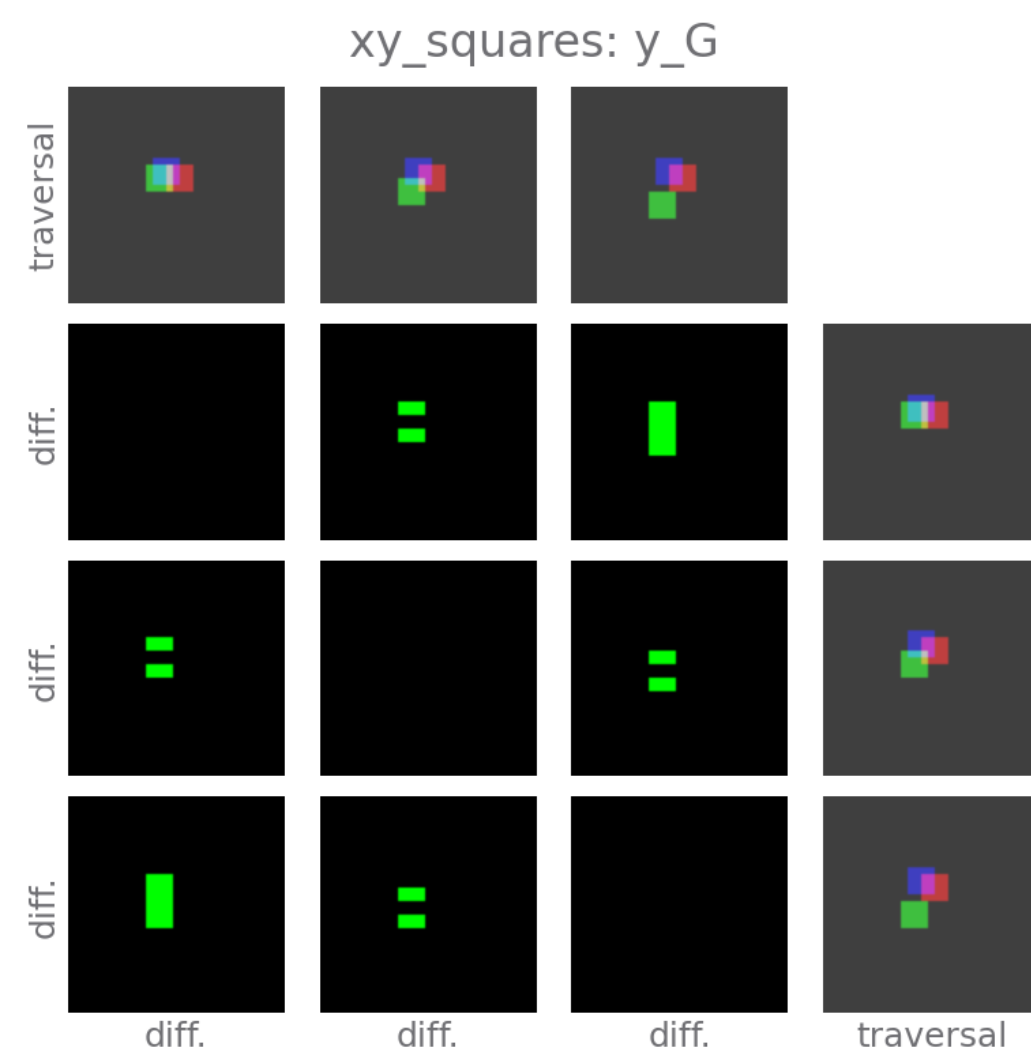
2px overlap

1px overlap

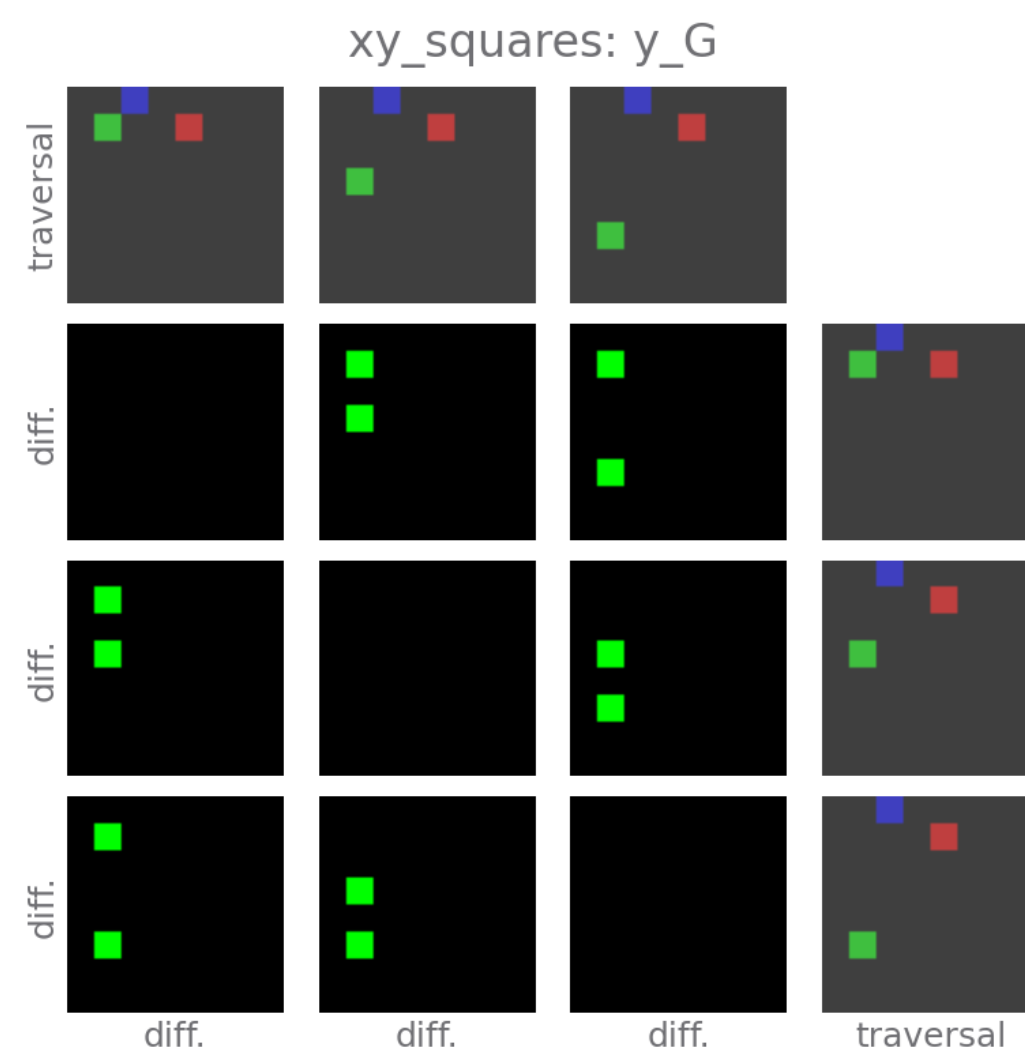
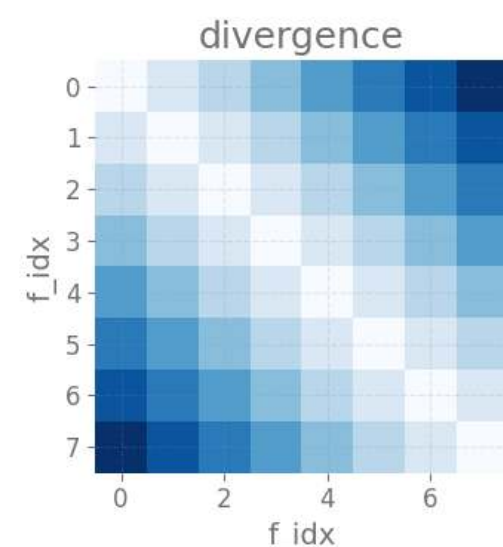
0px overlap

# Increasing Overlap

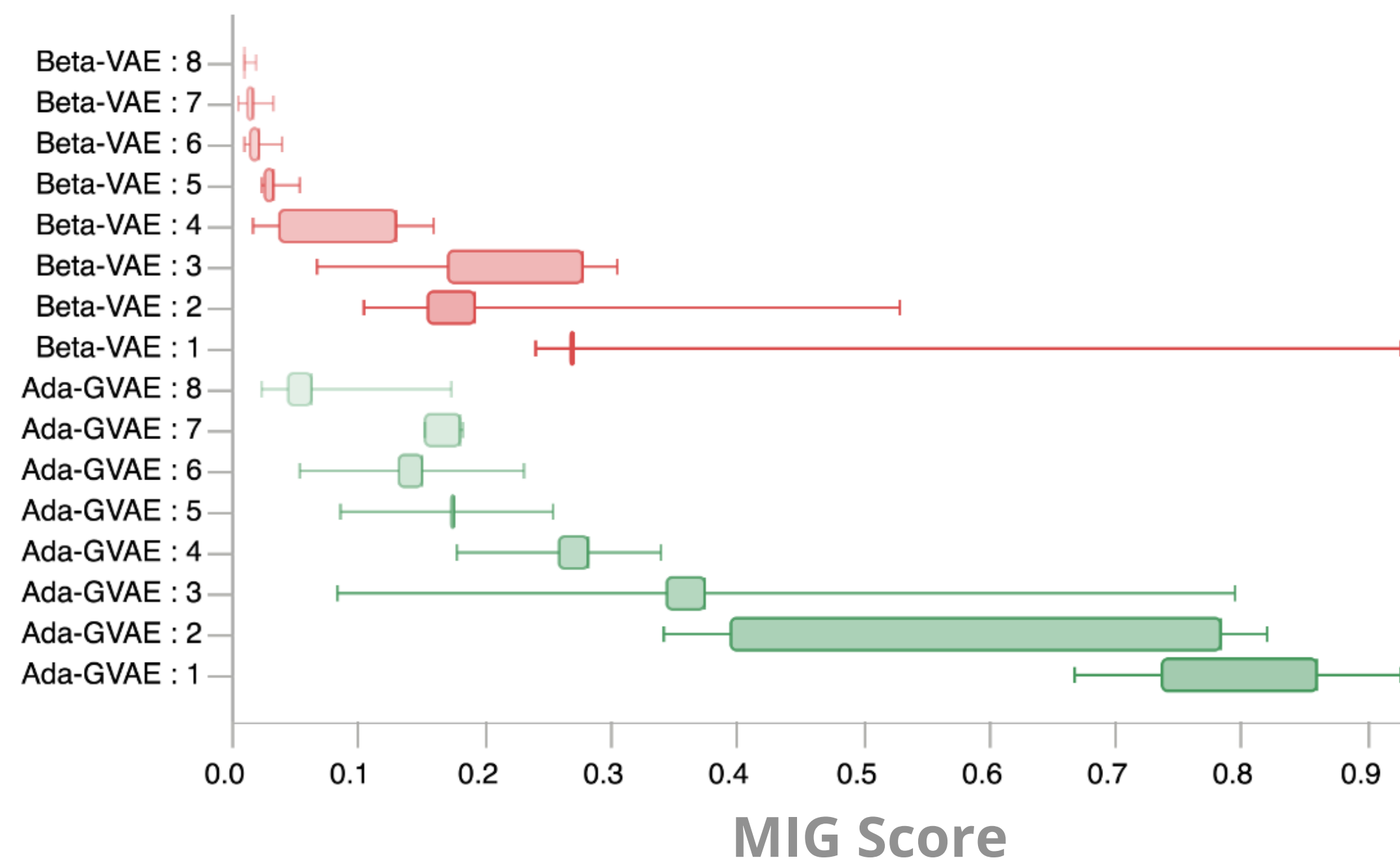
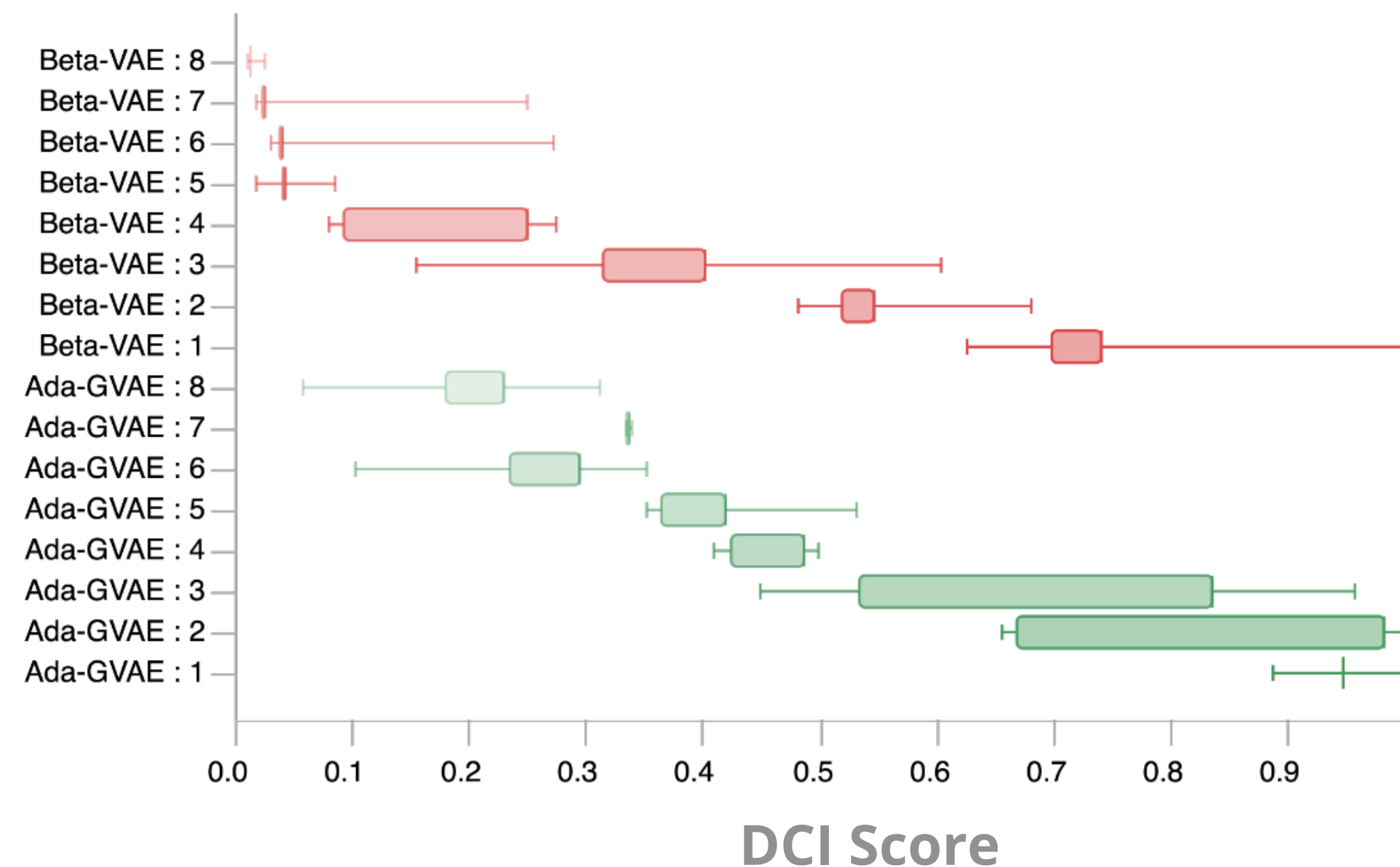
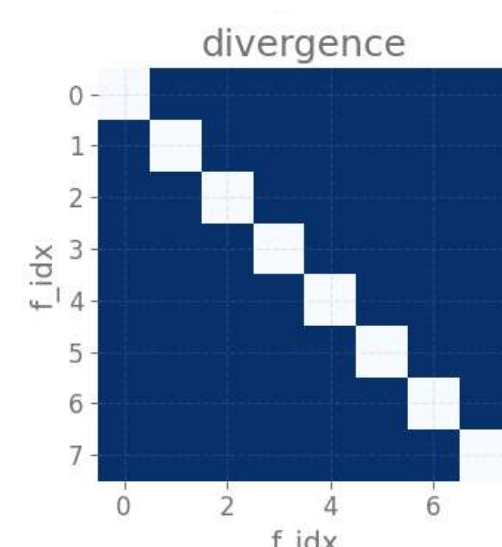
Constant Overlap Breaks SOTA



Overlapping Dataset  
7 Pixels Overlap



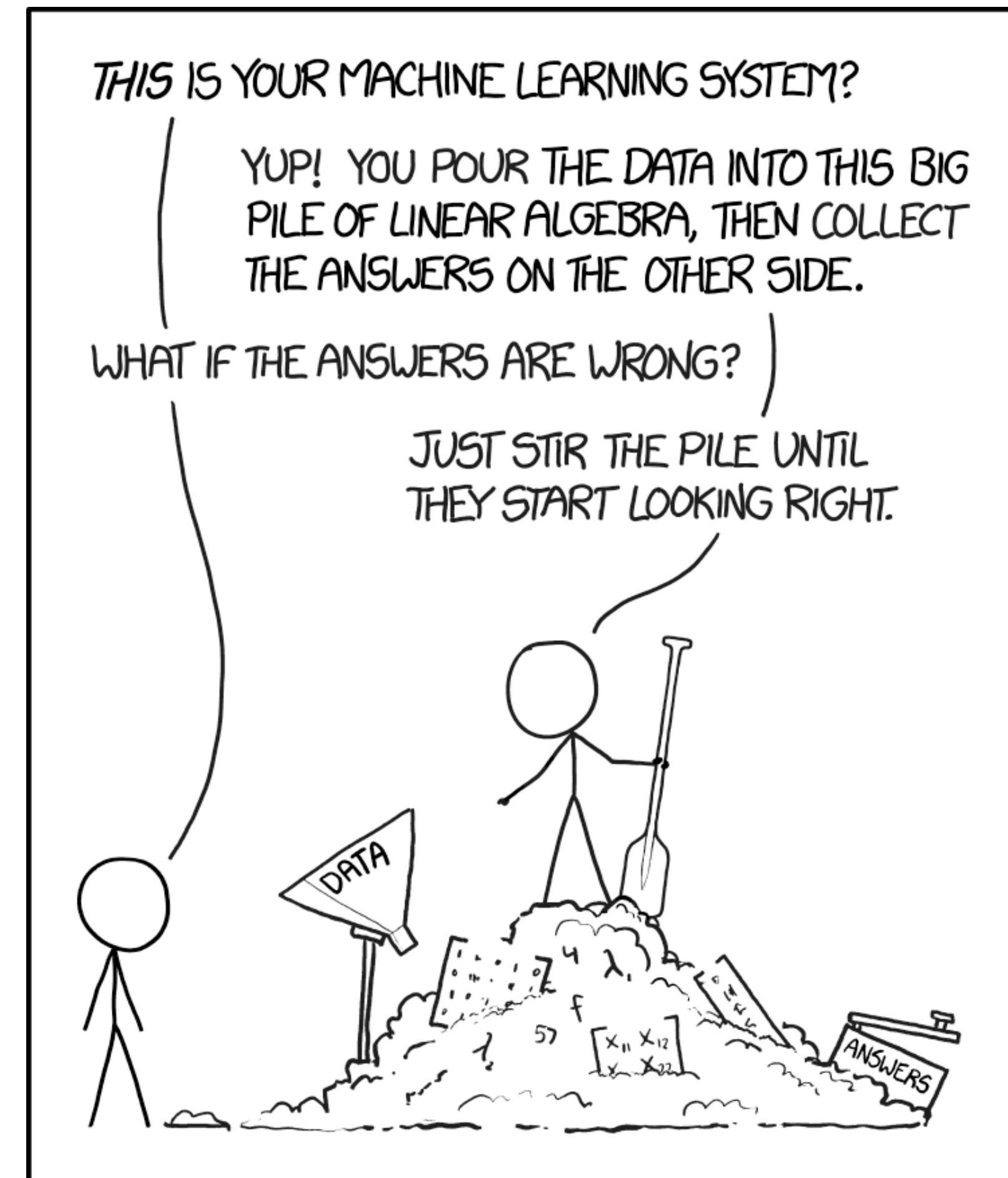
Original Adversarial Dataset  
0 Pixels Overlap



# Can We Disentangle The Adversarial Data?

We Could Use Metric Learning?

- Are VAEs Local Metric Learners?
  - Linear auto-encoders perform **PCA**
  - Observation **distances correspond** to ground-truth factors
  - Ordering is induced by **minimising sampling error** of reconstruction loss
- Can we use **metric learning** directly for disentanglement? (No decoder)
  - Start with **supervised** approach **first**. Then we can work on unsupervised.

