Latent Representation Disentanglement: beta and factor VAEs

Probabilistic Machine Learning

De Nardin Carlo

Index

01 Introduction to Disentanglement

02 Datasets and Metric

03 VAEs (standard, beta, factor)

04 Results

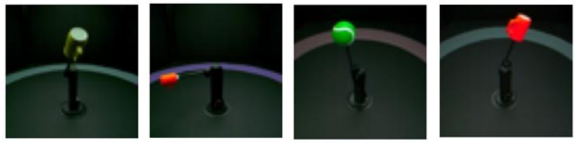
Disentangle VAEs

- The main objective for VAEs is not just data reconstruction, but also learning an **interpretable and compact** latent representation by understanding and separating the data's underlying generative factors.
- A representation is **disentangled** when a change in a single latent dimension (z_i) corresponds to a change in a single generative factor (f_i), leaving the others unchanged.
- Standard VAEs, however, often create **entangled** representations. In these models, a single latent dimension controls a mixture of factors, turning the latent space into a space in which is difficult to interpret and control these factors.

Datasets and Metric

- The *dSprites* dataset consists of 64x64 binary 2D images of simple shapes, with the following known ground-truth generative factors: shape, scale, rotation, position X and position Y.
- The *MPI3D* dataset consists of 64x64 images of complex shapes, with the following known ground-truth generative factors: color, shape, size, camera height, background color, horizontal axis and vertical axis.

dSprites dataset



MPI3D dataset

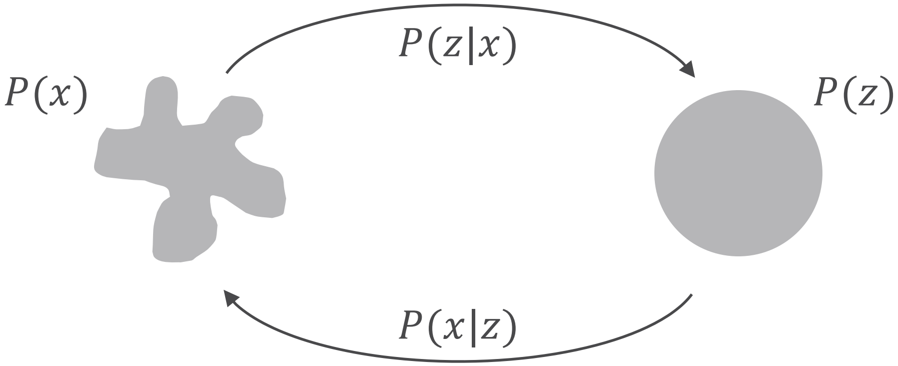
- The **Mutual Information Gap** (MIG) measures how much more information the single best latent dimension holds about this factor compared to the second best one.

$$MIG = \frac{1}{K}\sum_{k=1}^{K}\frac{I\left(z_{j^{(k)}}; v_k\right) - max_{j\neq j^{(k)}} I\left(z_j;\ v_k\right)}{H(v_k)}$$

$$I\left(z_j; v_k\right) = KL[p(z_j, v_k)|p(z_j)p(v_k)] = \sum\sum p\left(z_j, v_k\right)\log\frac{p(z_j, v_k)}{p(z_j)p(v_k)}$$

Models – VAEs

- VAEs are generative models that learn to map data into a probabilistic latent space and reconstruct back into the original data.



$$\mathcal{L}(\theta,\phi) = -\mathrm{E}_{z\sim q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + D_{KL}(q_{\phi}(z|x)||p(z))$$

- Standard VAEs aim for good reconstruction, but this often leads to entangled latent factors.

Models – Beta VAEs

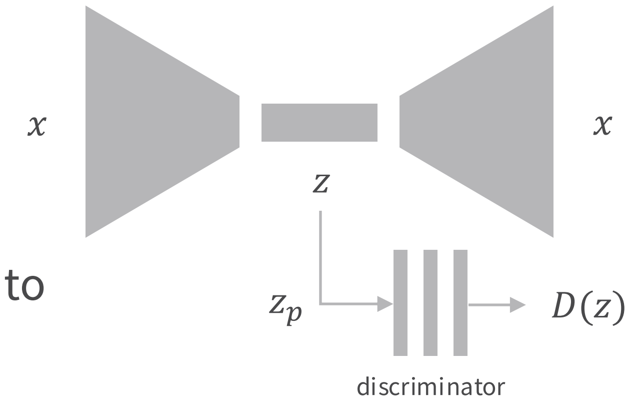
- The core idea is to give a greater weight to the KL regularization term. This enforces to learn a more efficient and disentangled latent representation by creating an information bottleneck.

$$\mathcal{L}(\theta,\phi) = -\mathrm{E}_{z\sim q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + \beta D_{KL}(q_{\phi}(z|x)||p(z))$$

$$\beta D_{KL}(q_{\phi}(z|x)||p(z)) = \beta I(x;z) + \beta D_{KL}(q(z)||p(z))$$

- A value of $\beta > 1$ penalizes the mutual information $I(x;z)$, forcing a trade-off: better disentanglement is achieved at the cost of a lower quality reconstruction.

Models – Factor VAEs



- Factor VAE introduces a new term, which is **Total Correlation**, in order to enforce the independence of the latent dimensions.

$$TC = D_{KL}(q(z)||\prod_{j=1}^{D} q(z_j))$$

- To approximate the TC a discriminator is trained to distinguish whether a latent vector **z** comes from the aggregated posterior rather than from the product of its marginals.

$$V(D) = E_{z\sim q(z)}[logD(z)] + E_{z\sim \prod_j q(z_j)}[\log(1 - D(z))] \qquad \mathcal{L}_D = -V(D)$$

- The encoder minimizes Total Correlation by encouraging independence between latent dimensions. It achieves this goal by fooling the discriminator until it can no longer distinguish the original latent vector from the permutated ones.

