

Accounting for the Sequential Nature of States to Learn Features for Reinforcement Learning

Nathan Michlo*, Devon Jarvis, Richard Klein^[1], Steven James^[2]

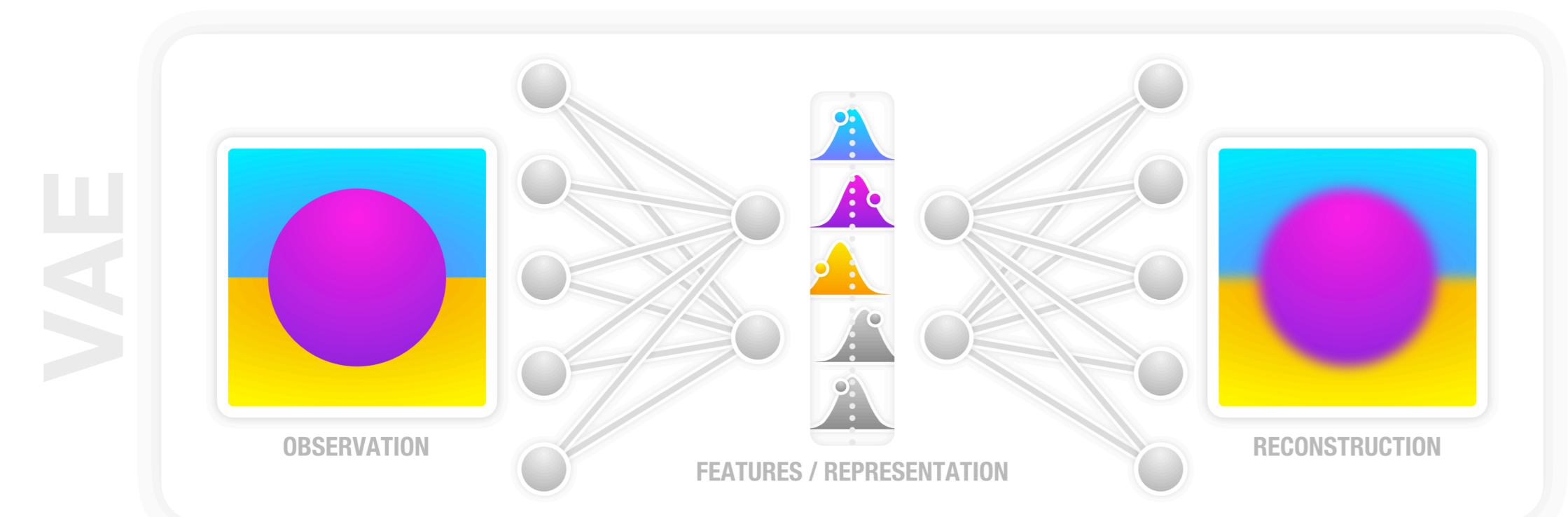
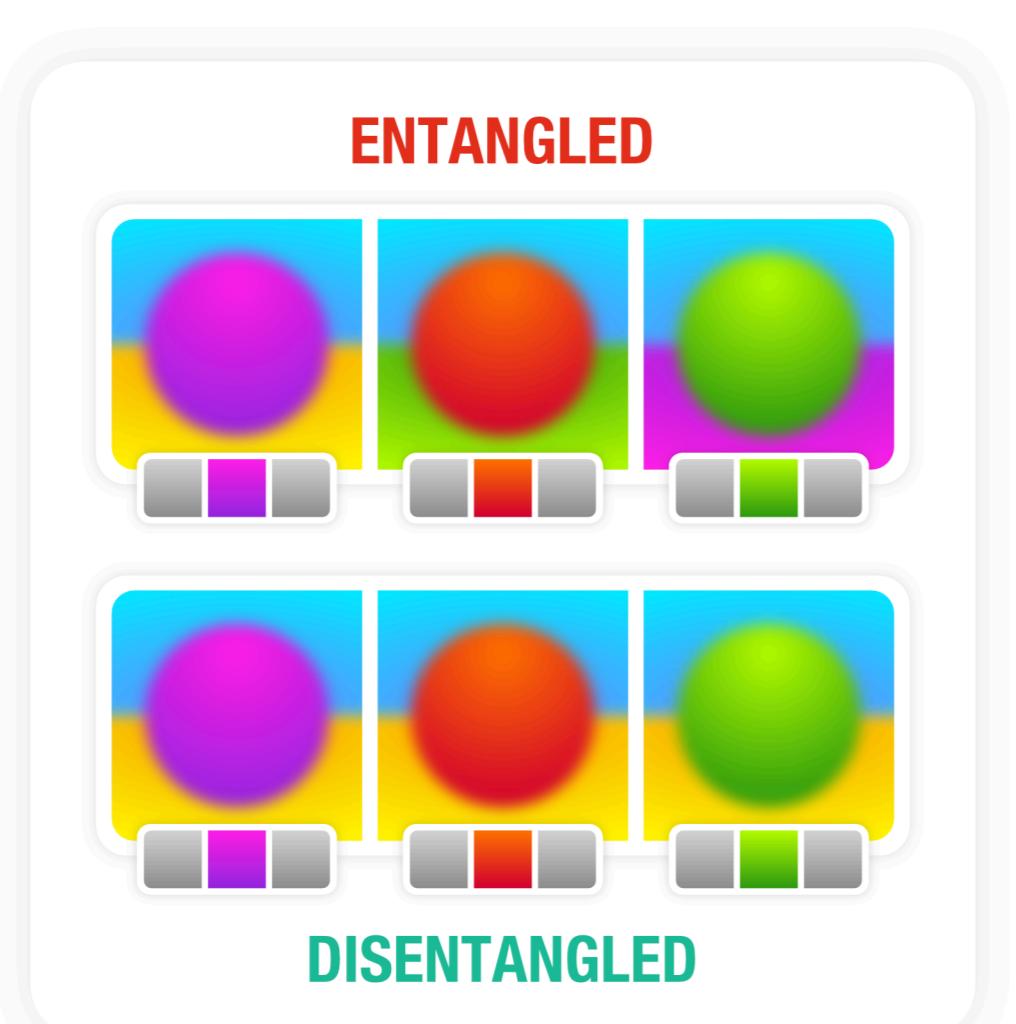
[all] University of the Witwatersrand, Johannesburg, South Africa [1] Prime Lab, Johannesburg, South Africa [2] Rail Lab, Johannesburg, South Africa