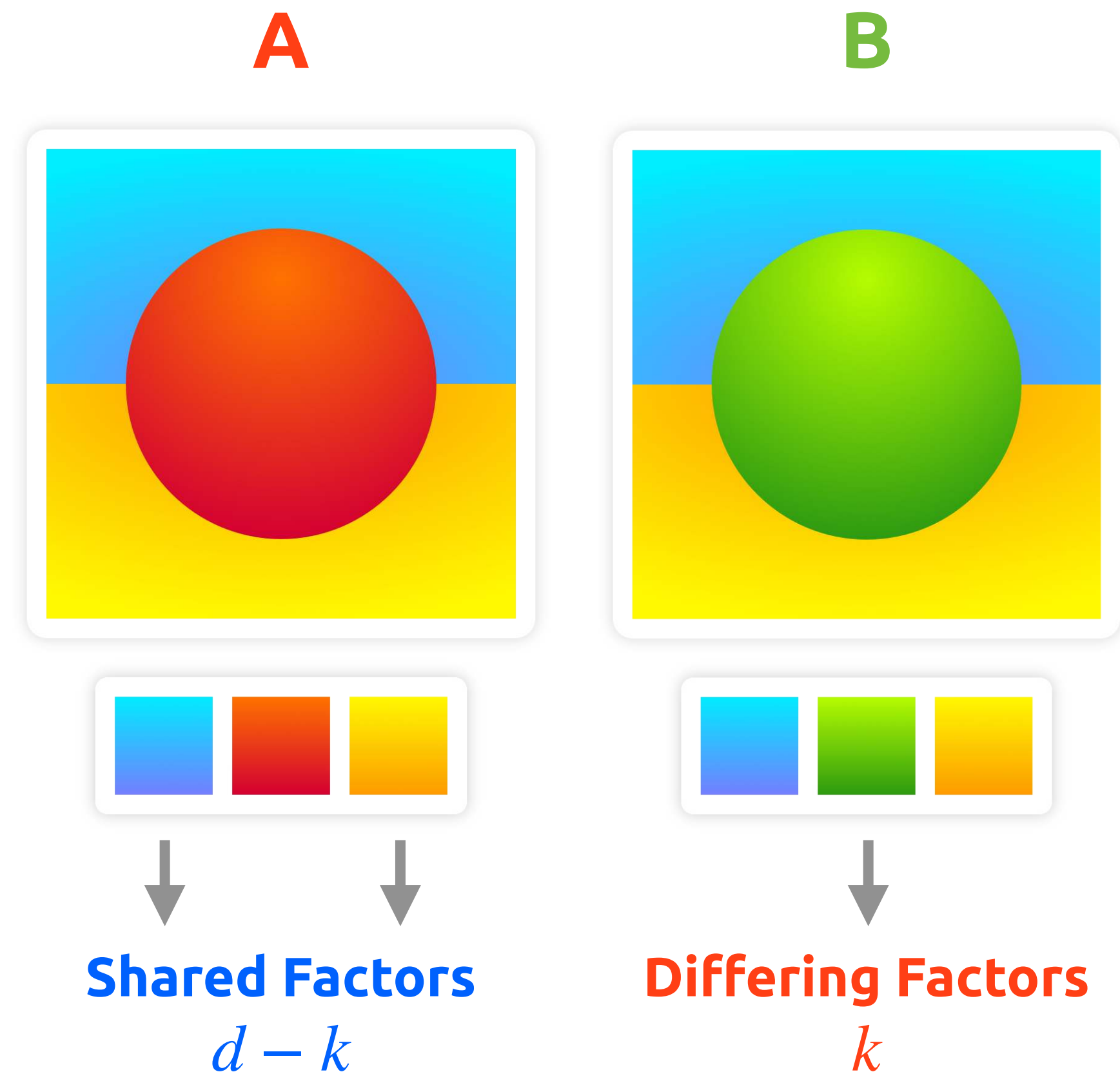


## **-4.** Background: **Adaptive VAEs**

# Adaptive VAEs

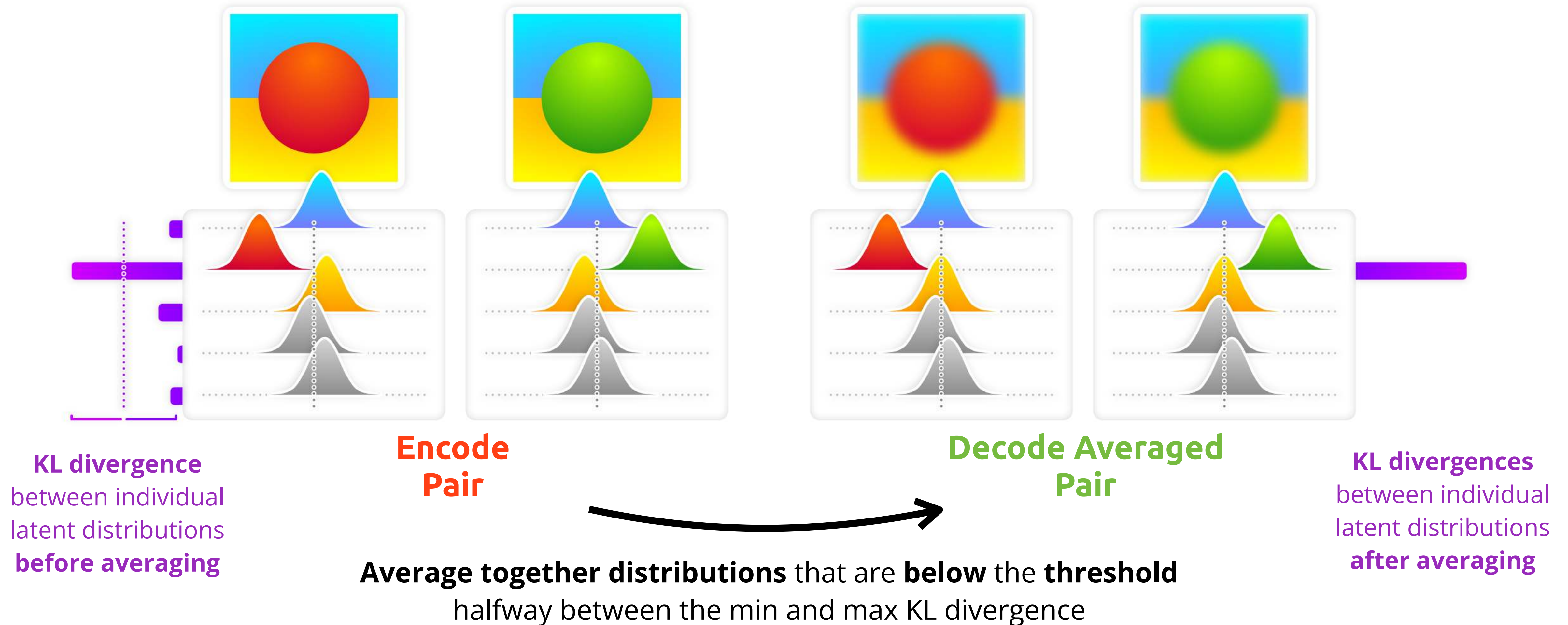
Weakly-Supervised Disentanglement Without Compromises <sup>[1]</sup>

- **Estimates which factors have changed** between observation pairs.
  - **Encourages** the desirable property of partially **shared representations**
  - Variants are: Ada-GVAE & Ada-ML-VAE changing how latents are averaged
- **Weakly-Supervised** by sampling pairs
  - Assumes **at least one factor** in a pair is **shared**



# Adaptive VAEs

## Algorithm Overview - Averaging



## **-5. Supervised Adaptive Triplet**

# Triplet Loss

Background: Standard Triplet Loss

## Goal

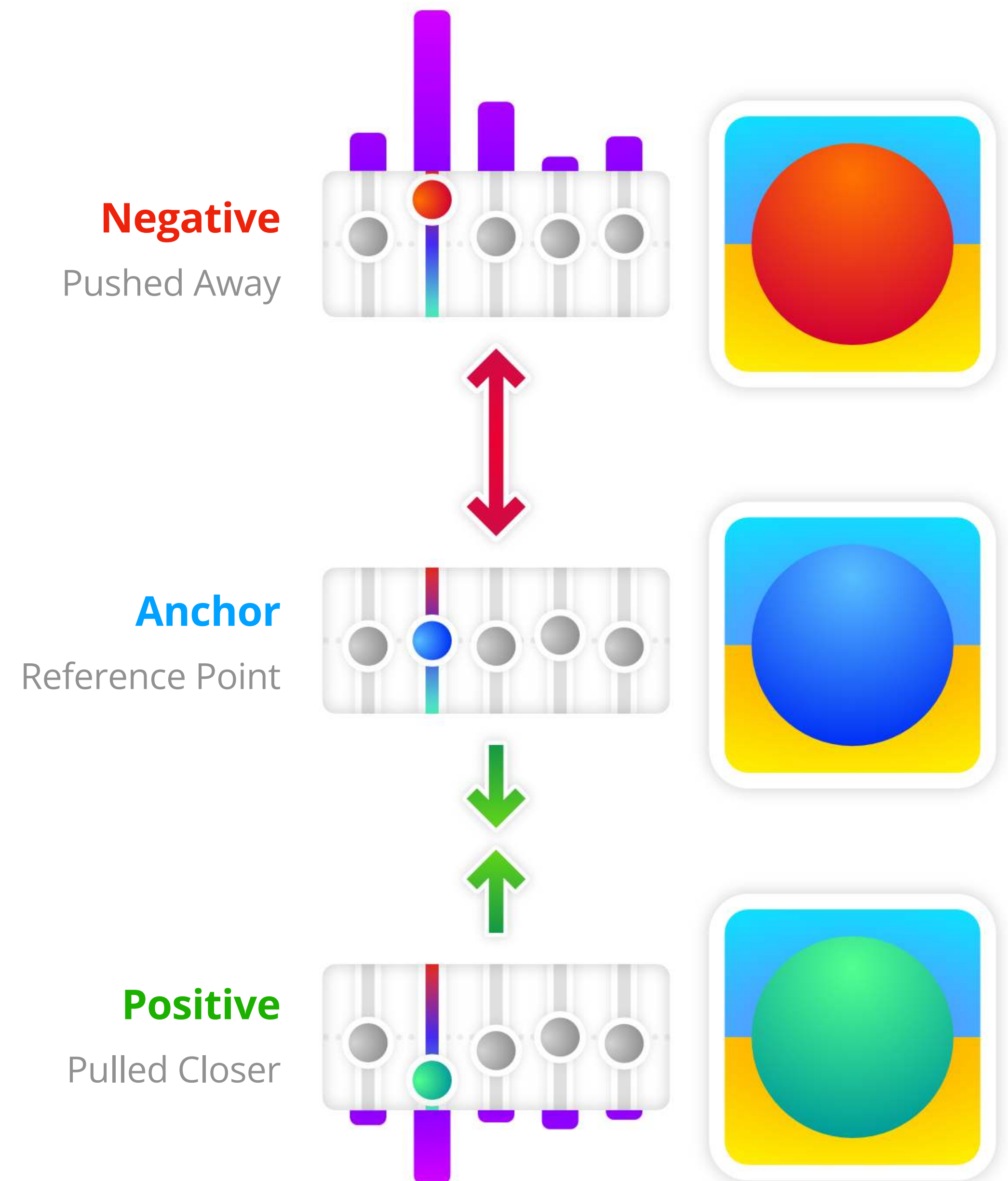
- Learn **embeddings** that correspond to some notion of similarity between **labeled observations**

## How

- Push **negative** representation further away
- Pull **positive** representation closer together
- Negative has no effect beyond **margin**
- Applied to mean of distributions:  $\mathbf{z} = \mu$

$$\mathcal{L}_{triplet} = \max( 0, \psi + d(\mu^{(a)}, \mu^{(p)}) - d(\mu^{(a)}, \mu^{(n)}) )$$

**\*NB\*** we no longer need a decoder, we just use it to visualise everything



# Adaptive Triplet Loss

Weight Shared Elements Less

## Goal

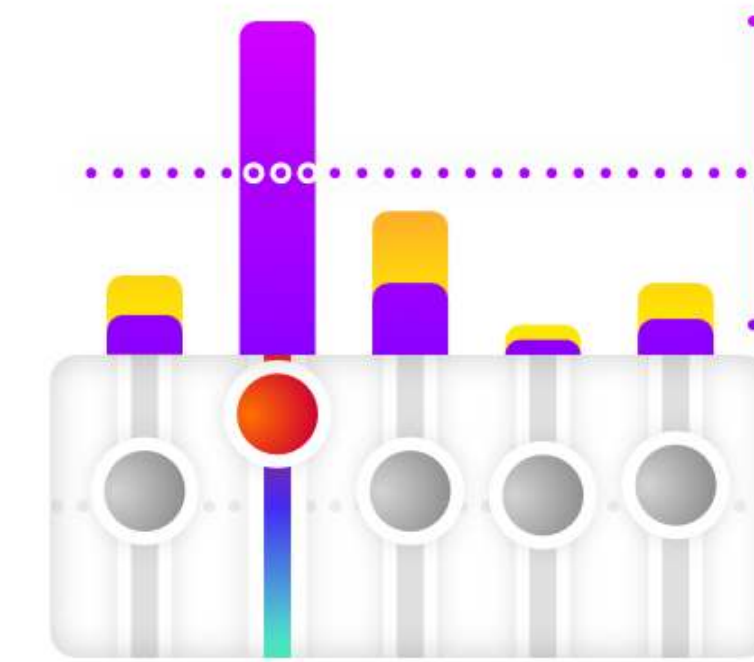
- Introduce axis alignment into triplet loss

## How

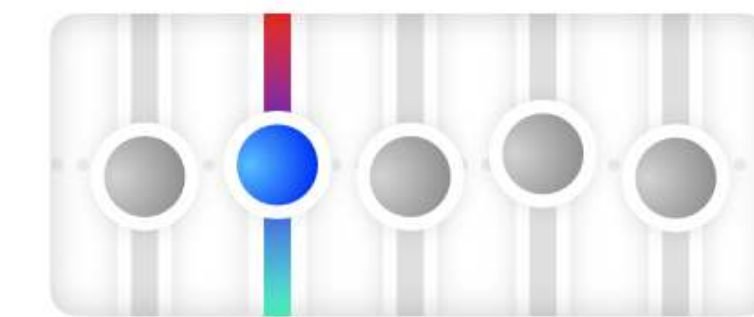
- **weight**  $\omega$  shared elements of **anchor-negative** distance less
- Unchanged **anchor-positive** “fixes mistakes”
- Replace **L2** distance with **L1** distance. Better in high dimensions and lets latent variables die

$$\mathcal{L}_{ada-triplet} = \max( 0, \psi + d(\mu^{(a)}, \mu^{(p)}) - d(\omega \odot \mu^{(a)}, \omega \odot \mu^{(n)}) )$$

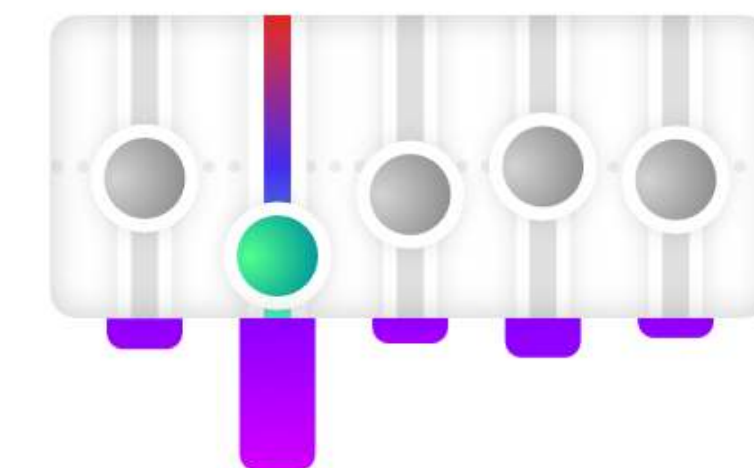
**Negative**  
Weighted & Pushed Away



**Anchor**  
Reference Point

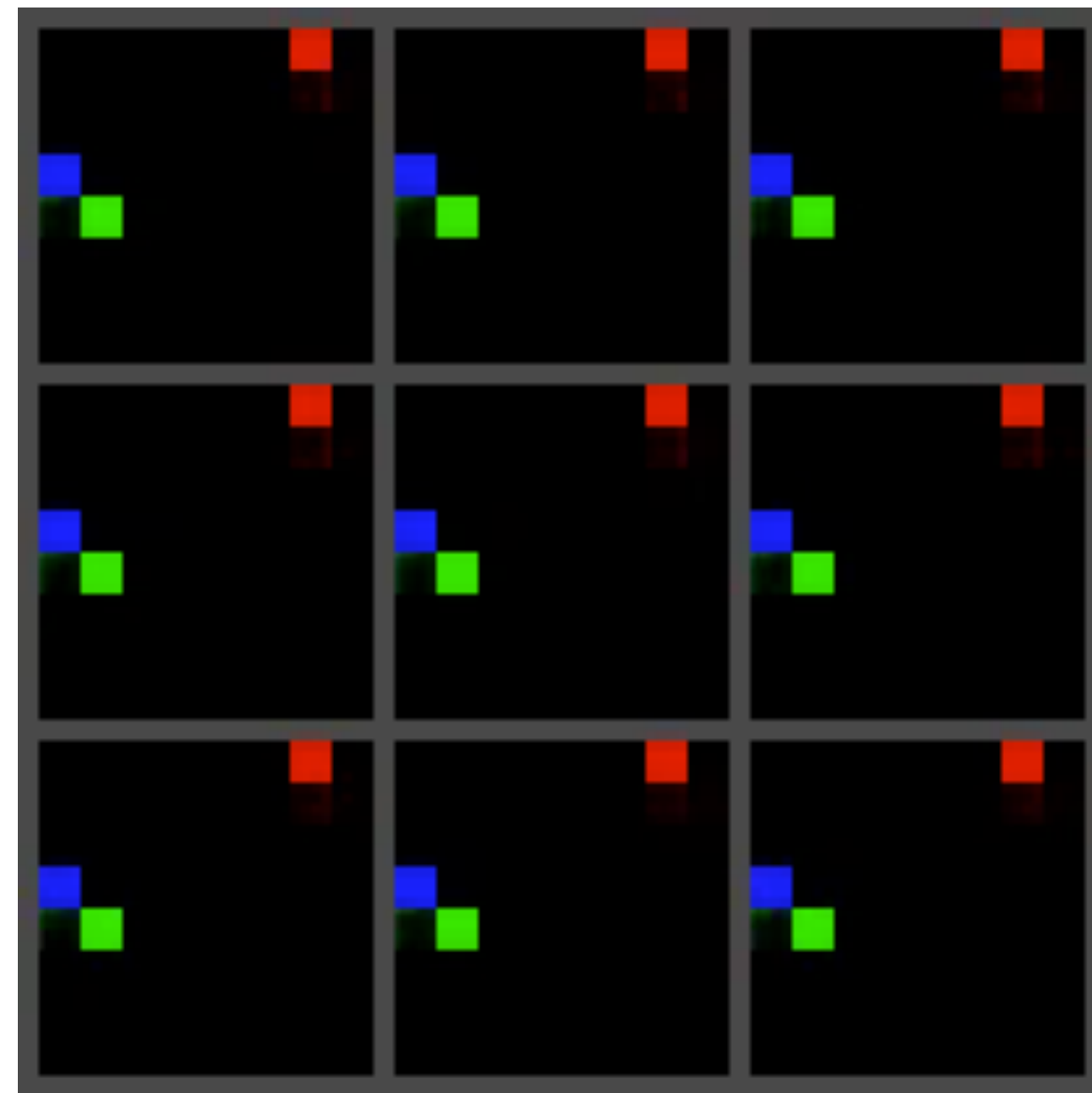


**Positive**  
Pulled Closer

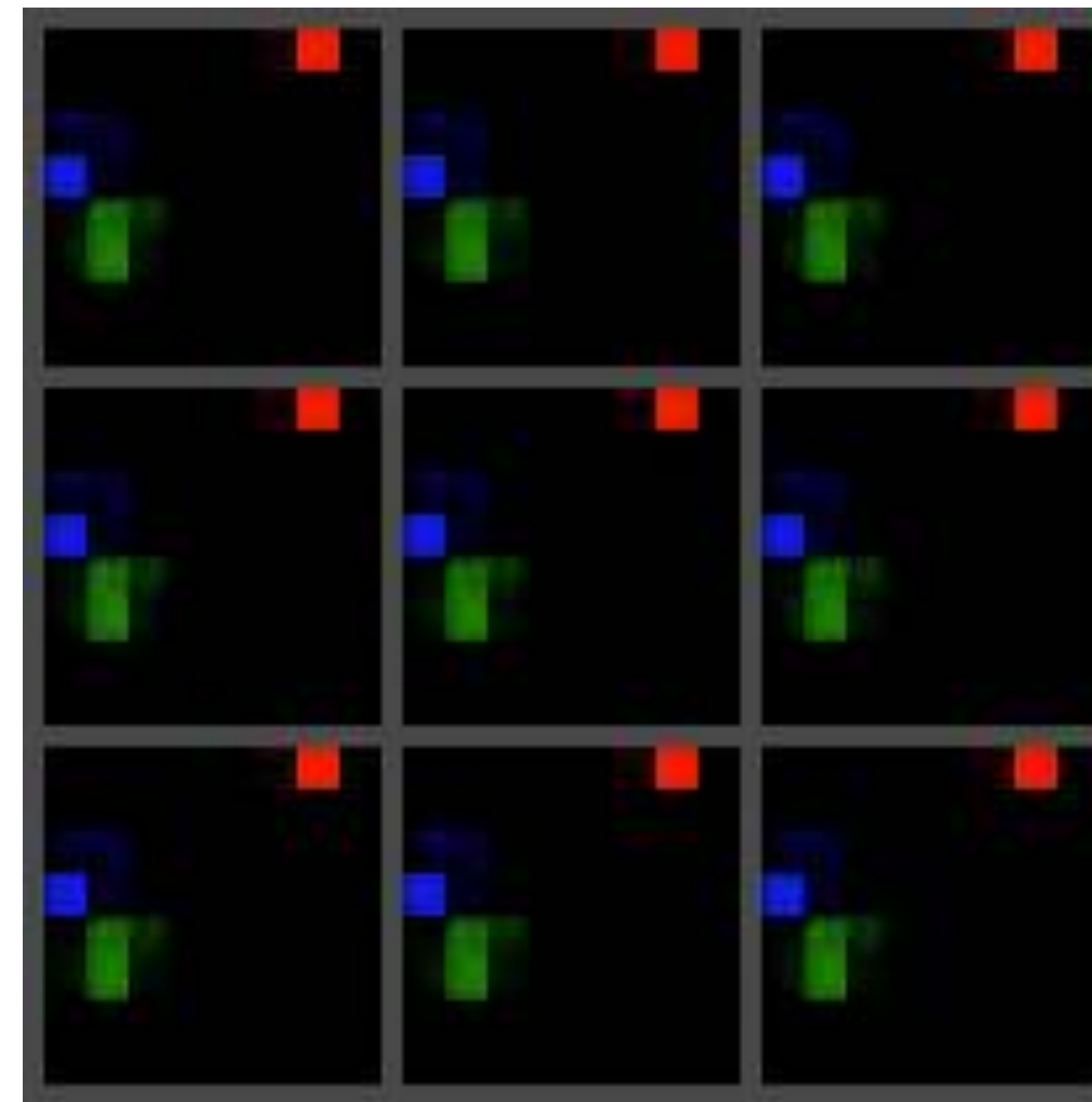


# Supervised Triplet Results

Detached Decoder & Slowly Increasing Averaging Threshold



Supervised **Triplet**

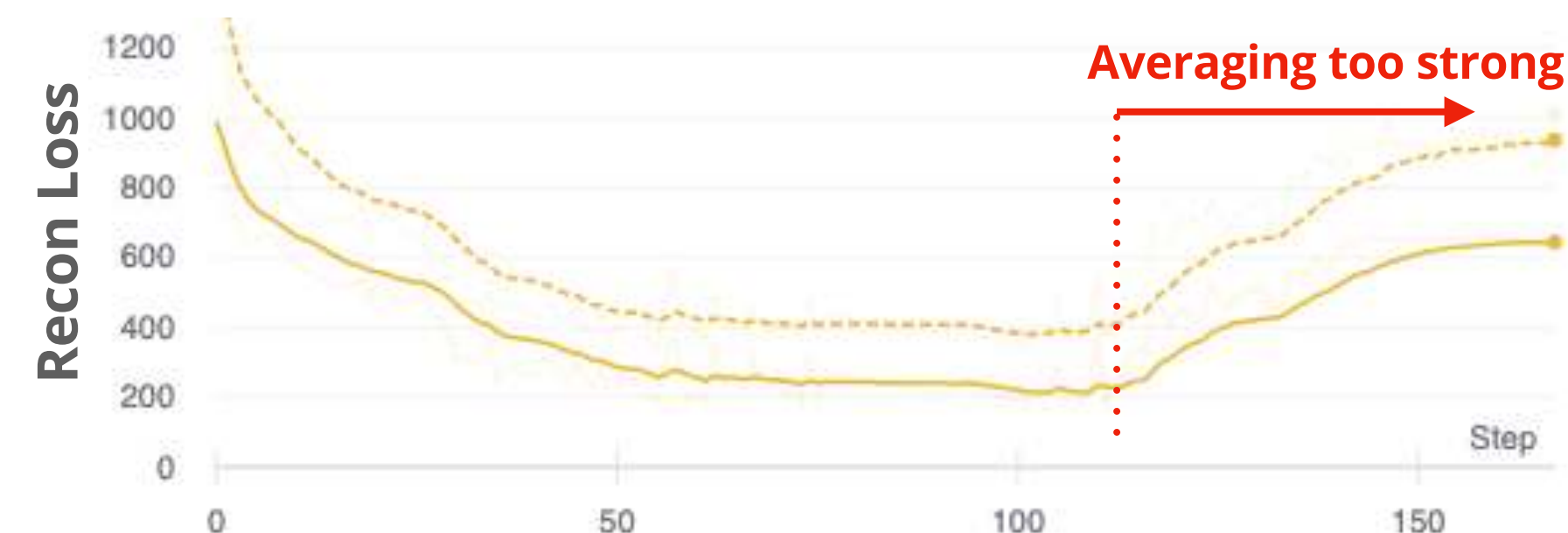


Supervised **Adaptive Triplet**



Supervised **Adaptive Triplet**

Squares move in the same pattern in different blocks - **not axis aligned**, but aligned along some arbitrary axis



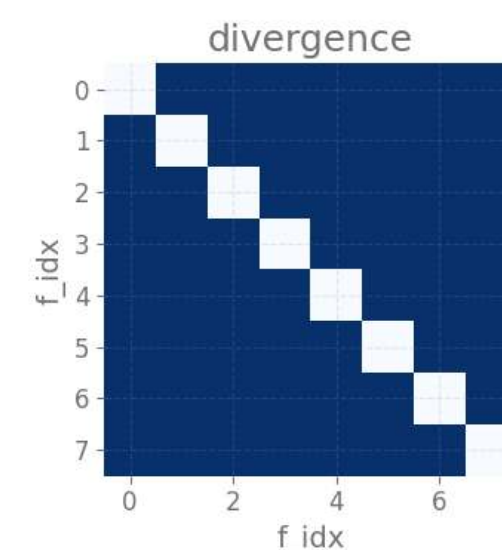
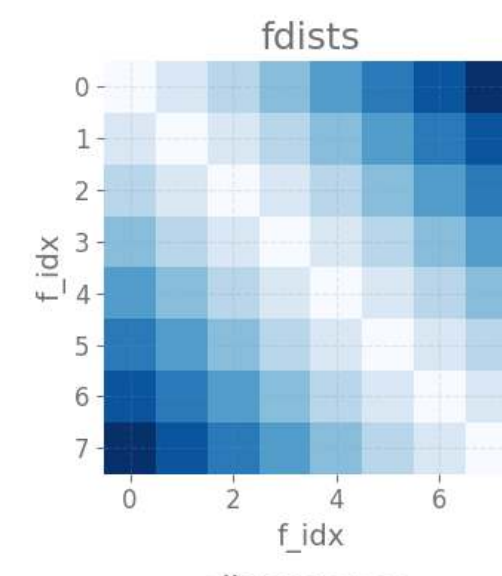
Increasing Ave. Threshold & Decreasing Ave. Weights

## **-6. Unsupervised Adaptive Triplet**

# Unsupervised Adaptive Triplet

## Order Triplets Using Visual Distance

- Use visual distance to sample triplets
  - **Measure visual overlap**/distance between anchor-positive and anchor-negative pairs
  - **Swap positive and negative** if distance is wrong
- Existing datasets are suited to this approach
  - **Adversarial dataset will still fail**
  - **Solution:** Move away from pixel-wise distance function!



### Supervised:

$$S_{\text{fdist}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = \|\mathbf{y}^{(a)} - \mathbf{y}^{(b)}\|_1$$

“Order triplets according to **Manhattan distance along factors**”

### Unsupervised:

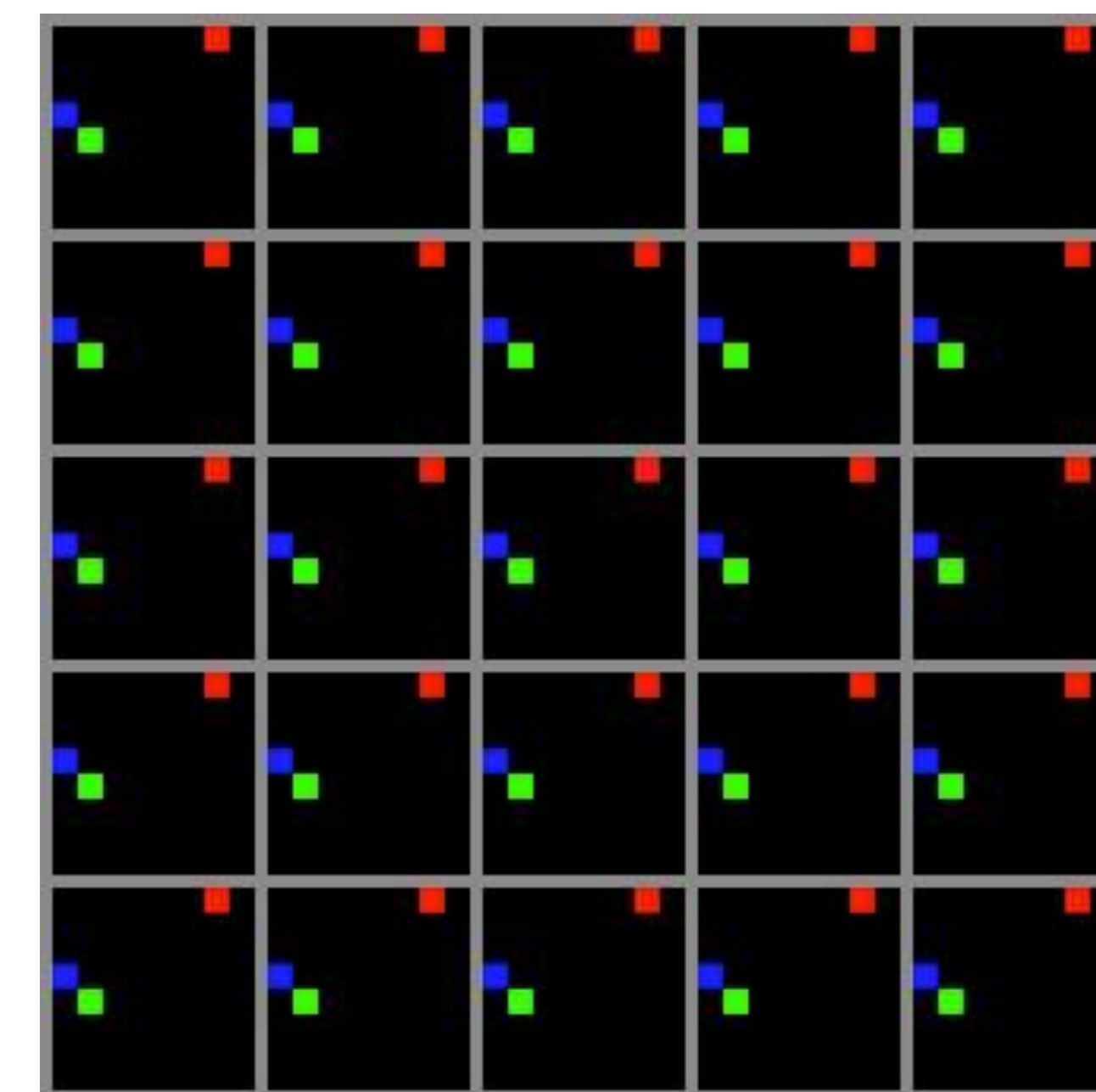
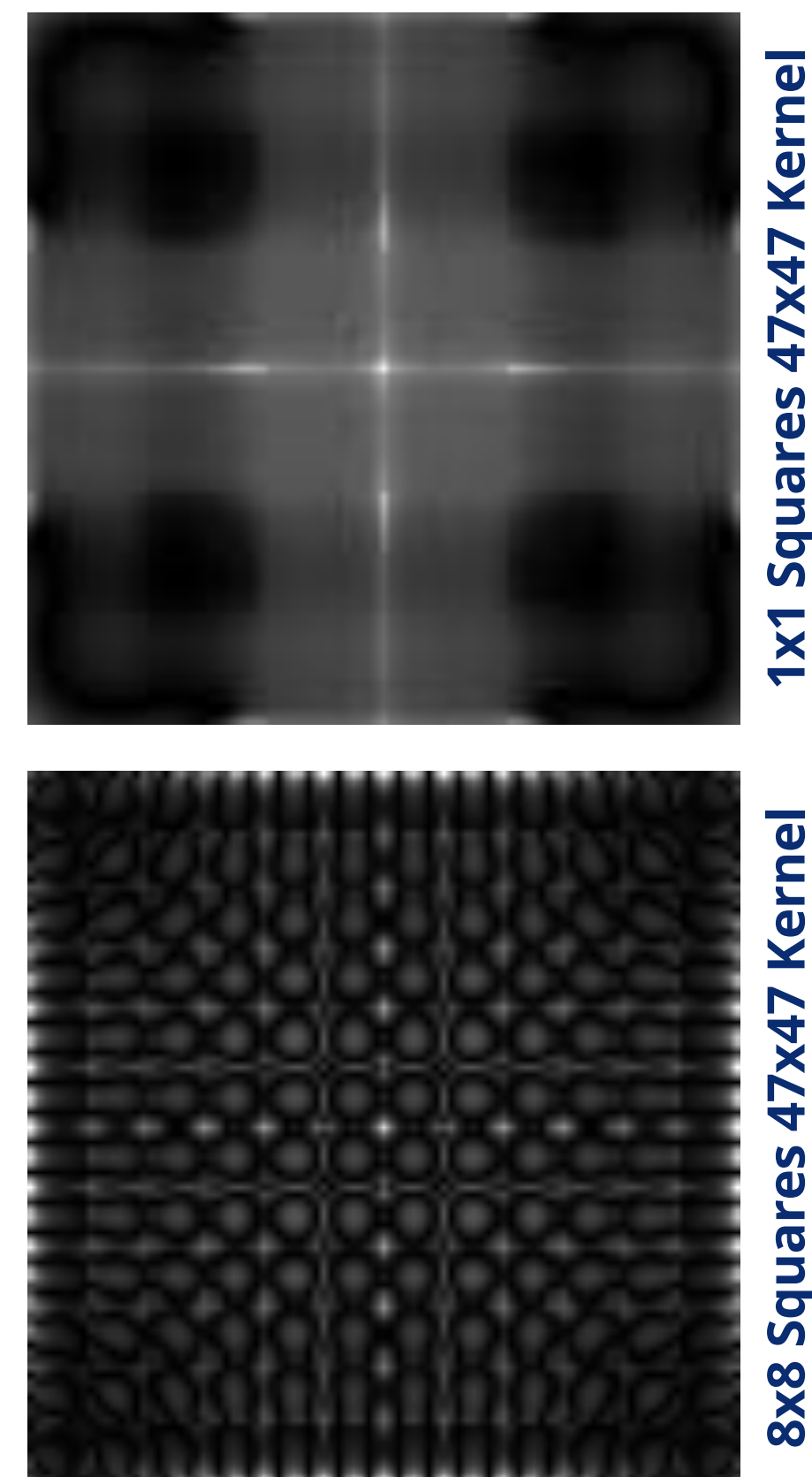
$$S_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = \mathcal{L}_{\text{rec}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)})$$

“Order triplets according to reconstruction loss / **visual distance**”

# Changing The Distance Function

Learn Distance Functions That Will Disentangle Certain Types Of Data

- **To disentangle** the adversarial dataset we need to **change** away from a pixel-wise **distance function** / loss
  - **Augment** the image **inside** the **distance function**.
- Optimise **spearman rank correlation** between factor distances and data visual distances
  - How well the latent ordering corresponds to the ground truth ordering

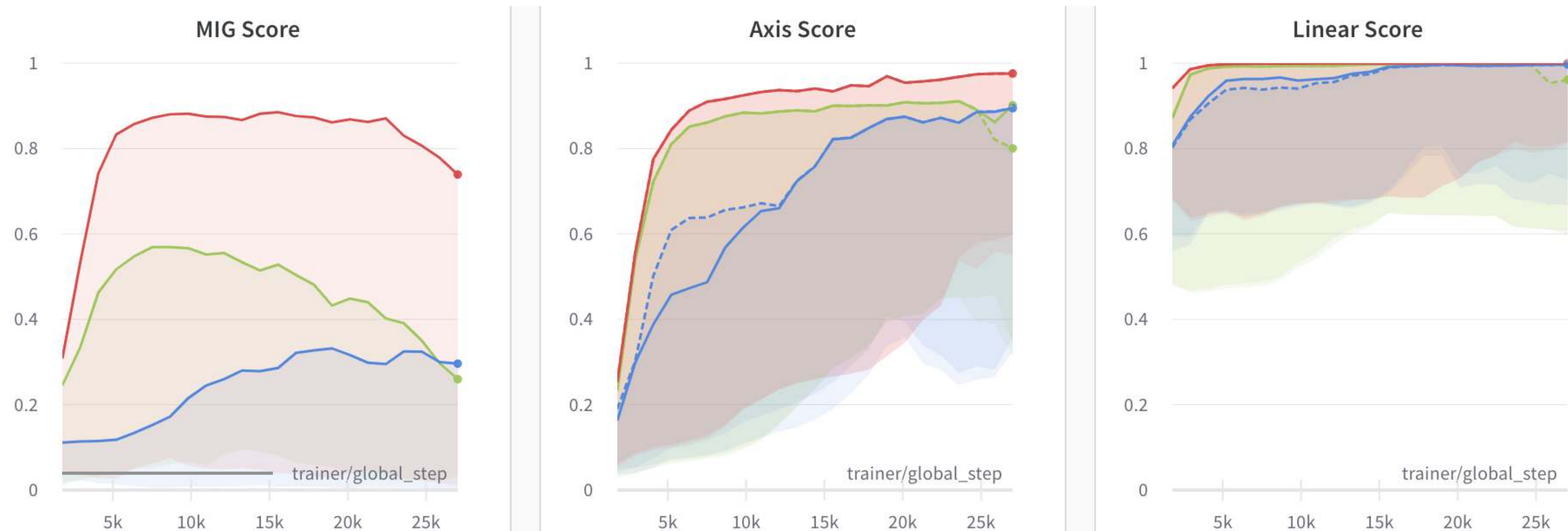


**Unsupervised Adaptive Triplet**  
1x1 Squares 47x47 Kernel

**\*NB\*** decoder just trained for visualisation

# Introducing Overlap

Results With Different Augmentations Before Distance Function



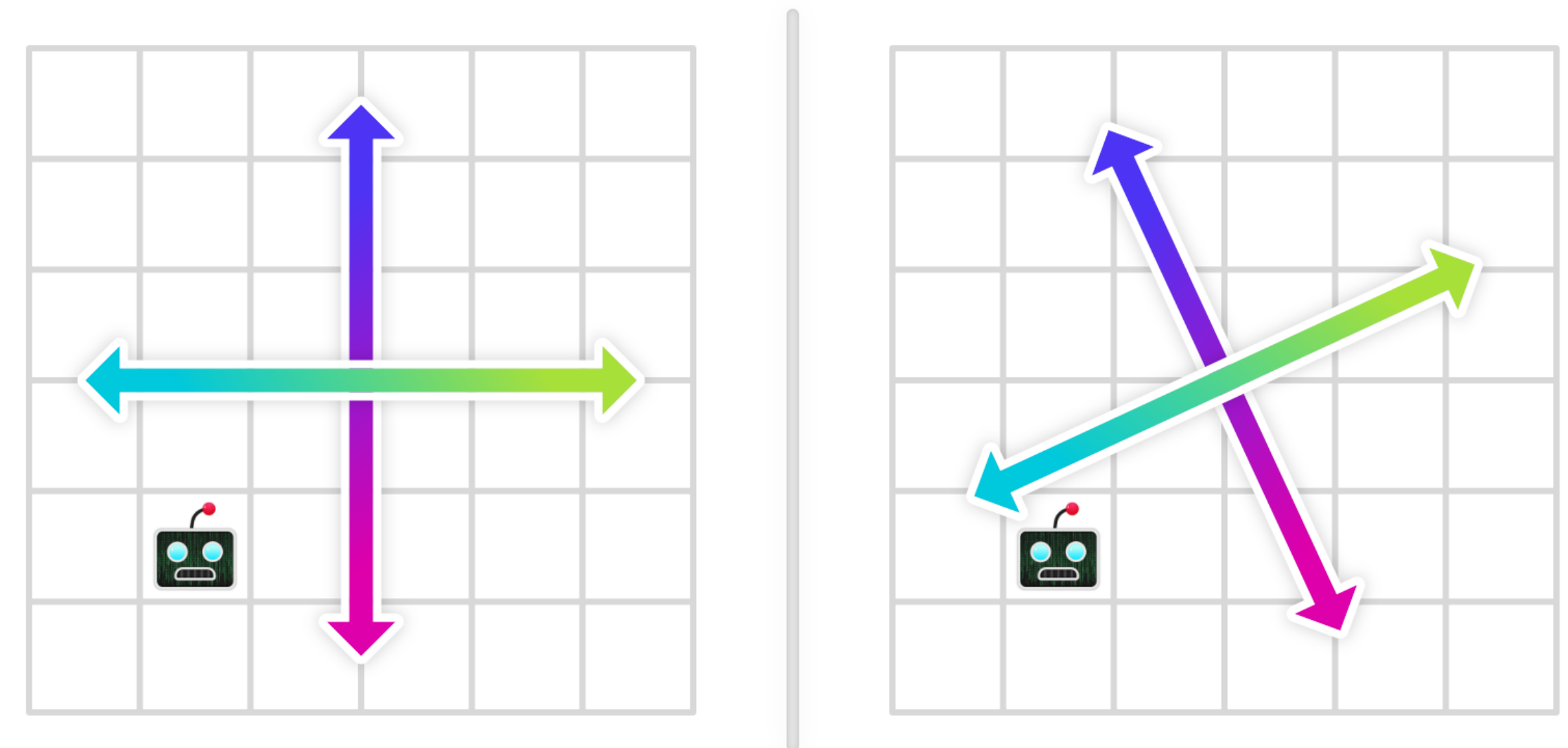
## **-7. Take Aways**

# Take Aways

How Can We Apply All Of This?

Disentanglement With VAEs is not always what it seems.