We use sequential states to learn better features for downstream tasks when we cannot rely on overlap in data

Introduction

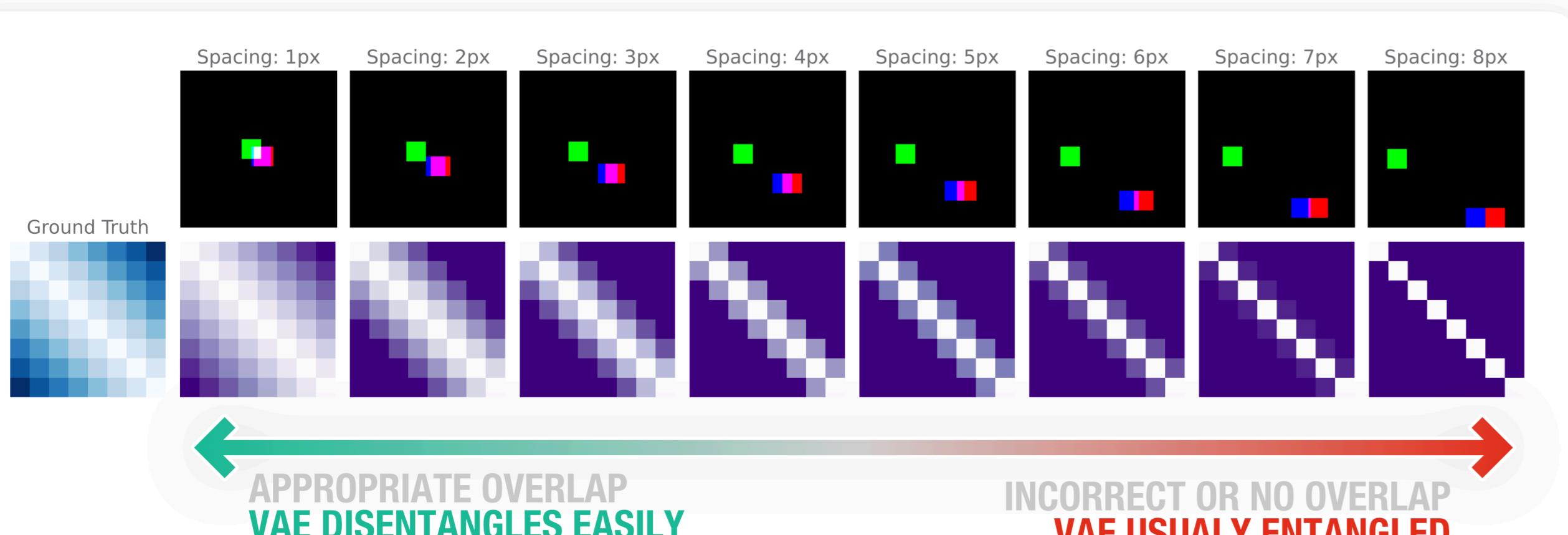
- Variational Auto-Encoders (VAEs) learn **disentangled** representations from data, which are useful features for downstream Reinforcement Learning (RL) tasks.
- VAEs may fail and produce **entangled** representations if data does not overlap eg. Chess piece on a chess board.
- Can we fix this by using the sequential nature of RL states?