- You need to **introduce biases** to disentangle
- VAEs find **axes in data**, but these **may not correspond to ground truth factors**. This is by chance with synthetic datasets
- VAEs often **fail on real world data**

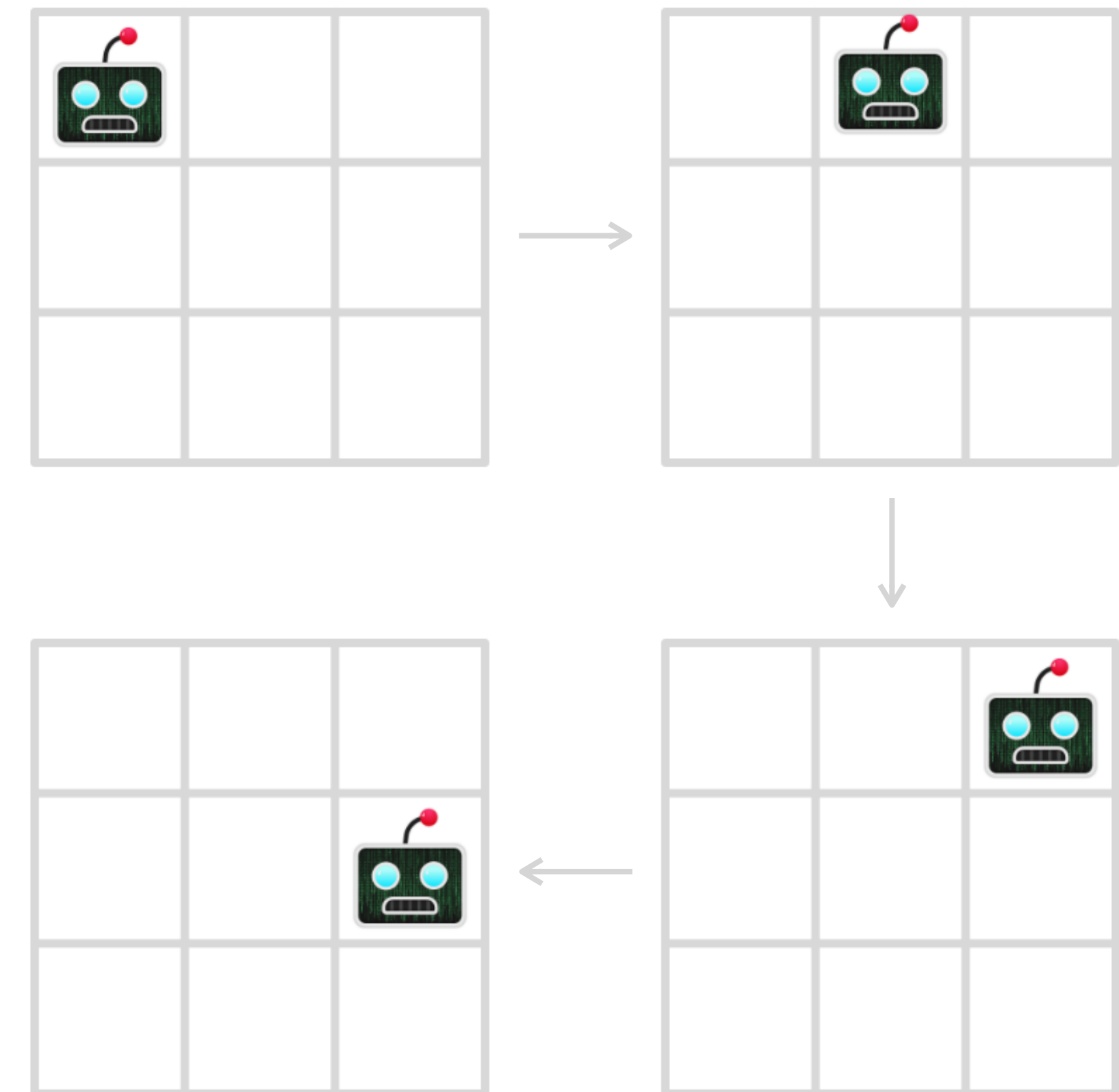


# Take Aways

How Can We Apply All Of This?

Be careful of the data you feed your models

- E.g. Reinforcement learning: consider **not skipping frames** if we need to learn a representation off of them. Missed **overlap may be important**.
- Can't expect to disentangle if the ground truth factors you want don't align with the distances in the data

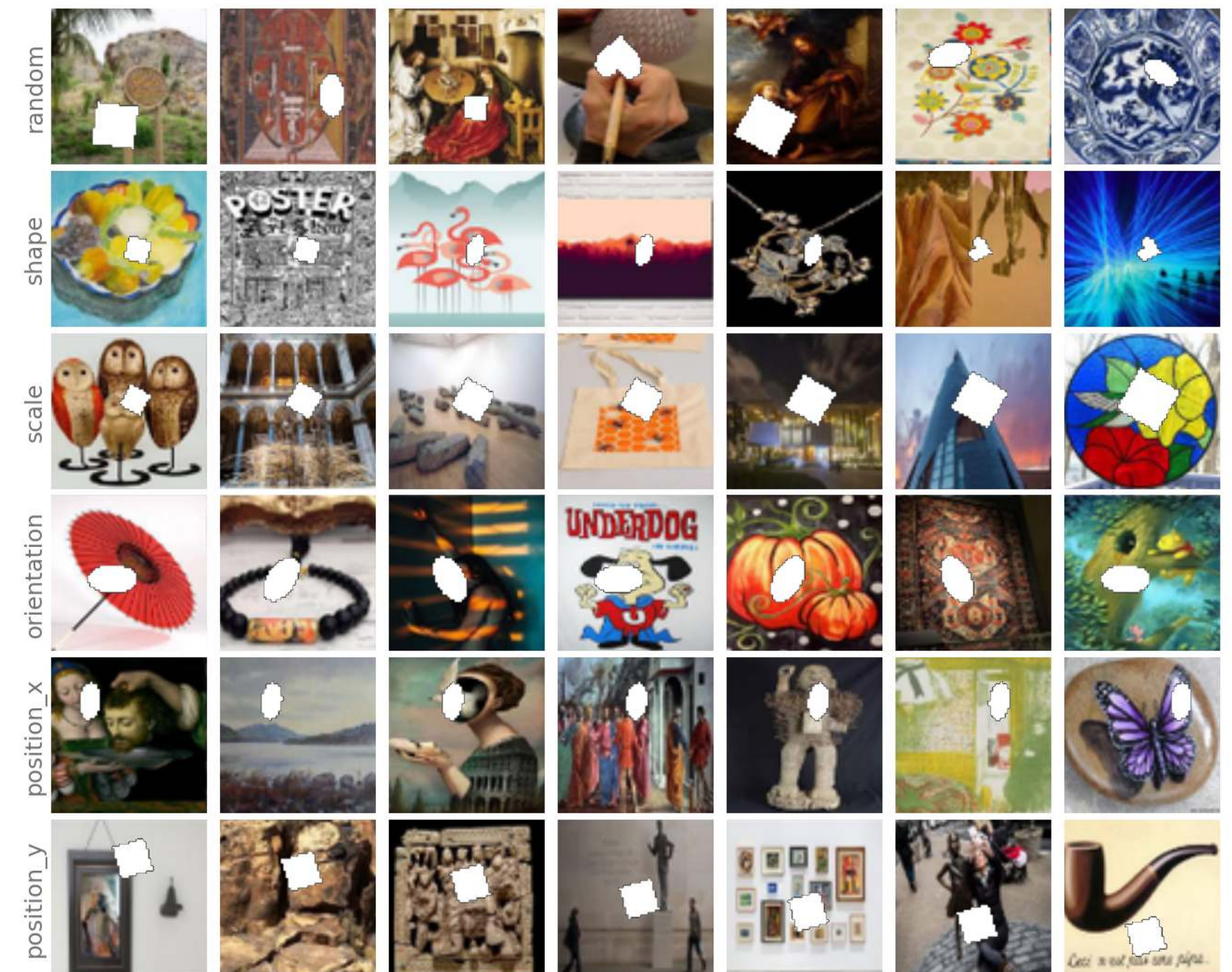


# Take Aways

# If You Have Extra Information Available, Use it!

## Use available information!

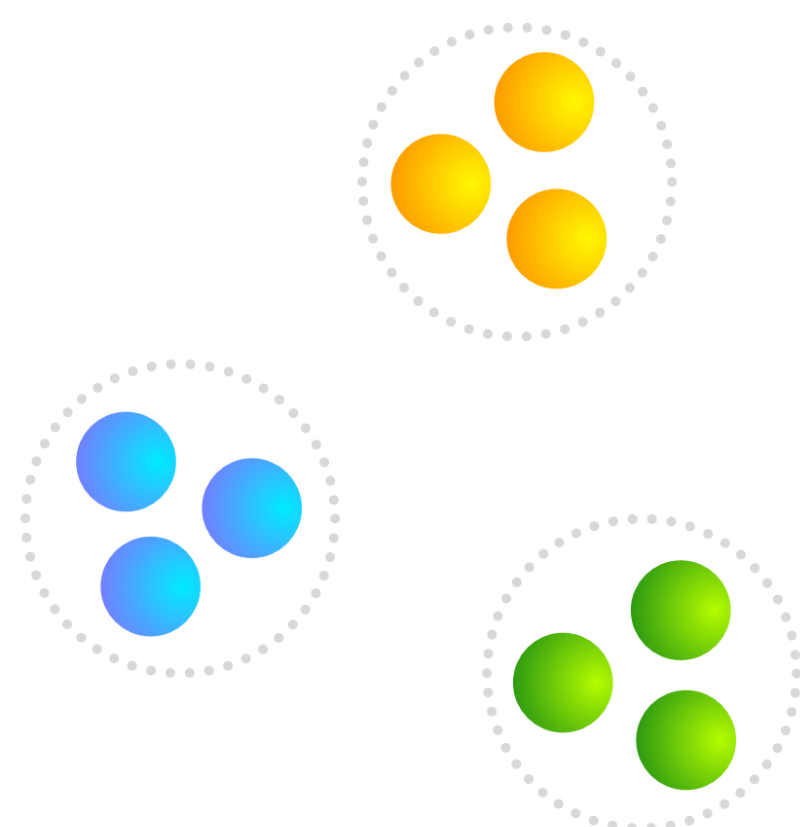
- If you have **sequences** of frames, **or location** information. You can use this to better learn representations
  - **Supervision**: E.g. Adaptive triplet
- VAEs might not be the best solution for your problem.
- Consider contrastive learning, eg. SimCLR



## **-8. More On Triplet**

# Supervised Triplet Choices

Triplet Intuition: Fine vs. Category Level Similarity



## Category-level Similarity

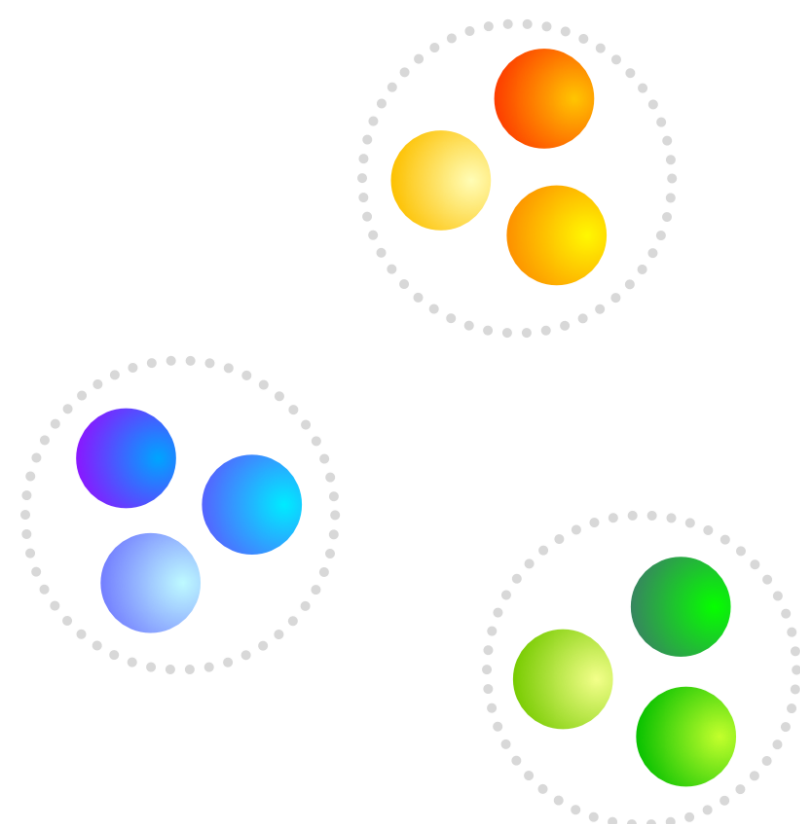
Distinguishes between categories

$$S_{\text{fdiff}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = k$$

$$S_{\text{fdiff}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = |\bar{\mathcal{S}}|$$

$$S_{\text{fdiff}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = \sum_{i=1}^m \left[ y_i^{(a)} \neq y_i^{(b)} \right]$$

eg. "Number of factors that differ"



## Fine-Grained Similarity

Distinguish between items within a category

$$S_{\text{fdist}}(\mathbf{x}^{(a)}, \mathbf{x}^{(b)}) = \|\mathbf{y}^{(a)} - \mathbf{y}^{(b)}\|_1$$

eg. "Manhattan distance along ground-truth factors"

Easy way to learn fine-grained similarity:

- Apply triplet loss to representations of an auto-encoder

## **-9. More On Adaptive VAEs**

# Adaptive VAEs

## Computing Deltas, Threshold & Shared Latent Distributions

### Deltas:

KL divergence between latents distributions

$$\delta_i = \tilde{D}_{\text{KL}} \left( q_{\phi}(z_i | \mathbf{x}^{(a)}), q_{\phi}(z_i | \mathbf{x}^{(b)}) \right)$$

### Threshold:

Half the min and the max delta

$$\tau = \frac{1}{2} \left( \min_{i \in [D]} \delta_i + \max_{i \in [D]} \delta_i \right)$$

### Shared Indices Set:

Indices of deltas less than the threshold

$$S = \{i \mid \delta_i < \tau, \forall i \in [D]\}$$

Compute the averaged posteriors

$$\tilde{q}_{\phi}(z_i | \mathbf{x}^{(a)}) = \begin{cases} \text{a} \left( q_{\phi}(z_i | \mathbf{x}^{(a)}), q_{\phi}(z_i | \mathbf{x}^{(b)}) \right) & \text{if shared: } i \in S \\ q_{\phi}(z_i | \mathbf{x}^{(a)}) & \text{if differing: } i \in \bar{S} \end{cases}$$

# Adaptive VAEs

## Algorithm Overview - Loss

$$L_{\text{AdaVAE}} = \frac{1}{2} L_{\beta\text{VAE}}^{(A)} + \frac{1}{2} L_{\beta\text{VAE}}^{(B)}$$

Adaptive VAE loss is just the average of the  $\beta$ -VAE loss for both observations in the pair.

The **averaged** latent distributions **are used** for the regularisation terms and sampling.

# Adaptive VAEs

## Differences From Original Paper

- Original paper uses **asymmetric deltas** to compute the threshold and average
  - Instead use **symmetric** deltas which are the average of the forward and backward KL divergences between corresponding latent distributions:

$$\delta_i = \tilde{D}_{\text{KL}}(q_{A_i}, q_{B_i})$$

Symmetric KL Divergence:

$$\tilde{D}_{\text{KL}}(p, q) = \frac{1}{2}D_{\text{KL}}(p \parallel q) + \frac{1}{2}D_{\text{KL}}(q \parallel p)$$

- The name used here does not correspond to the original.
  - **Ada-GVAE** and **Ada-ML-VAE** indicate the choice of **averaging function** for the latent distributions in the original work.
  - Frameworks are **exactly the same otherwise**.
  - The more simple **Ada-GVAE** is (usually) **better**.

$$\mathcal{N}\left(\frac{1}{2}(\mu_i^{(a)} + \mu_i^{(b)}), \frac{1}{2}(\sigma_i^{2(a)} + \sigma_i^{2(b)})\right)$$

# **-10. Problems With VAEs**

# VAEs In Practice - Problems With Posterior

Or Rather The Multivariate Normal Distribution With Diagonal Covariance

- Normal distribution is **rotationally invariant**
  - Breaking this property is important for disentanglement
- Often  $\sigma$  is **discarded** and only  $\mu$  kept for inference.
  - More advanced frameworks like the Ada-GVAE **may learn incorrect distances** due to its involvement in the KL divergence

- KL Divergence between posterior and prior is **effectively just l2 regularisation**

$$\mathcal{L}_{\text{reg}} = \text{KL}(\mathcal{N}(\mu, \sigma) || \mathcal{N}(1, 0))$$

$$= \frac{1}{2}(\sigma^2 + \mu^2 - 1 - 2\ln(\sigma))$$

if  $\sigma \approx 1$  then

$$\approx \frac{1}{2}\mu^2$$

**\* We recognise but ignore these issues**

# Regularisation

## Structure Enforced By The Prior <sup>[1]</sup>

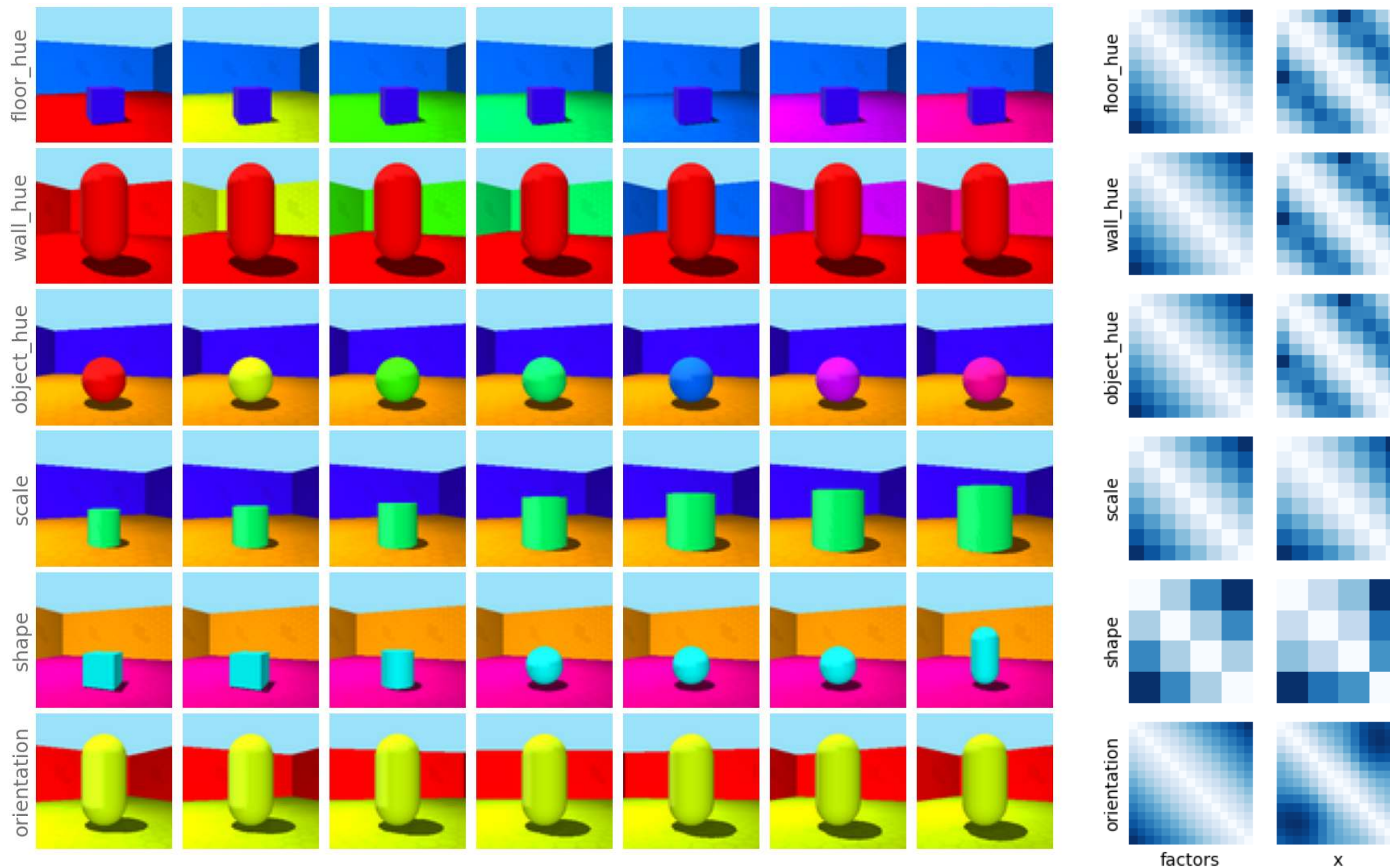
- Most common prior is the multivariate normal distribution with diagonal covariance.
  - Usually **justified by** encouraging **independence** of latent units.
  - **Rotational invariance**, does not encourage axis alignment any more than it encourages any arbitrary rotation of these factors.
- **Better priors** can enforce better structure and dependencies on latent space.
  - **Breaking rotational invariance**, even with simple manipulations, **is important**

[1] Mathieu, Emile, Tom Rainforth, N. Siddharth, and Yee Whye Teh. 2018. "Disentangling Disentanglement In Variational Autoencoders". *Arxiv.Org*. <https://arxiv.org/abs/1812.02833>.