VAEs Perceive Overlap In Data

VAEs perceive distances between observations based on the chosen reconstruction loss.

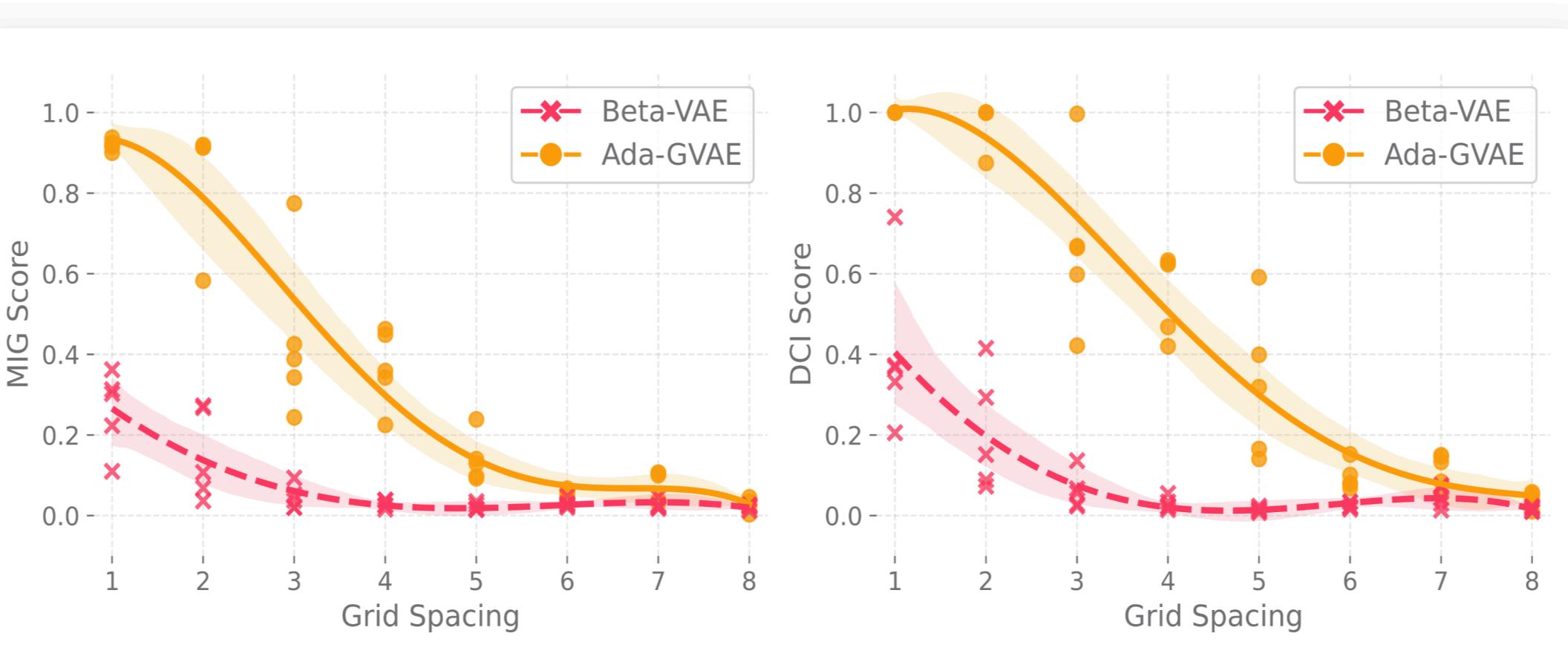
- We design a simple 8x8 gridworld domain with adjustable spacing of x and y ground-truth factors to test this.



- Train VAEs over gridworld datasets with varying spacing. More overlap in the data gives better disentanglement.

Our approach outperforms **standard triplet** and **VAE baselines**, when trained over our non-overlapping gridworld domain.

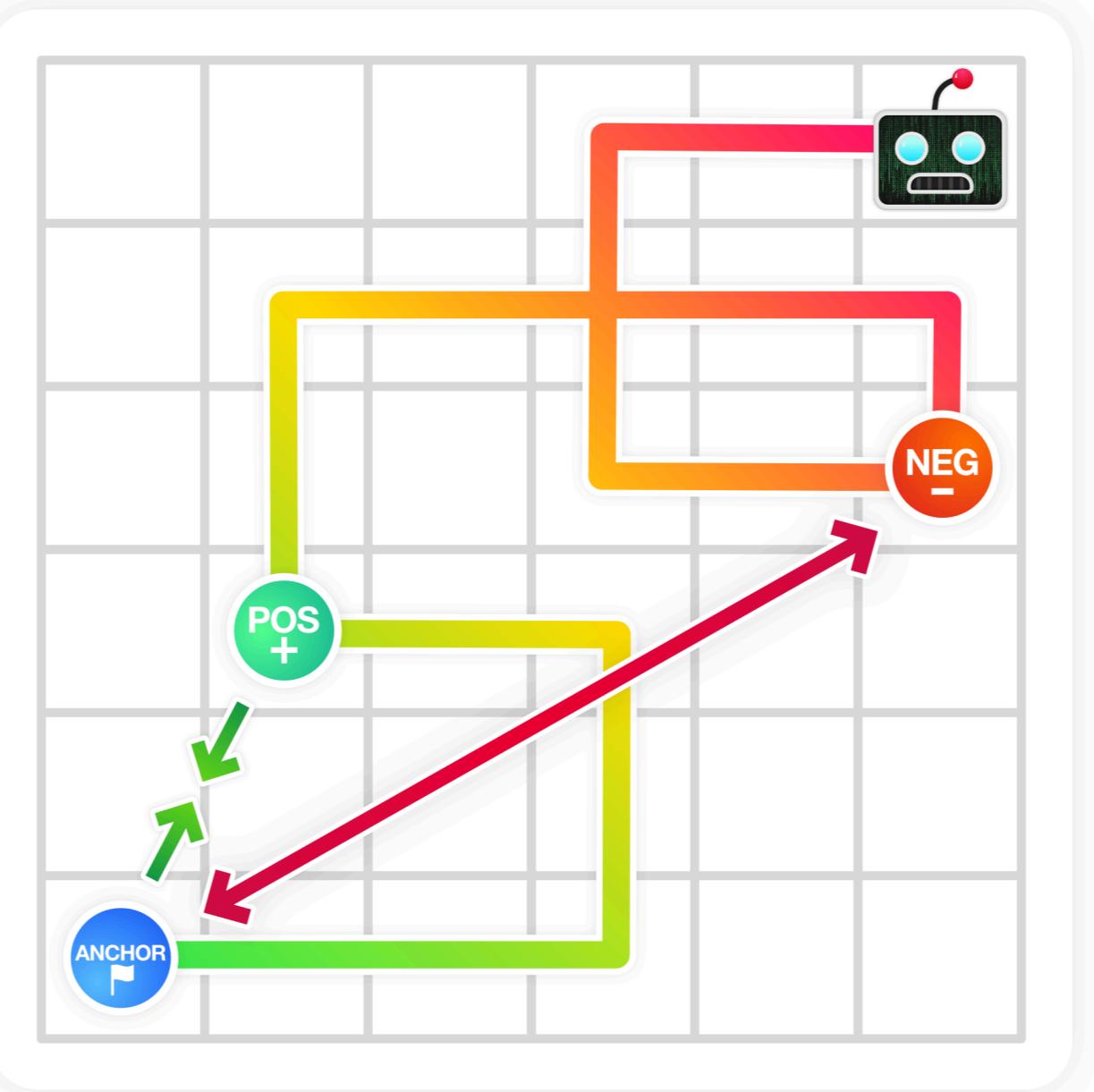
- We compare ground-truth constructed triplets versus those constructed from the replay buffer.