# **-11. Problems With Existing Data**

# Existing Ground-Truth Datasets

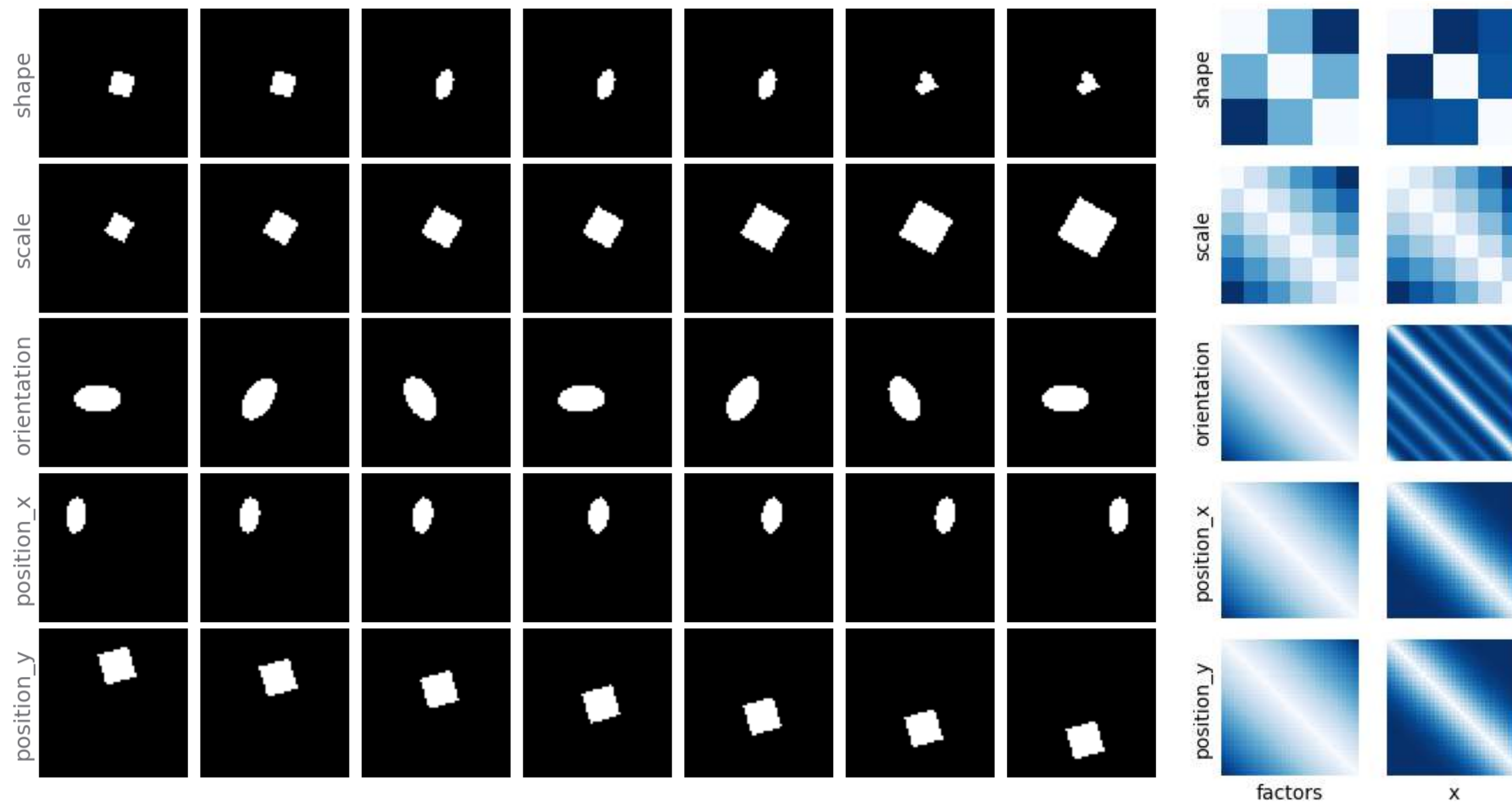
## Shapes3D



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

# Existing Ground-Truth Datasets

## DSprites



- Ground-truth factors with only 3 or less items are always disentangled

# Existing Ground-Truth Datasets

## Typical Datasets

elevation



azimuth



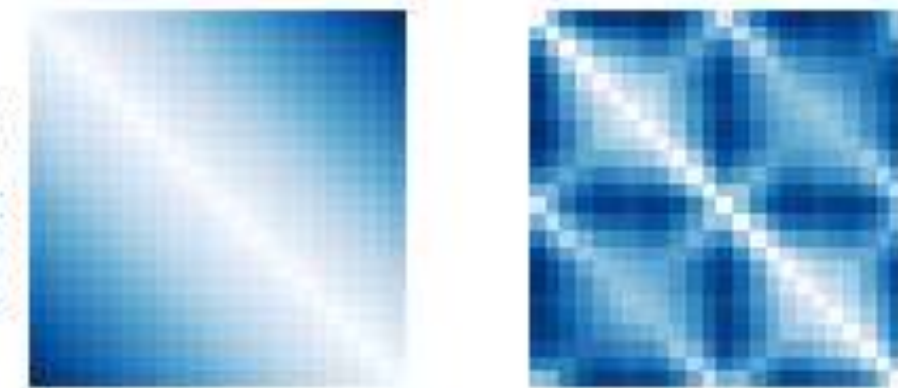
object\_type



elevation



azimuth



object\_type



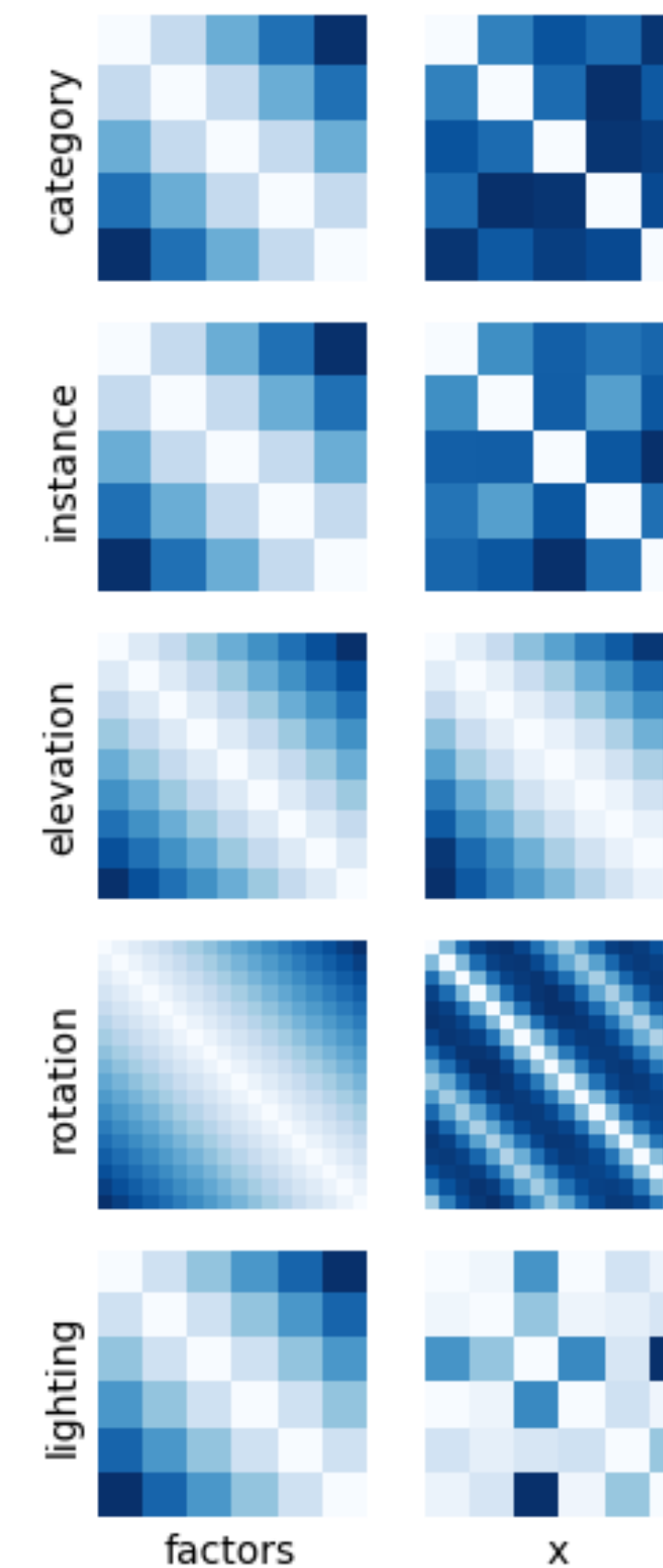
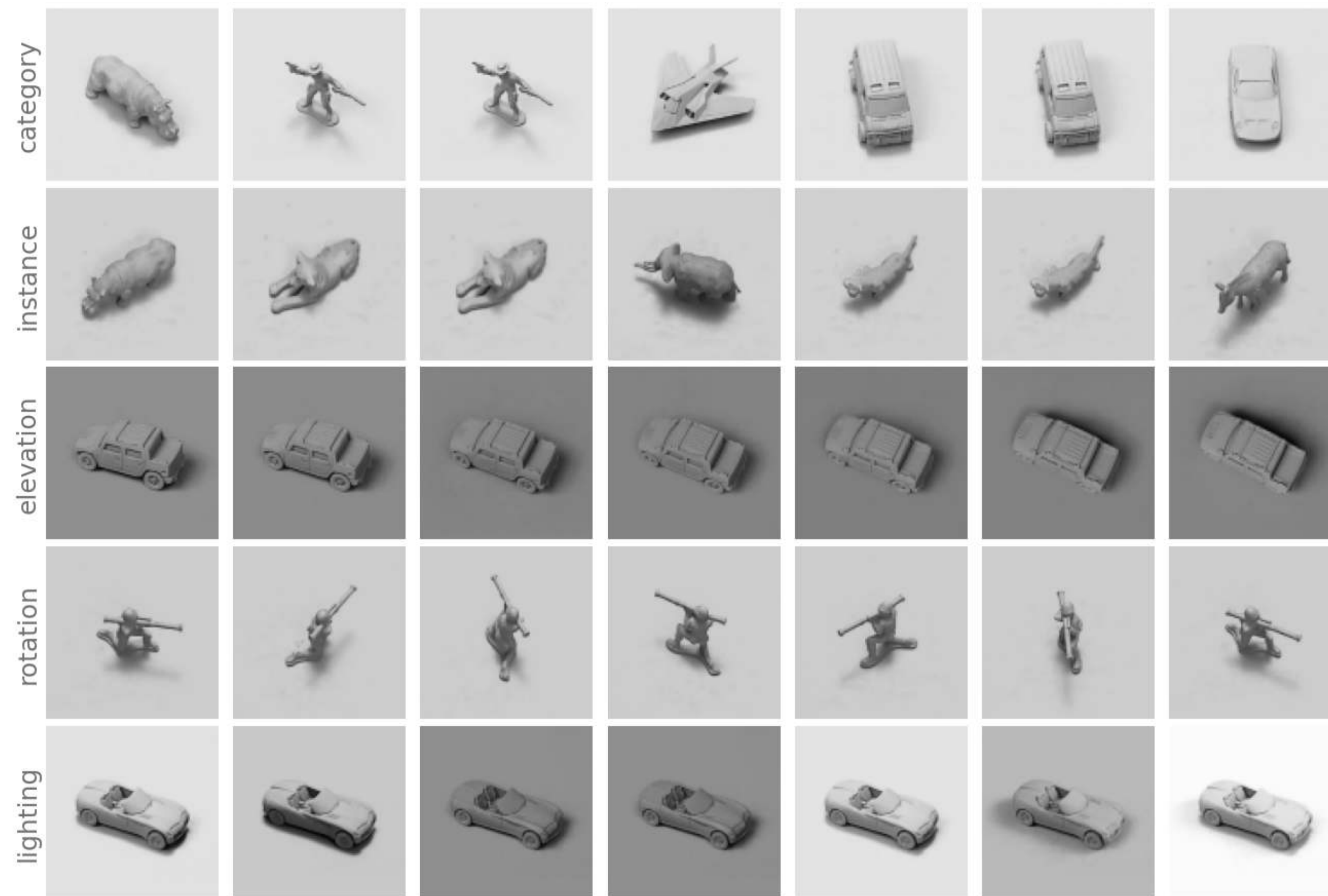
factors

x

- Factors (azimuth & object\_type) are easier to decompose into sub-factors

# Existing Ground-Truth Datasets

## Typical Datasets

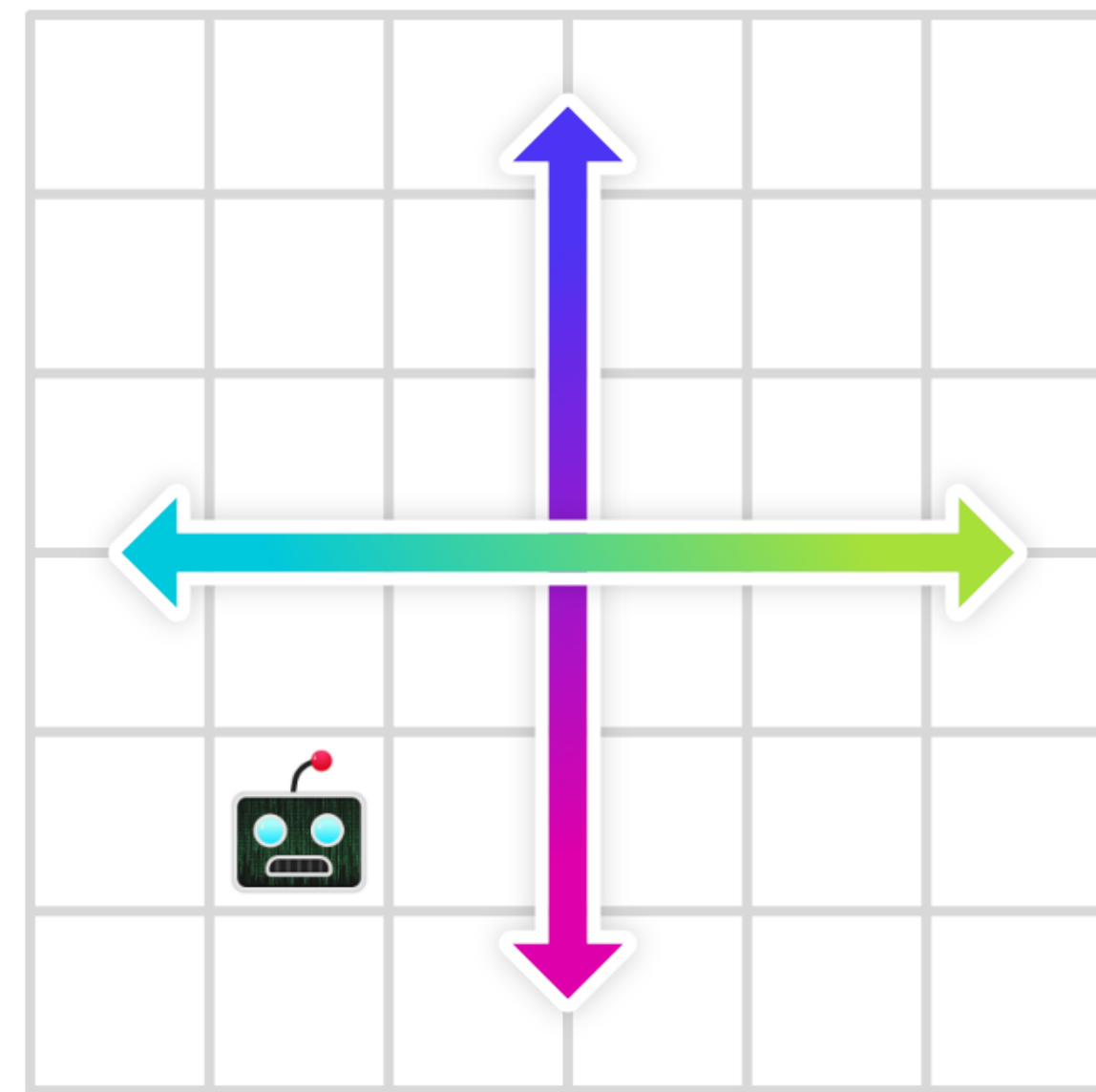
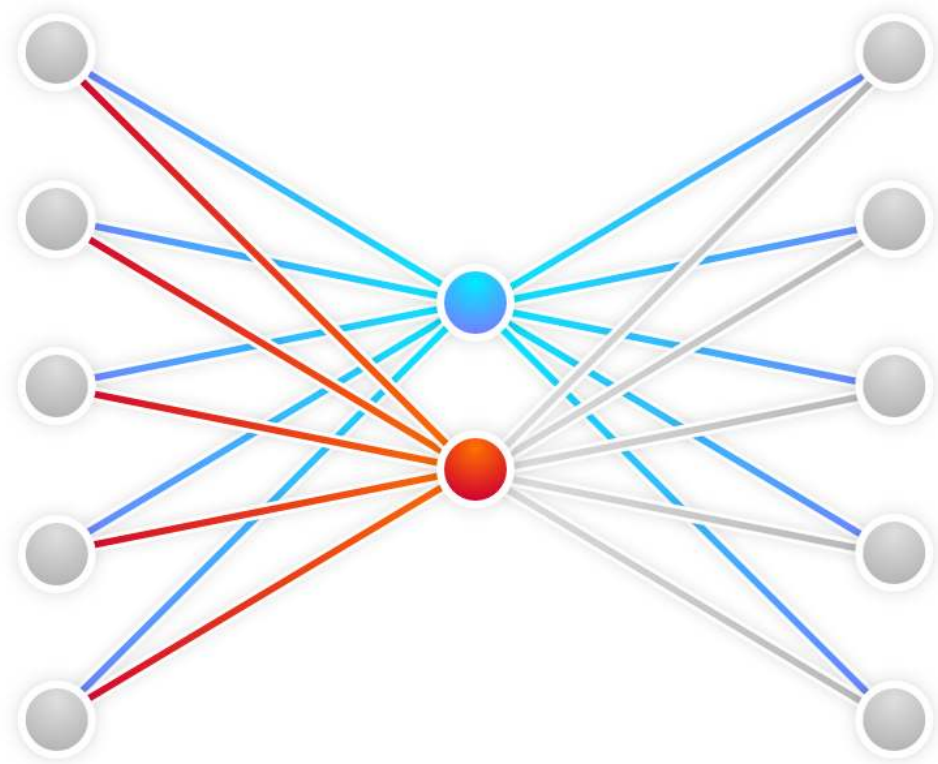


- Factor (lighting) is easier to decompose into sub-factors
- Other factors do not correspond to ground truth

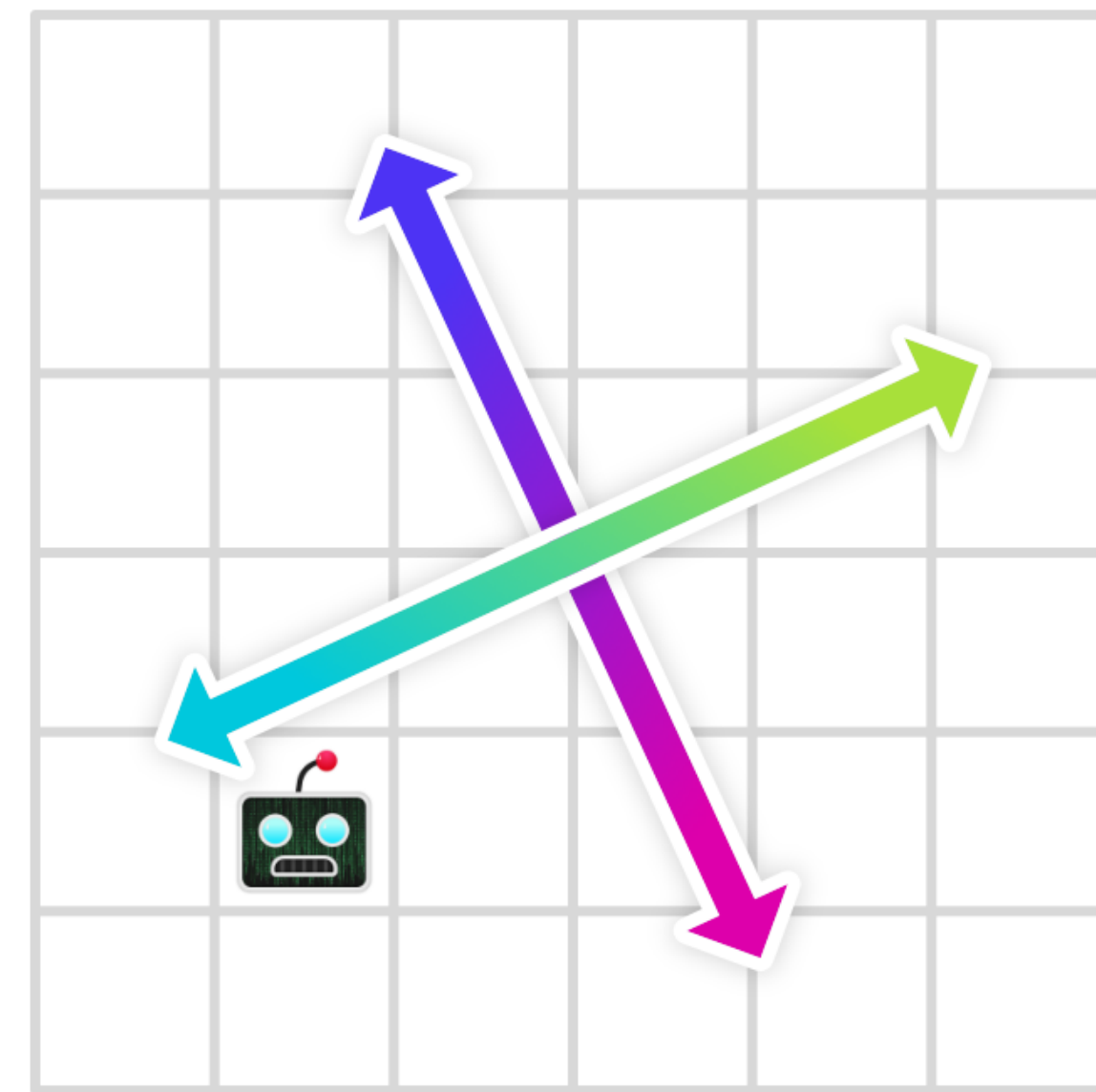
# **-12. Measuring Axis-Alignment**

# Axis Alignment Metrics

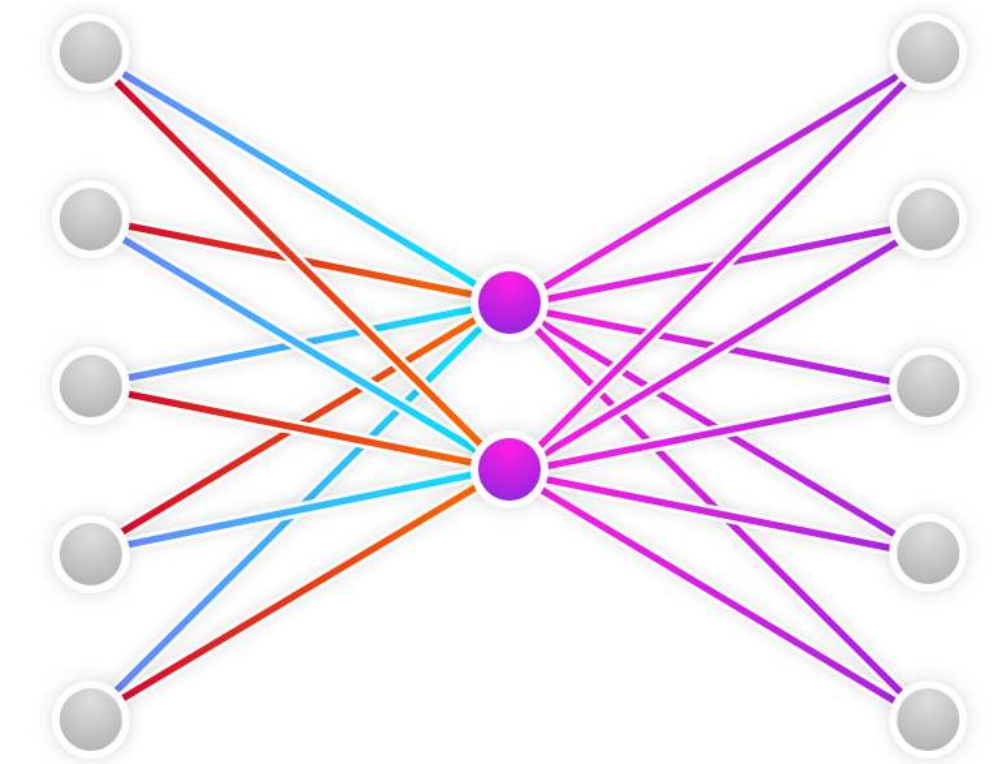
Disentanglement Scores



Disentangled  
Axis Aligned



Entangled  
Arbitrary Axis



# Axis Alignment Metrics

## The Nitty-Gritty

$$\mathcal{F}_{\text{linear},i}(\mathcal{Z}_i) = \max_{u \in [D]} \left( \frac{\mathbb{E} [\sqrt{\lambda_u}] }{\sum_{v=1}^D \mathbb{E} [\sqrt{\lambda_v}]} \right)$$
$$\mathcal{F}_{\text{linear}} = \left( \prod_{i=1}^F \mathcal{F}_{\text{linear},i} \right)^{\frac{1}{F}}$$