Using Sequential States

How can we remedy this problem for environments that do not contain sufficient overlap in data?

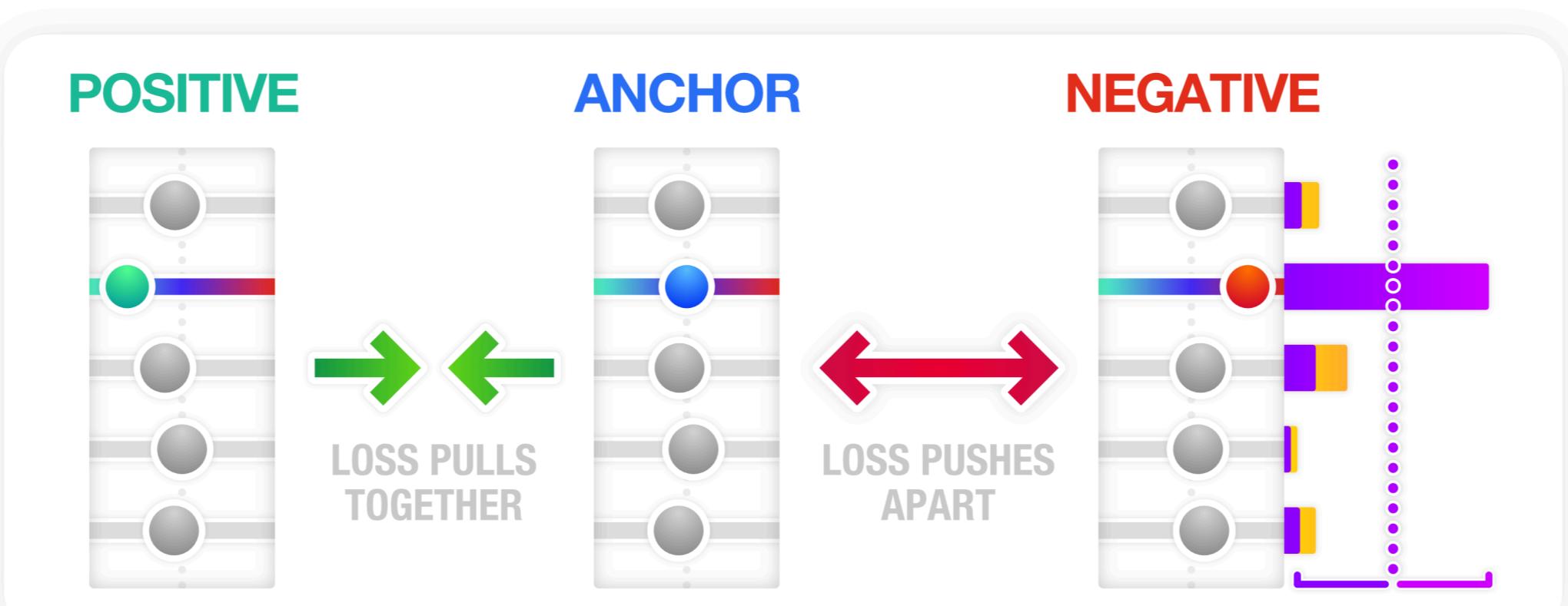
- Use an RL replay buffer and make the assumption that: *states further in time on average have more differing ground-truth factors*.
- Allows us to construct triplets for metric learning, where distances between representations of **anchor-positive** observations should be less than those of the **anchor-negatives**.



Adaptive Triplet Loss

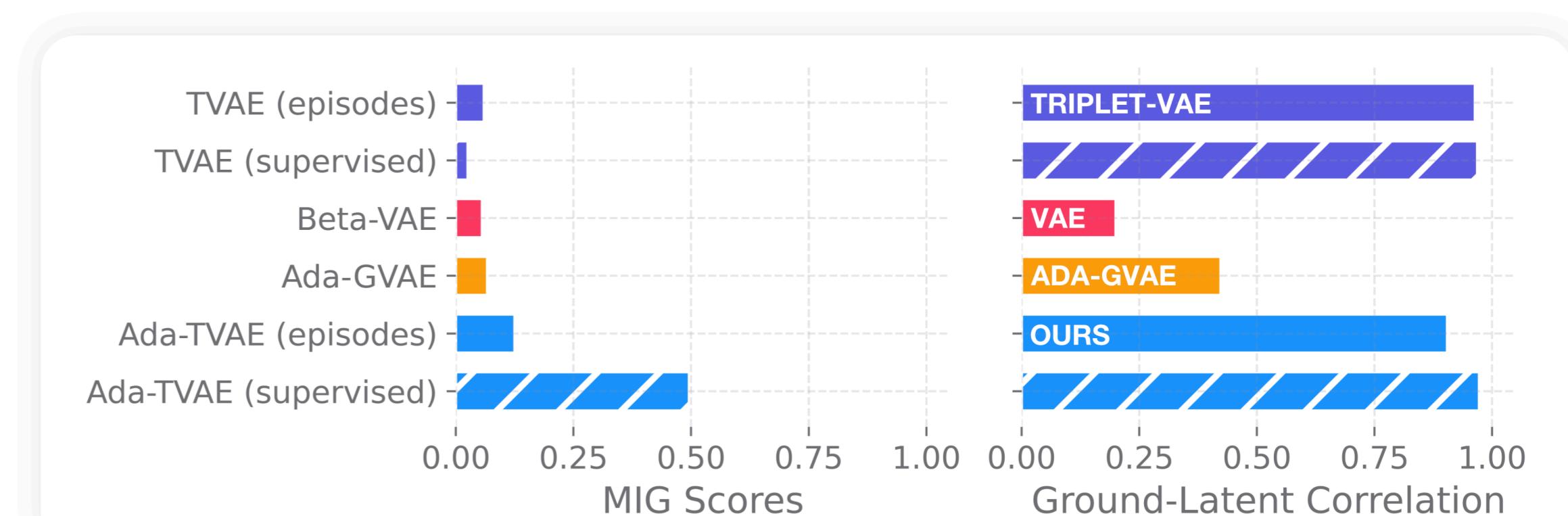
Standard triplet loss does not encourage disentanglement. So we instead encourage differences to be encoded in a subset of latent variables.

- We **weight** elements of the **anchor-negative** term less if their **absolute differences** are less than half way between the **min and the max**.



Our approach outperforms **standard triplet** and **VAE baselines**, when trained over our non-overlapping gridworld domain.