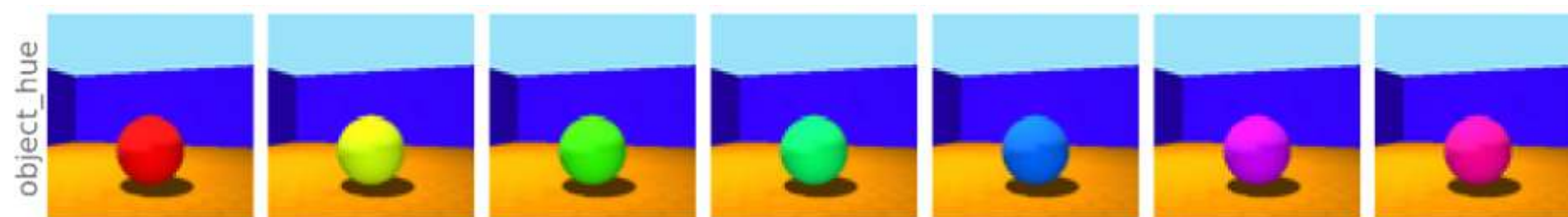
**Linearity Score:** Deviation from the best principle component

- Largest **singular value** over sum

$$\mathcal{F}_{\text{axis},i}(\mathcal{Z}_i) = \max_{u \in [D]} \left( \frac{\mathbb{E} [\text{Std}[z_u]] }{\sum_{v=1}^D \mathbb{E} [\text{Std}[z_v]]} \right)$$
$$\mathcal{F}_{\text{axis}} = \left( \prod_{i=1}^F \mathcal{F}_{\text{axis},i} \right)^{\frac{1}{F}}$$

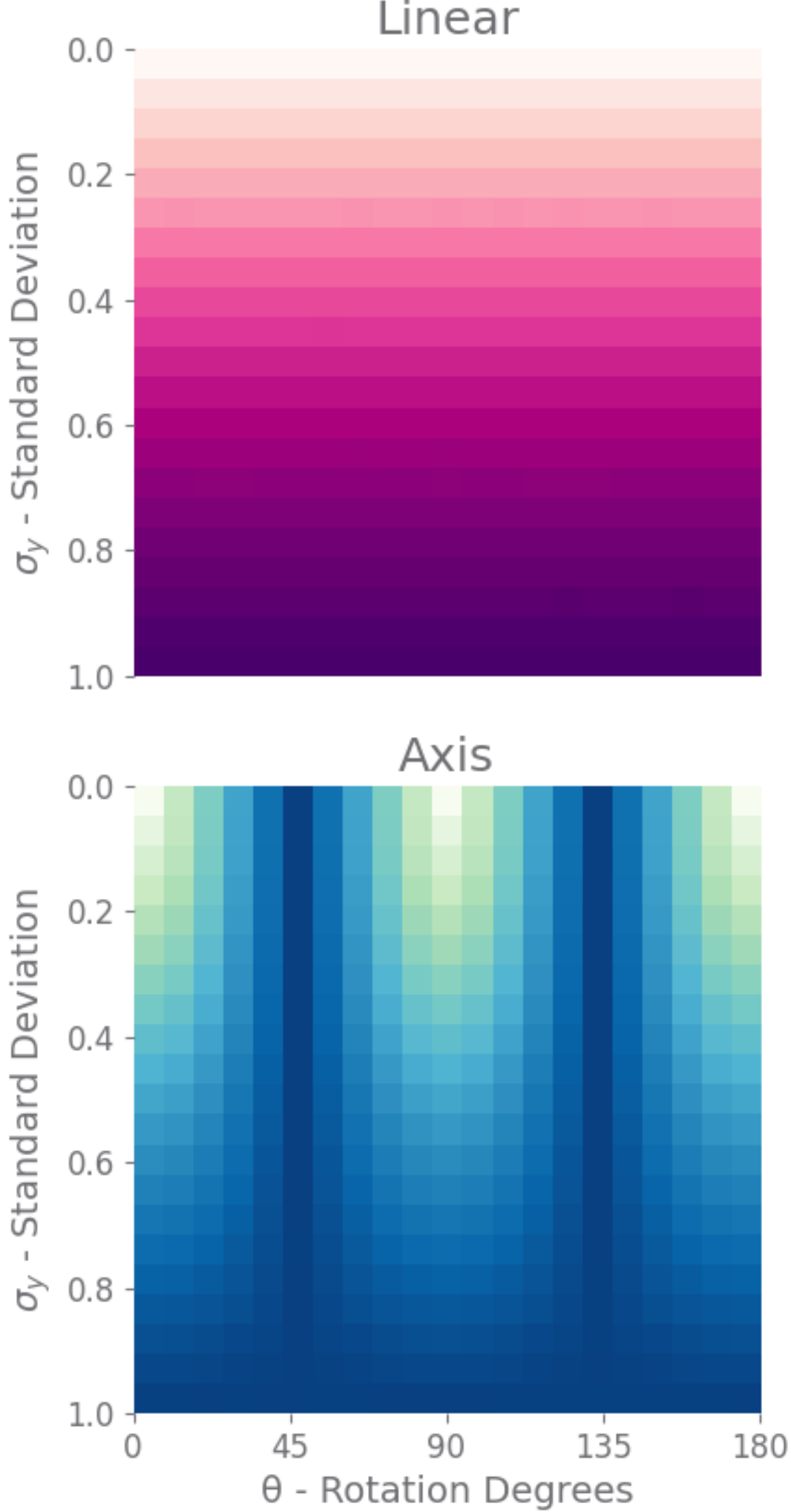
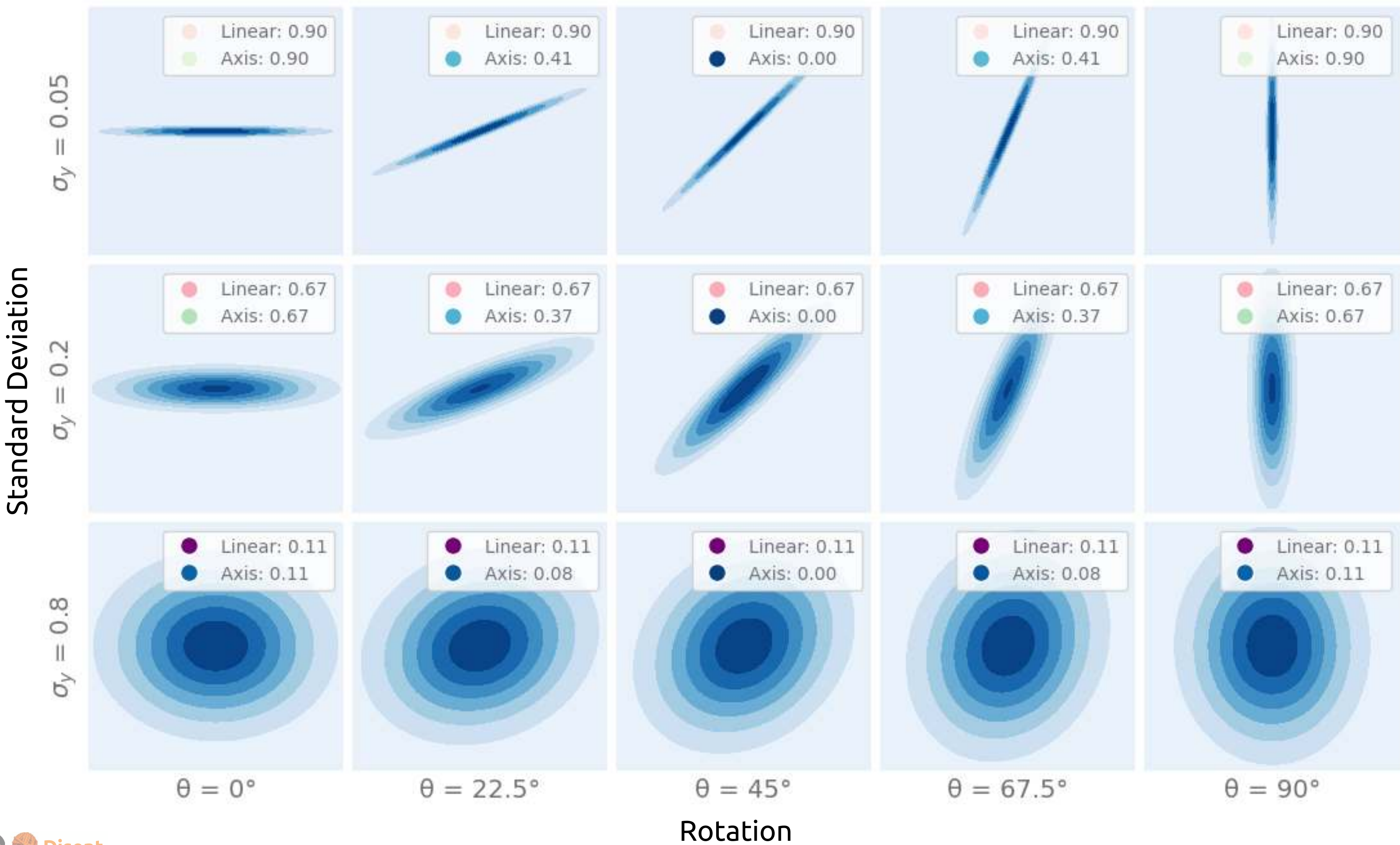
**Axis Score:** Deviation from the best standard basis / axis

- Largest **standard deviation** over sum



Score is **computed for each** ground-truth **factor** using random traversals. Then the geometric mean is taken.

# Axis Alignment Metrics



# **-13. Generating Adversarial Data**

# Modified Adversarial Dataset

How Can We Generate Adversarial Data?

## 1. **Modify** The Dataset

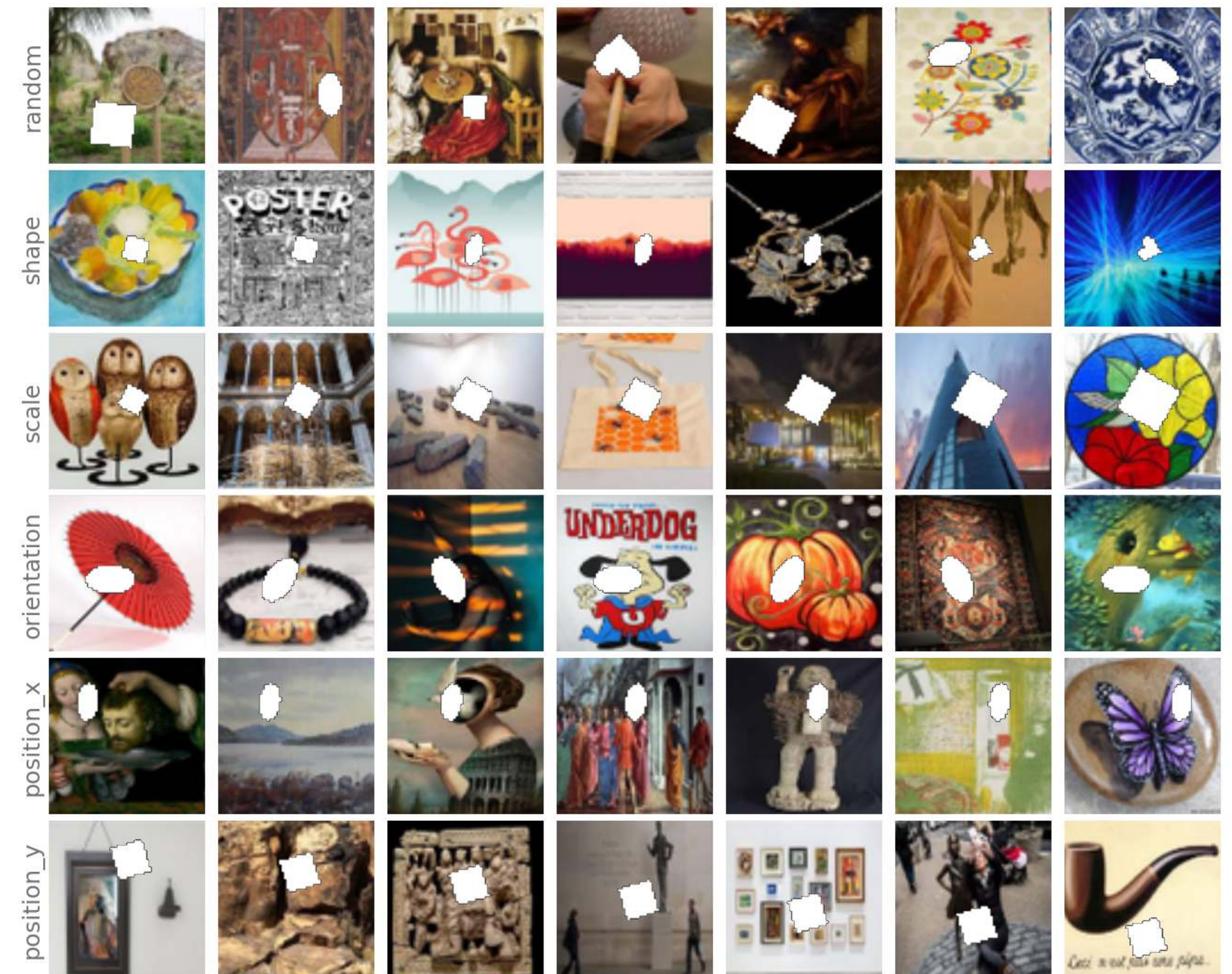
- Directly optimise the dataset for certain properties?

## 2. **Remove Elements** From Dataset

- Use a genetic algorithm to create a mask over dataset elements?

## 3. Use **Real World** Data<sup>[1]</sup>

- Incorporate real world data into the disentanglement benchmarks?
- I still need to test this!



DSprites - ImageNet

[1] I need to go on a hunt for this citation! But it makes sense based on our explanations so far!

# Modified Adversarial Dataset

## Triplet Loss To The Rescue

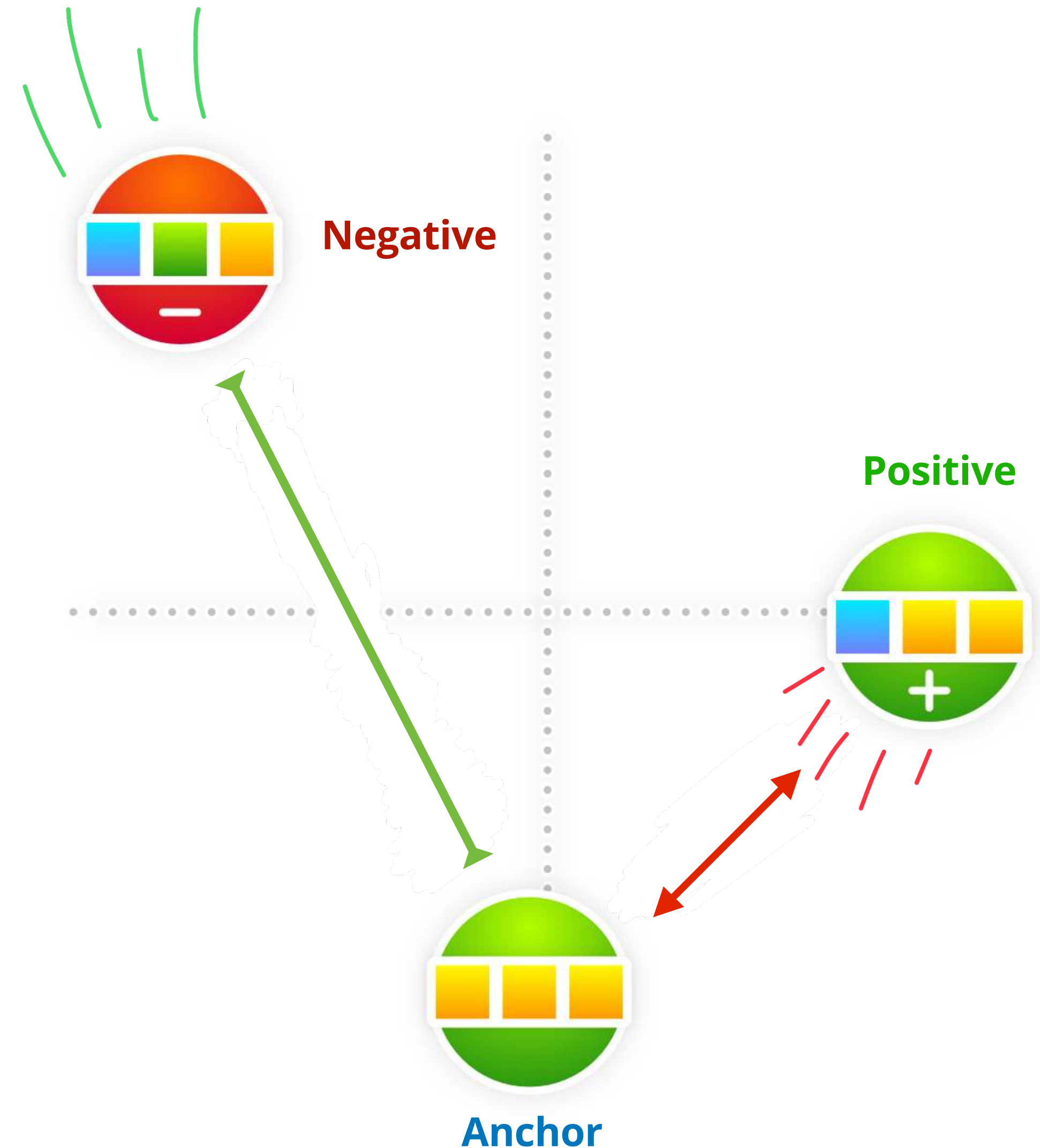
### Goal

- **Invert overlap** between observations according to ground-truth factors. Push close factors apart and pull far factors together

### Inverted Triplet Loss

- **Pull negative closer**
- **Push positive further away**
- **Positive has no effect beyond margin**

$$\mathcal{L}_{\text{invert}} = \max(0, \underbrace{\psi}_{\text{blue}} + \underbrace{\mathcal{L}_{\text{rec}}(a, n)}_{\text{red}} - \underbrace{\mathcal{L}_{\text{rec}}(a, p)}_{\text{green}})$$

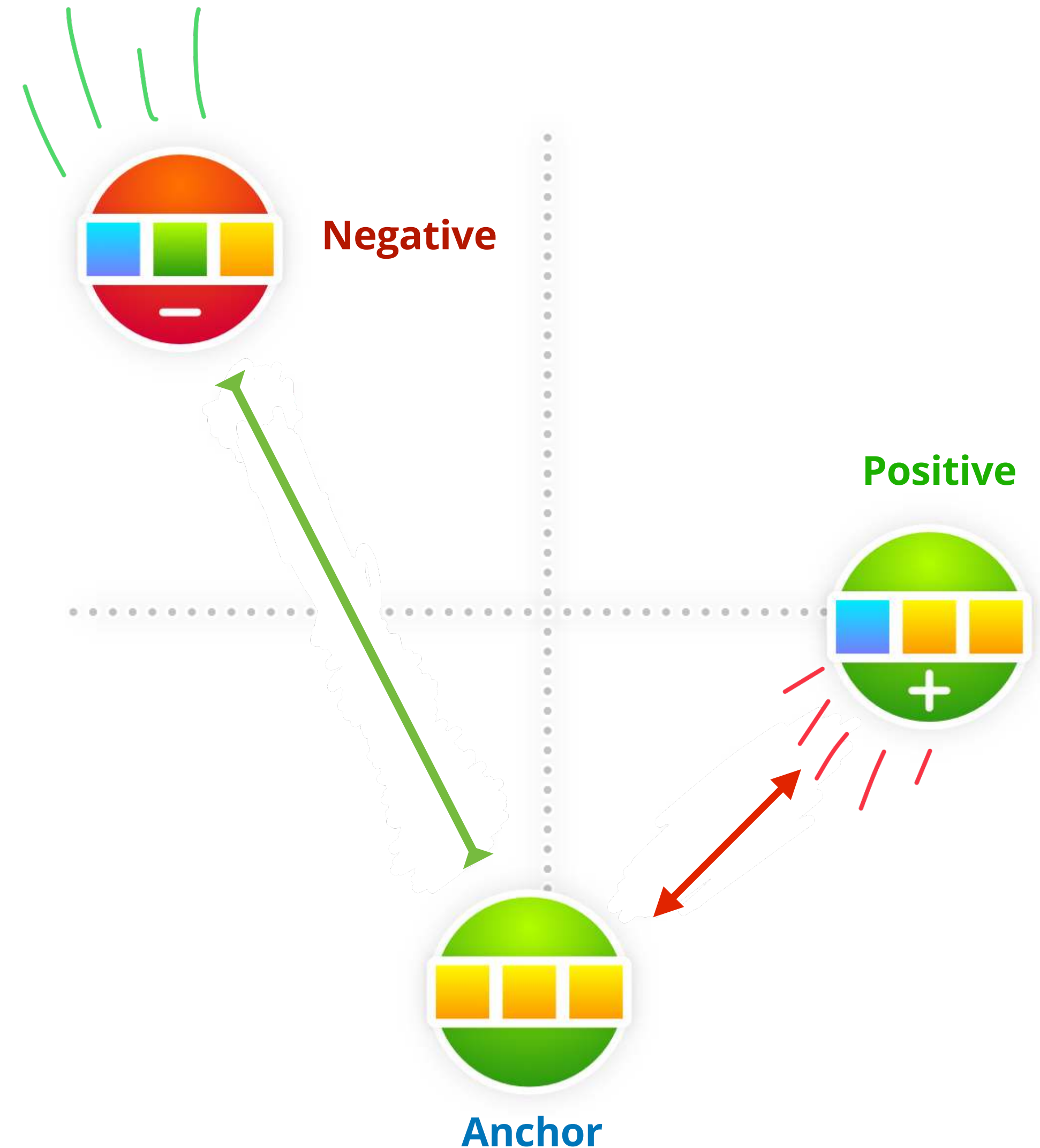


# Modified Adversarial Dataset

Applying Triplet Loss

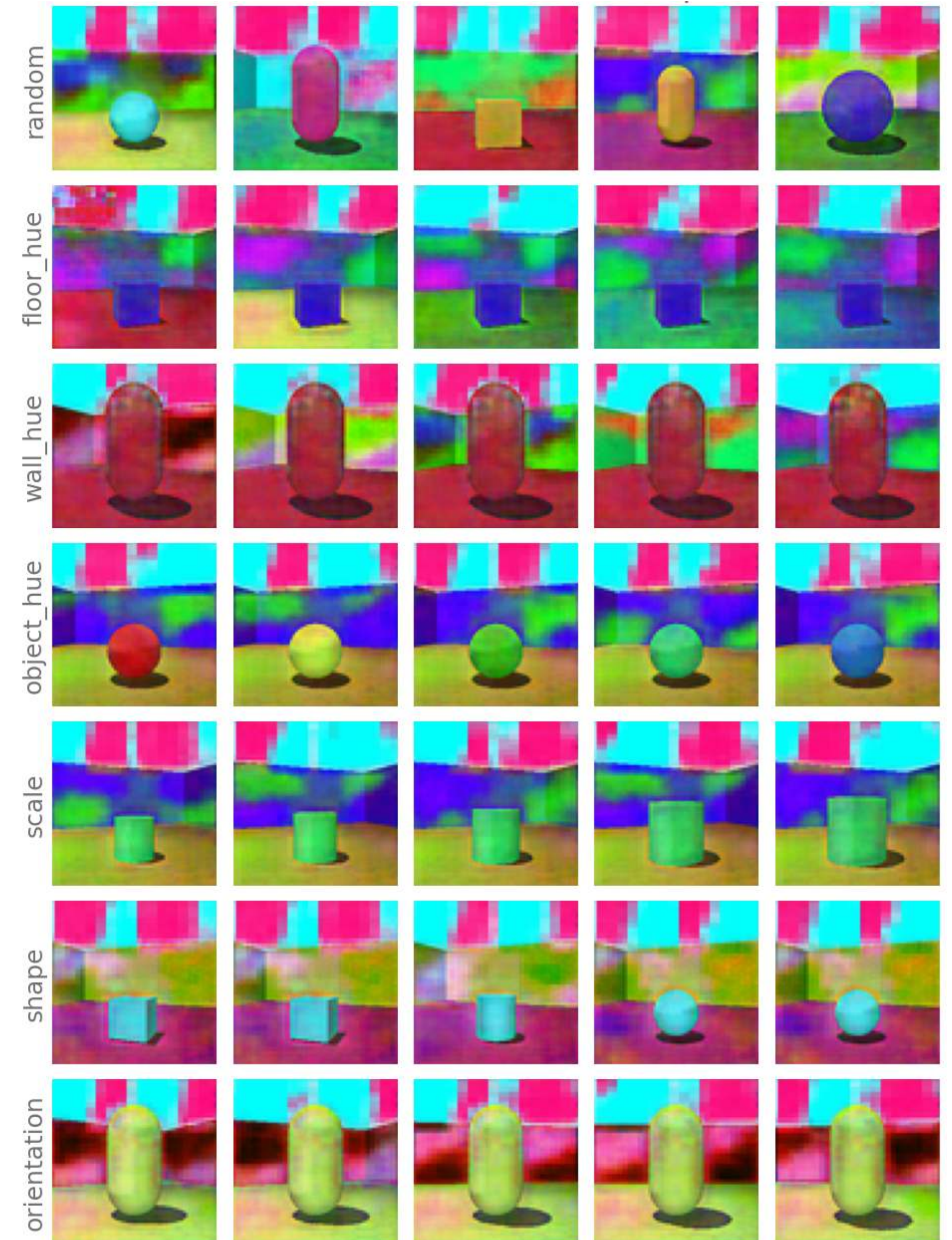
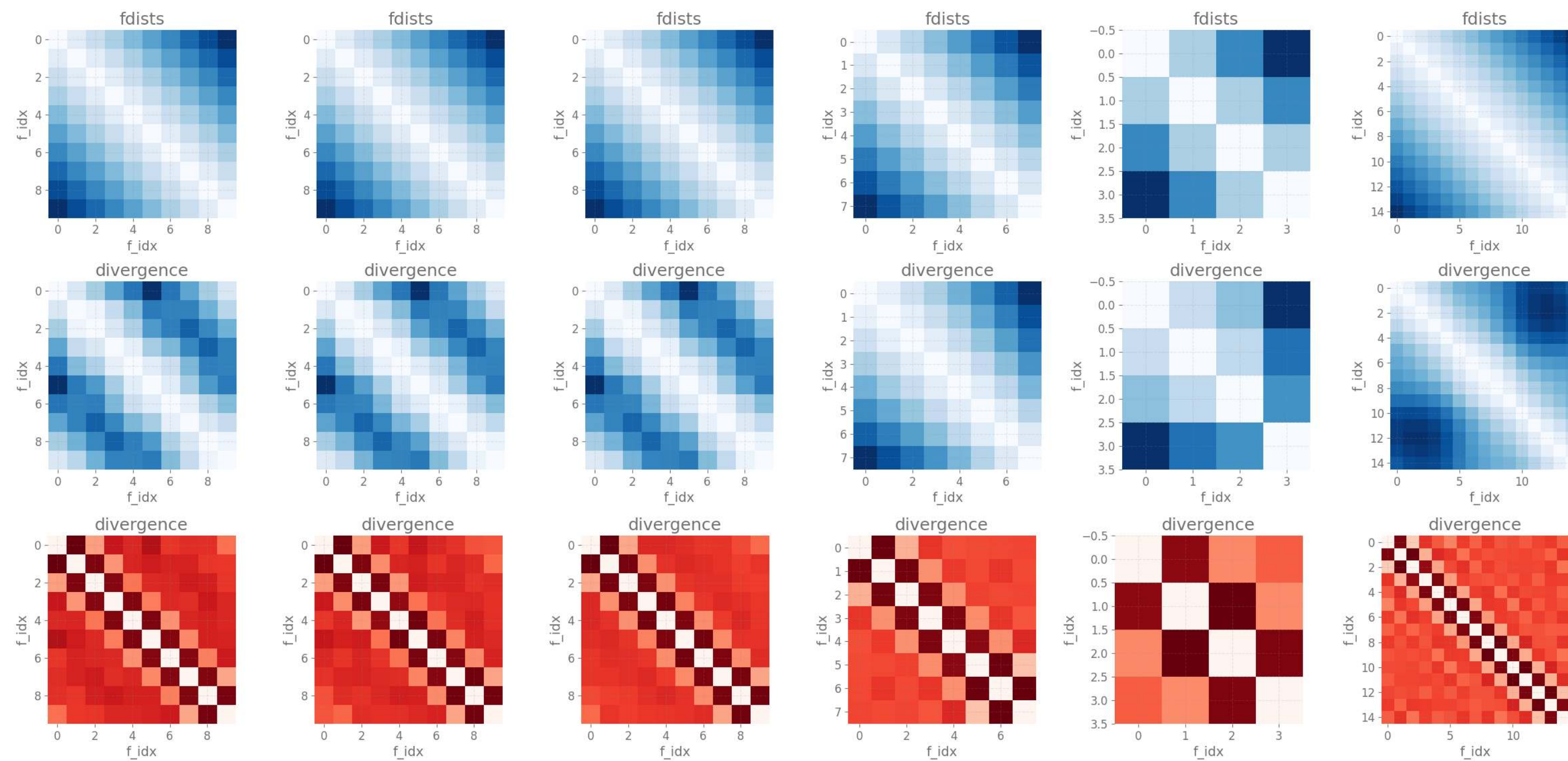
## Approximations

- Often **Infeasible** to load entire dataset into memory and apply optimiser to it in its entirety
- Approximate adversarial dataset with **neural network** that **encodes differences**  
 $\hat{x} = x + \delta(x)$
- **Regularise** differences **towards zero**
- Apply triplet over adversarial items



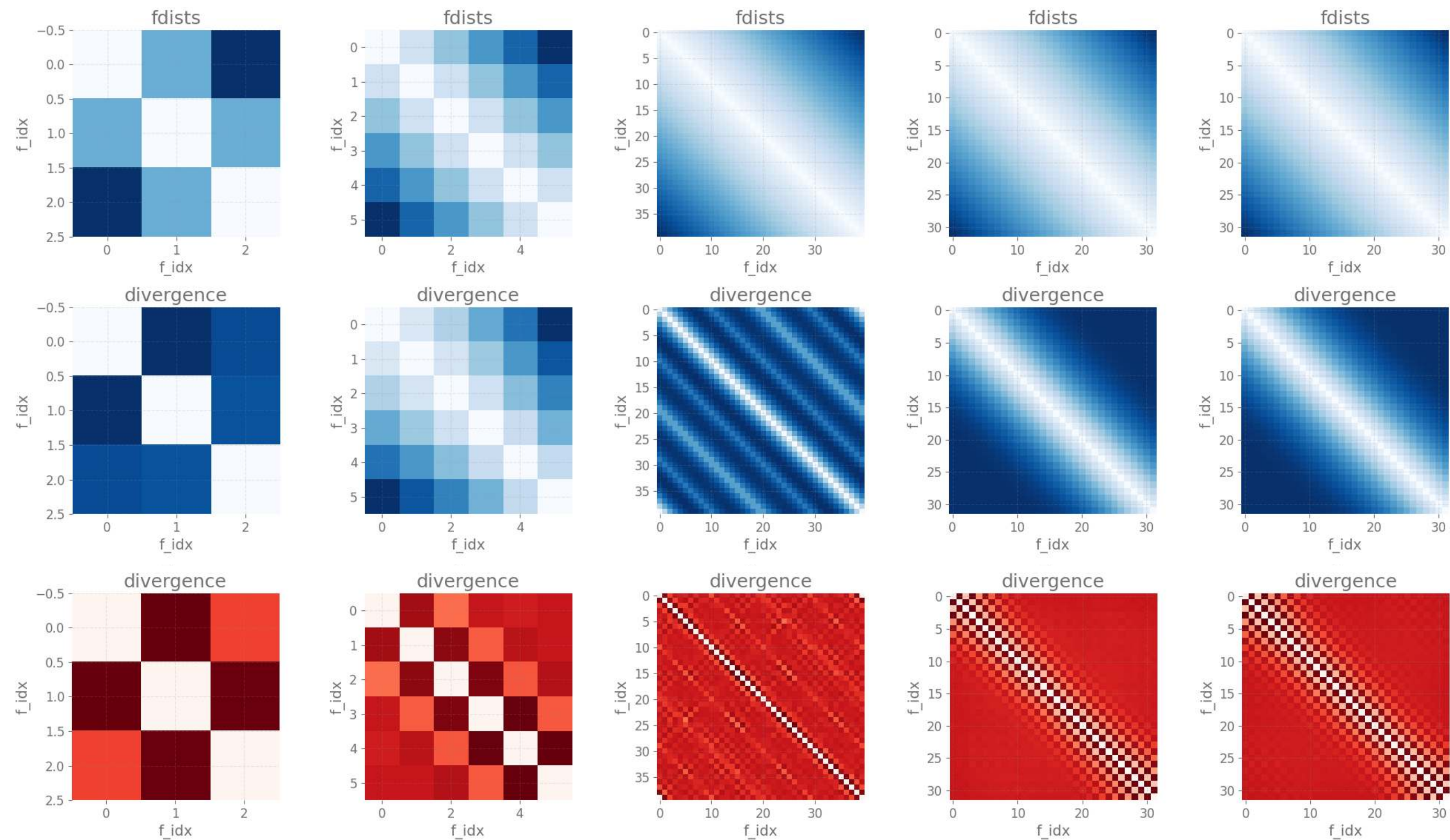
# Adversarially Modified

## Overview - Shapes3D



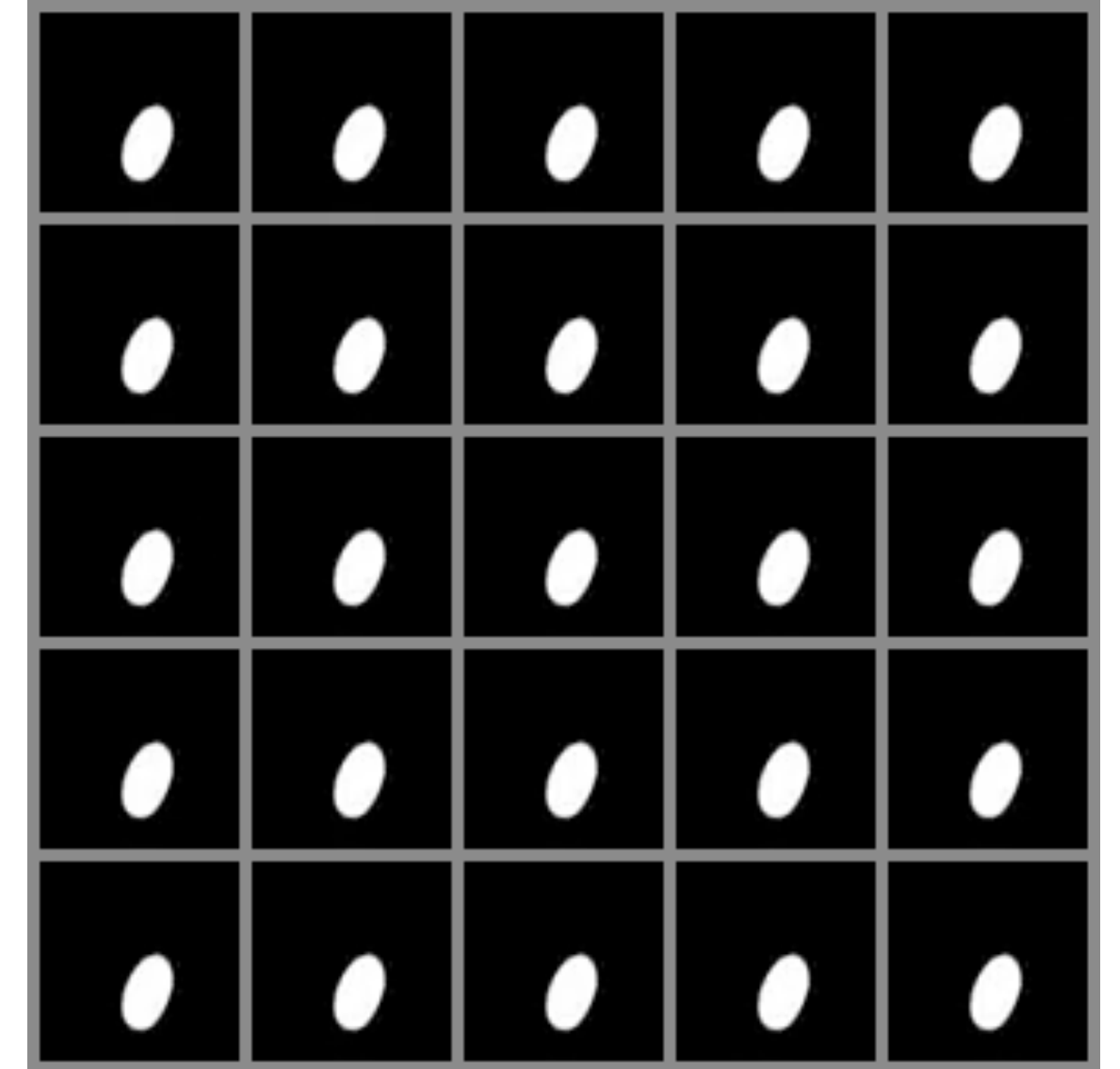
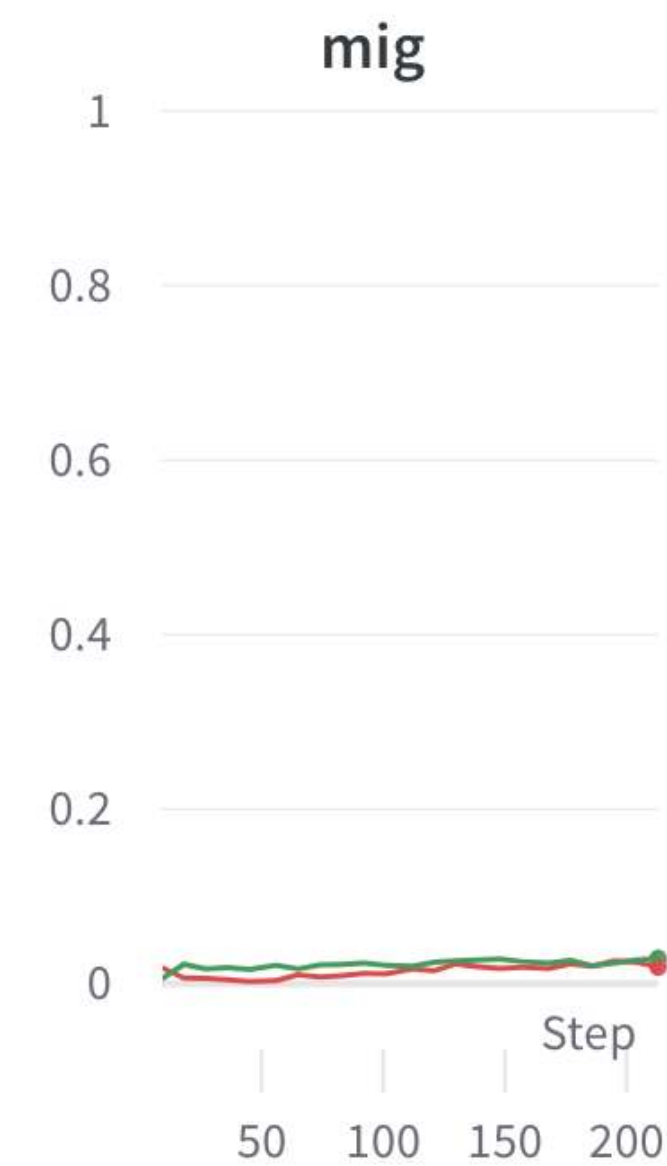
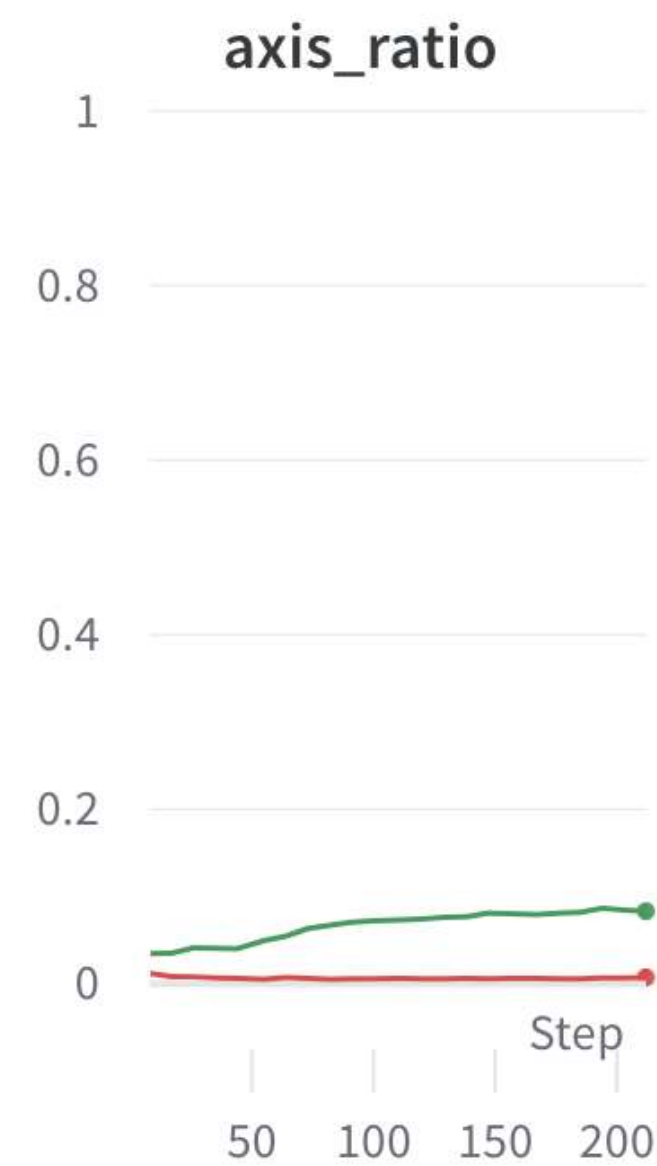
# Adversarially Modified