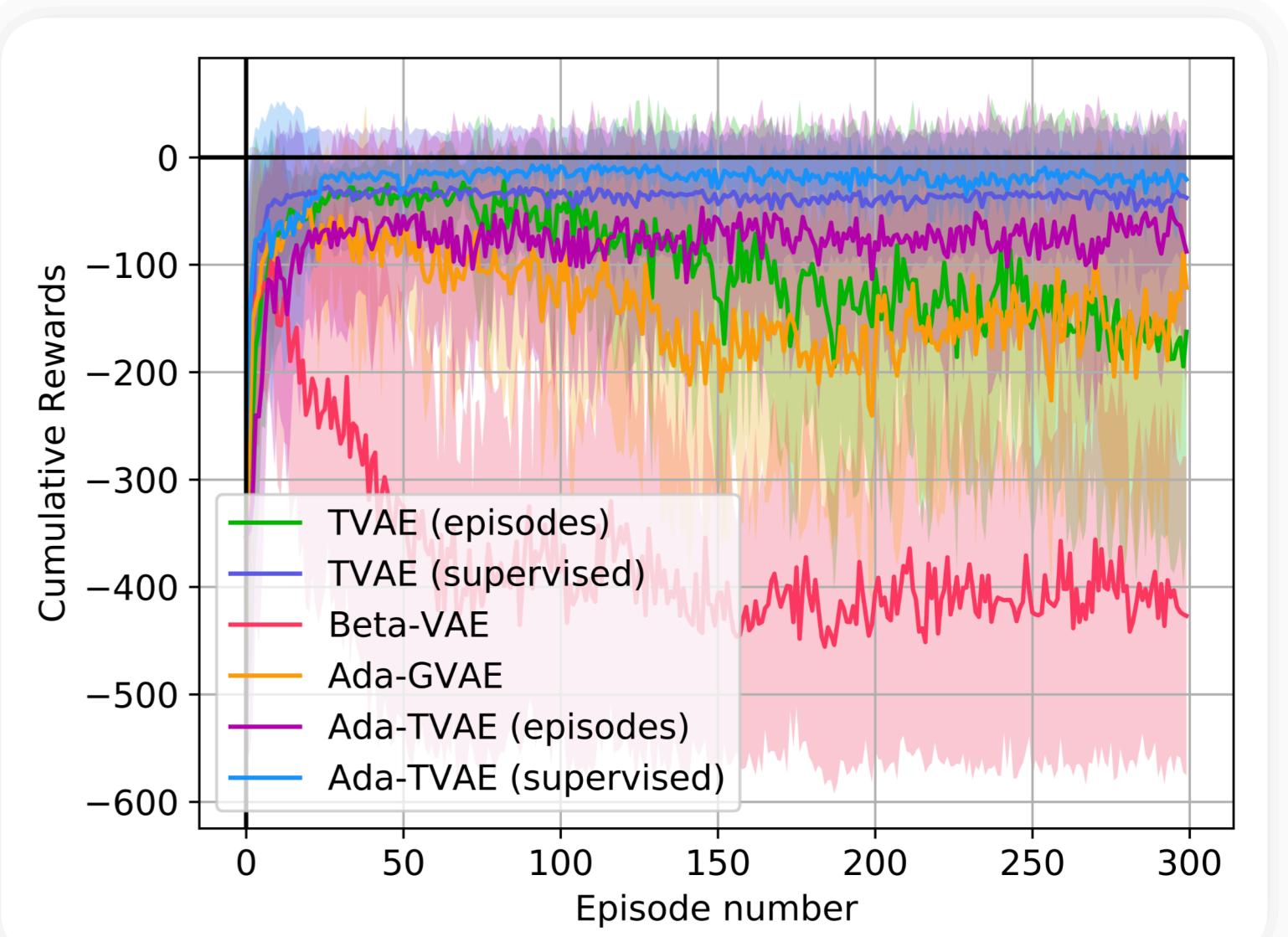
- We compare ground-truth constructed triplets versus those constructed from the replay buffer.



Downstream Tasks

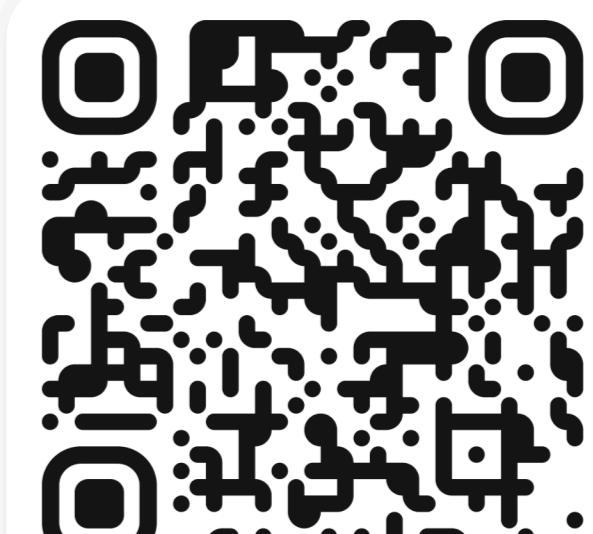
We use the previously trained VAEs to extract features used to make decisions by a downstream reinforcement learning agent

- The agent is tasked to move from one corner to another in the non-overlapping gridworld environment.
- Our adaptive triplet VAEs perform best.



Conclusion

Using sequential state information can improve the features learnt in problematic environments where data is non-overlapping and traditional methods fail



SCAN FOR PAPER & RESOURCES
or visit github.com/nmichlo