## Overview - DSprites



# Adversarially Modified

Results - DSprites



# Masked Adversarial Dataset

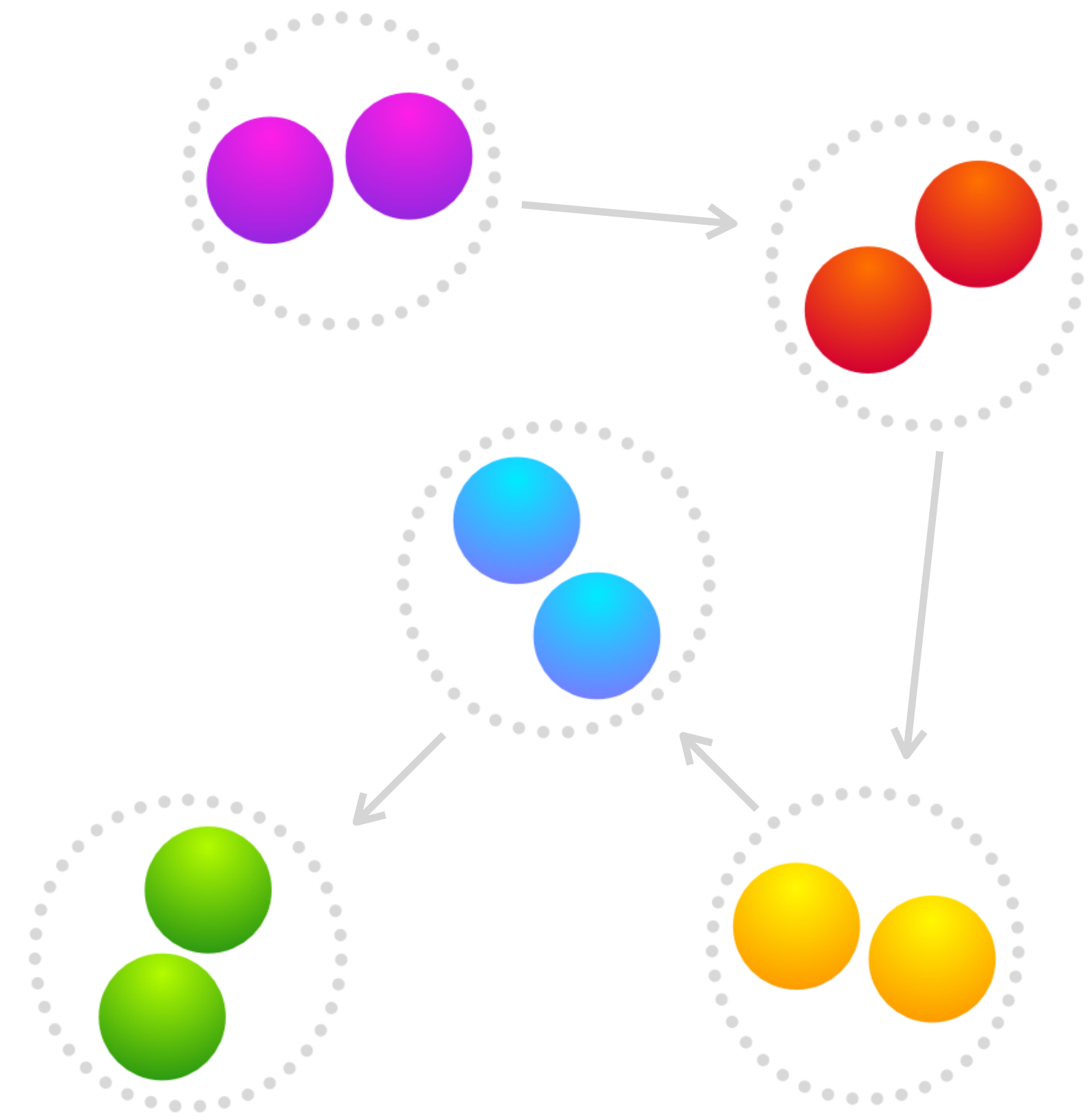
Subset of the Original Dataset?

## Goal

- Is there a **subset of the dataset** that can discourage disentanglement?
- Multi-objective problem — **overlap score** & dataset **usage ratio**

## How

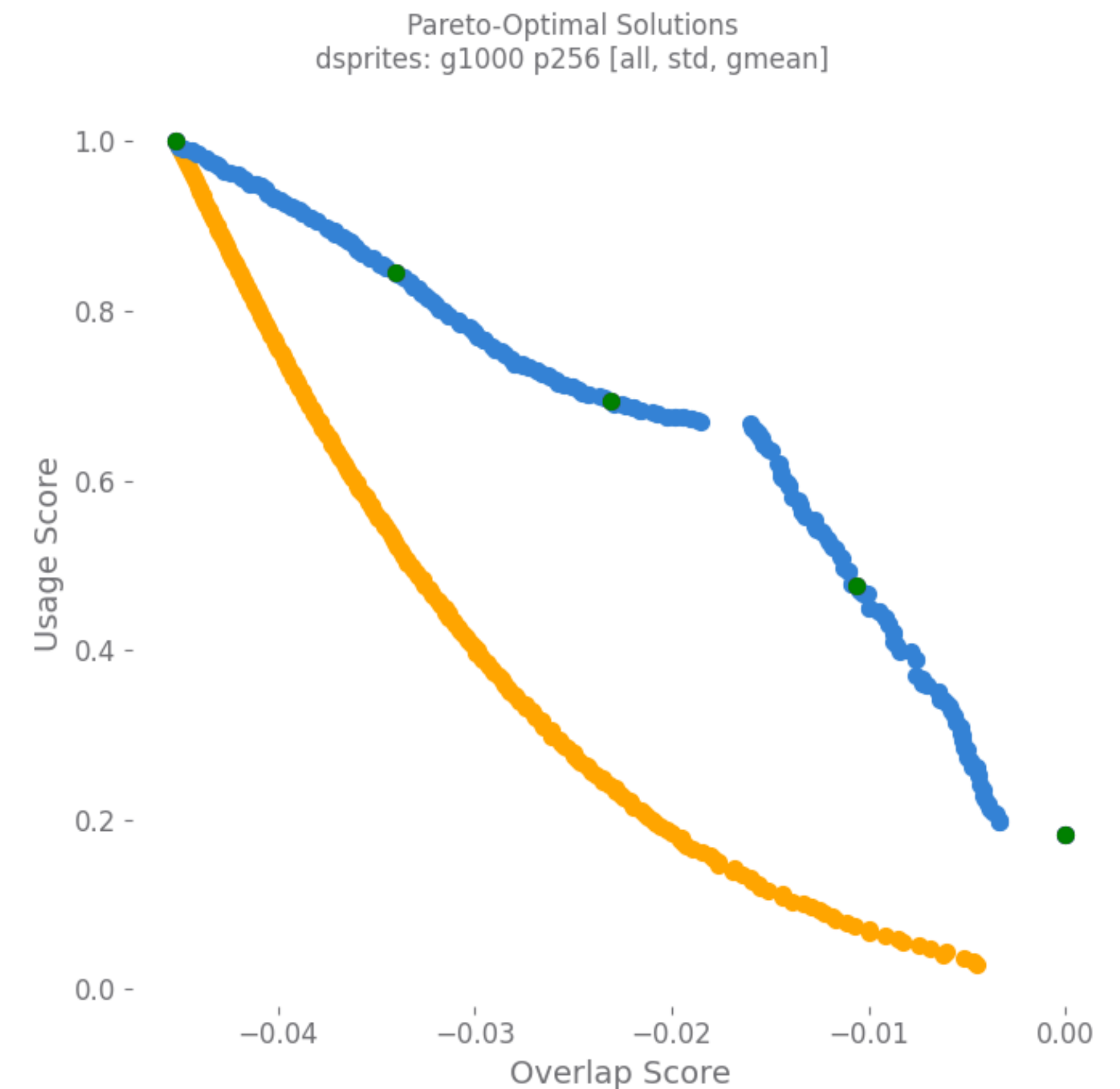
- **Intractable** in most cases to compute overlap between all pairs in a dataset —  $O(N^2)$  even if we precompute distances
- Use a **GA** to optimise for constant overlap?



# Masked Adversarial Dataset

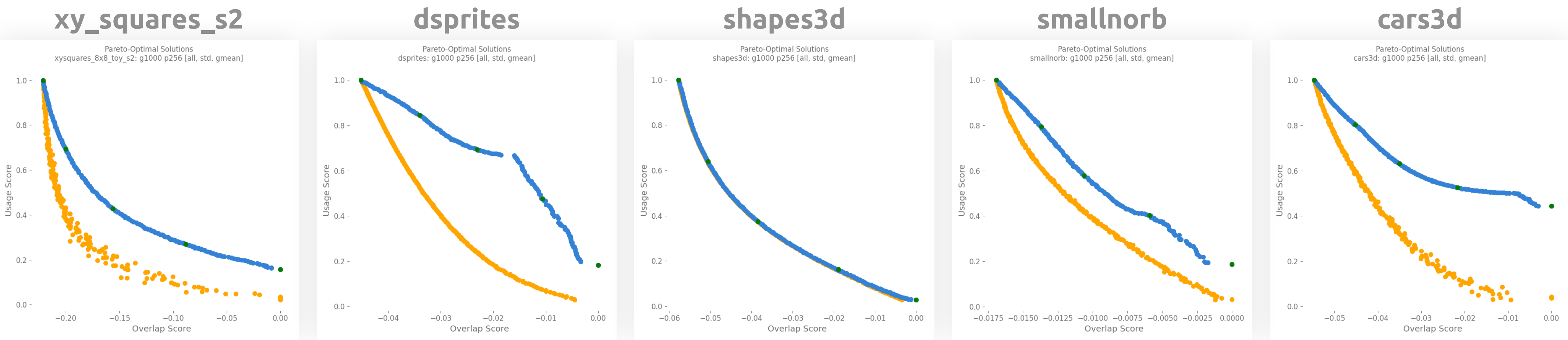
## Genetic Algorithm

- **Precompute** subset of **distance pairs**, hopefully this will capture dynamics of the dataset. A group of distances per element in the dataset.
- **Population** of dataset masks
- **Evaluate** each members overlap score as the average standard deviation of overlap values in a group across all groups
- **Mate and Mutate** members, flipping bits of the mask and using crossover
- **Select members** using NSGA-II to find the Pareto-optimal solutions of the multi-value objective



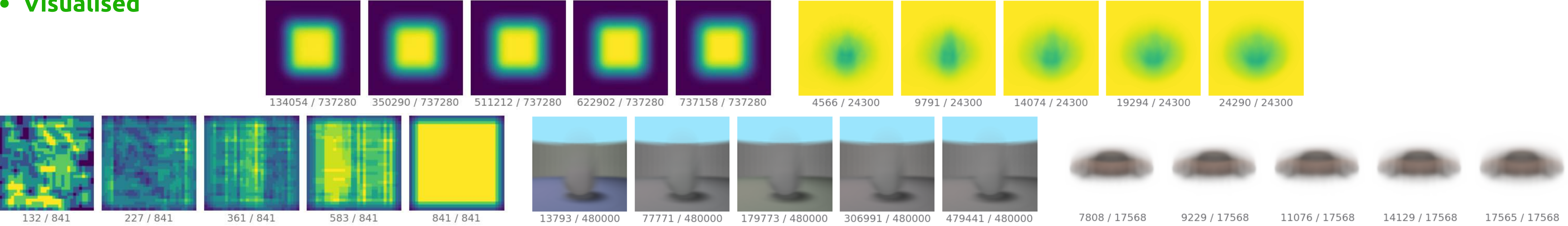
# Masked Adversarial Dataset

## Resulting Populations



Some datasets resist optimisation

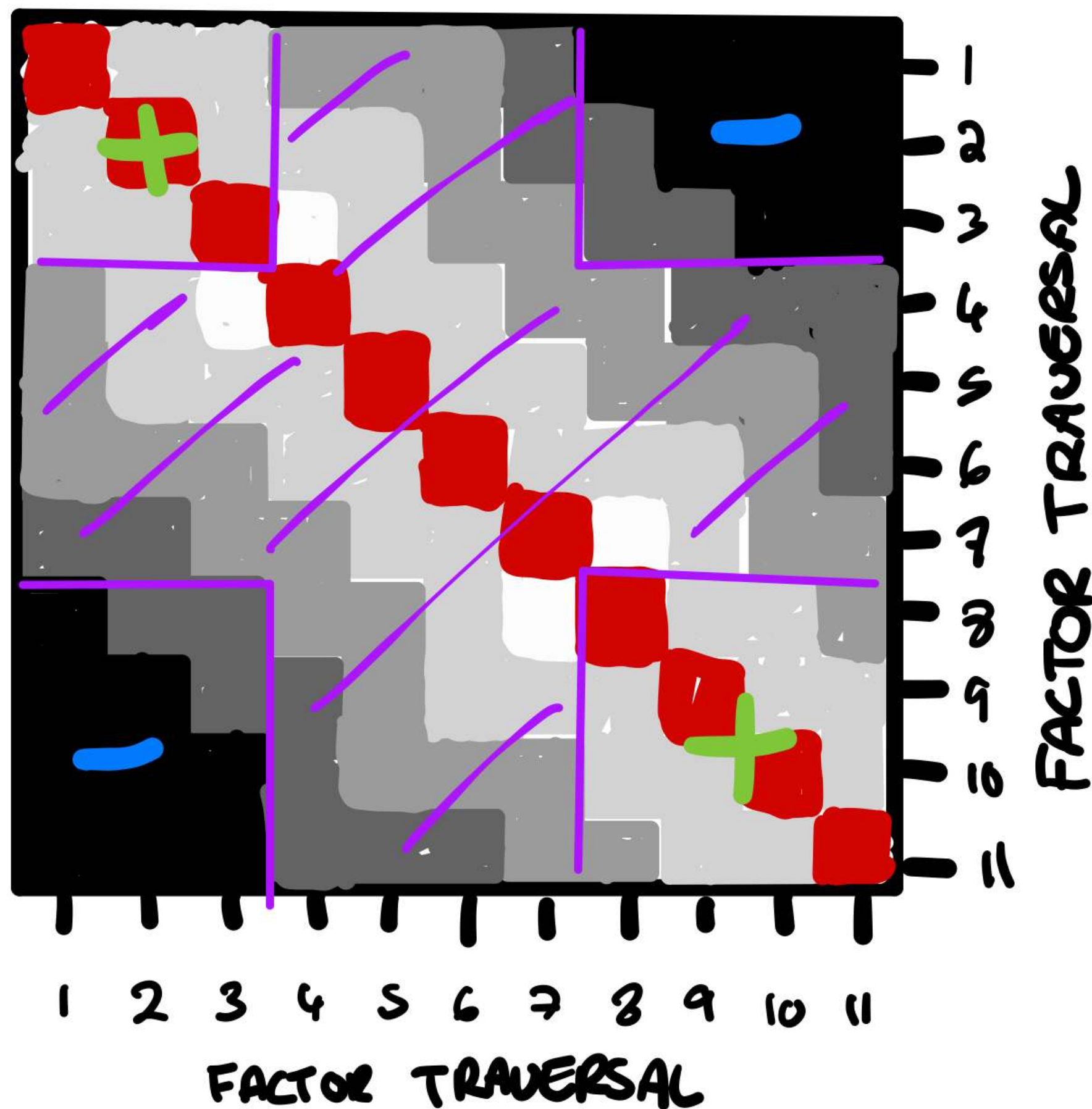
- Pareto-Optimal Solutions
- Random Solutions
- Visualised



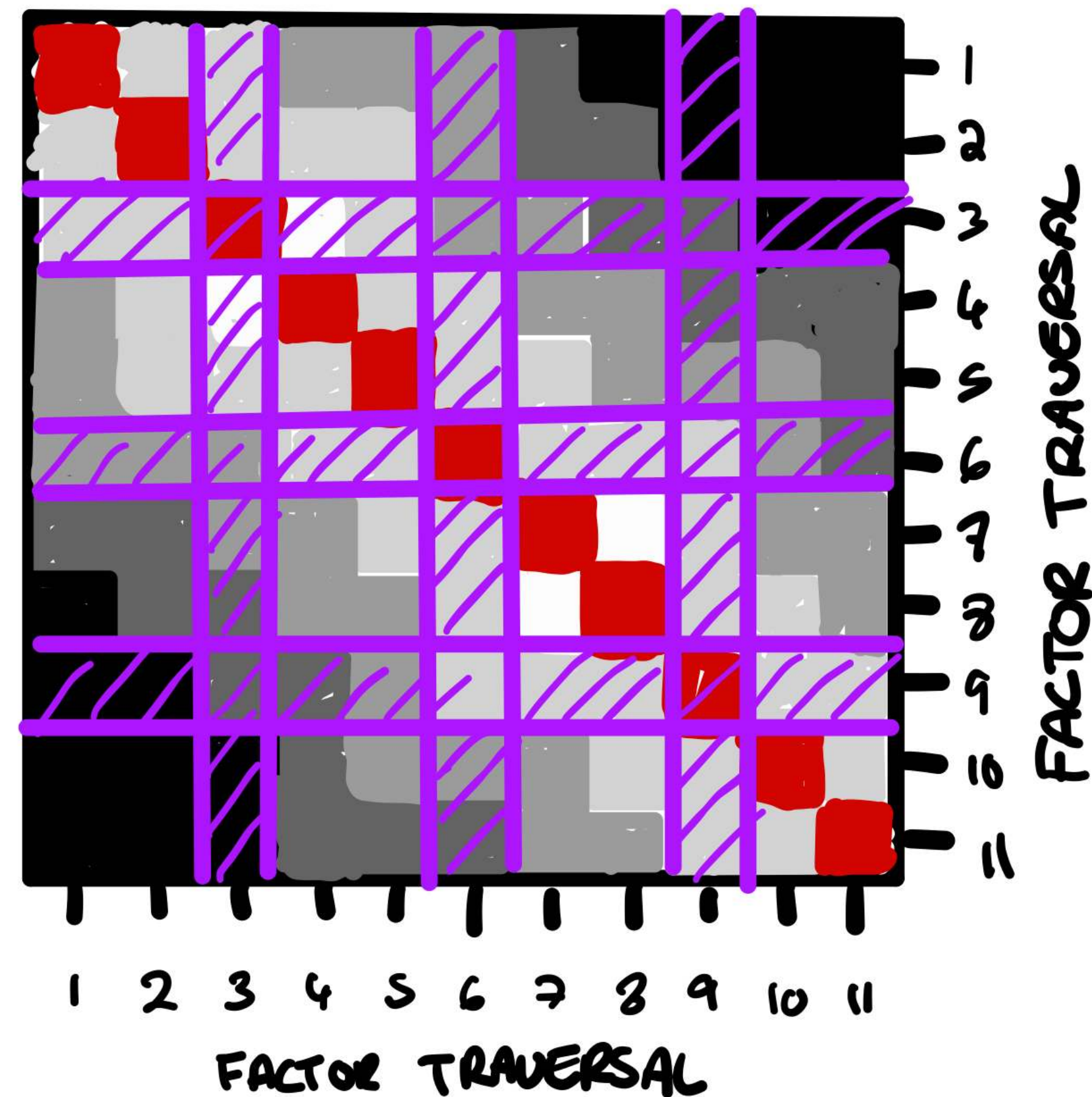
# Masked Adversarial Dataset

Are We Accidentally Cheating?

Discontinuities



Better, But Missing Examples





# Overlooked Implications of the Reconstruction Loss for VAE Disentanglement

Nathan Michlo<sup>[1,2]</sup>, Richard Klein<sup>[1,2]</sup>, Steven James<sup>[2]</sup>

<sup>[all]</sup> University of the Witwatersrand, Johannesburg, South Africa <sup>[1]</sup> Prime Lab, Johannesburg, South Africa <sup>[2]</sup> Rail Lab, Johannesburg, South Africa



We show how **data** & **recon. loss** lead to VAE disentanglement