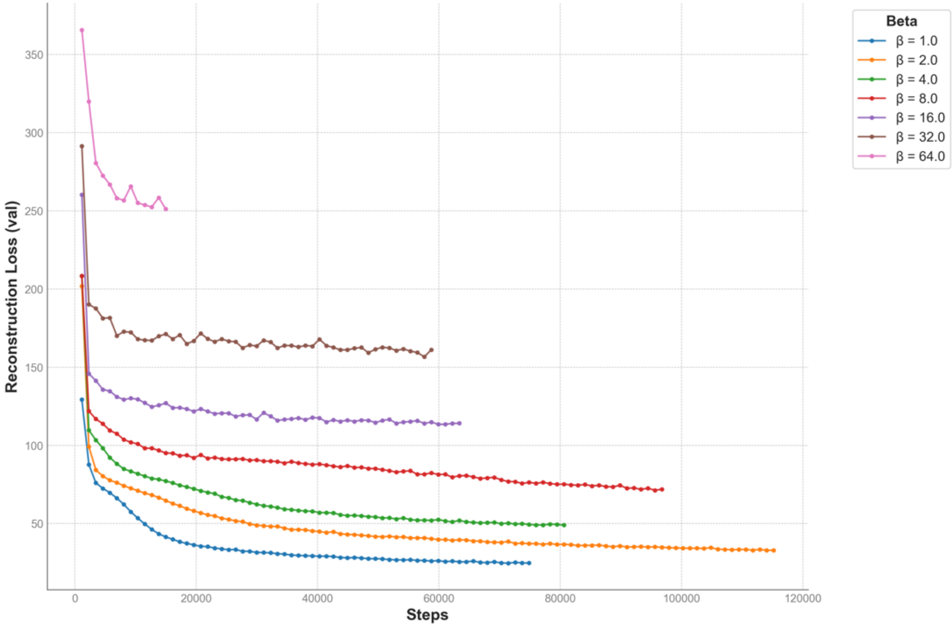
$$TC = \frac{1}{N}\sum_{i=1}^{N} log \frac{D(z^{(i)})}{1 - D(z^{(i)})}$$

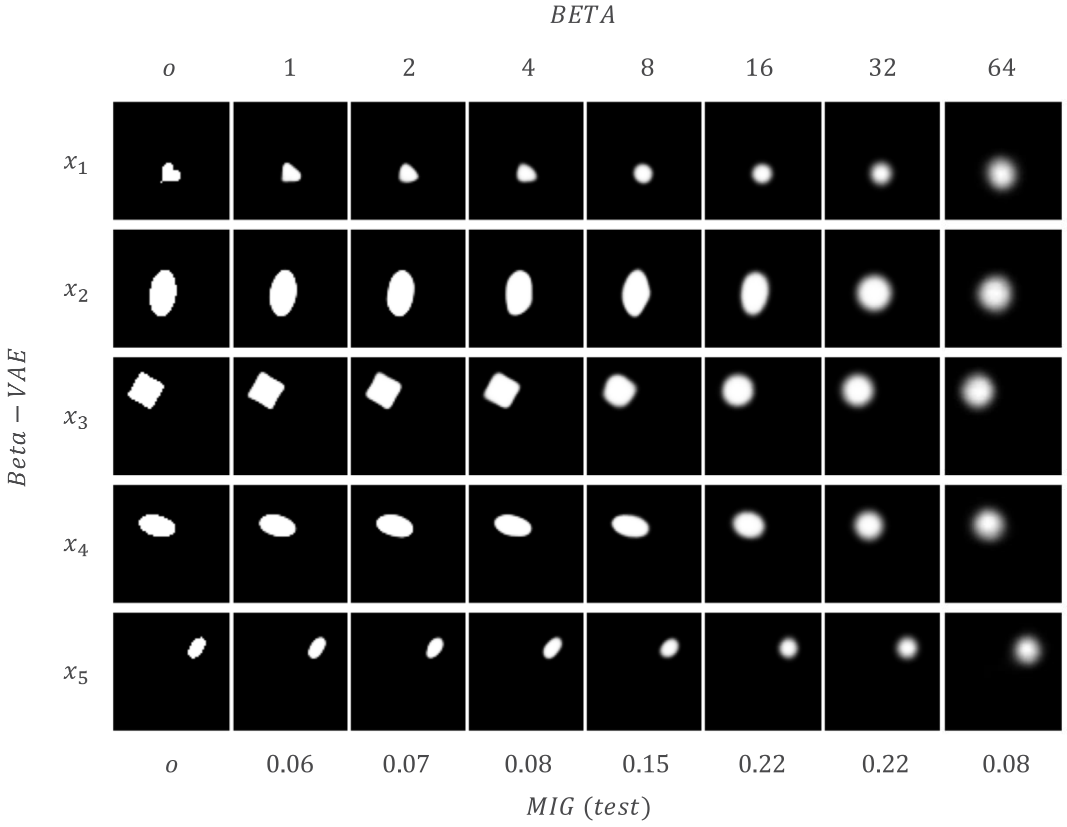
$$\mathcal{L}(\theta,\phi) = -\mathrm{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] + D_{KL}(q_\phi(z|x)||p(z)) + \boldsymbol{\gamma} D_{KL}(q(z)||\prod_{j=1}^{d} q(z_j))$$

Results – Beta VAE (dSprites)

Beta	Rec Error	KL
1	24.71	27.29
2	32.91	18.83
4	71.91	9.92
8	74.69	9.74
16	114.16	6.50
32	161.15	4.45
64	251.20	2.71

Table 01: validation loss

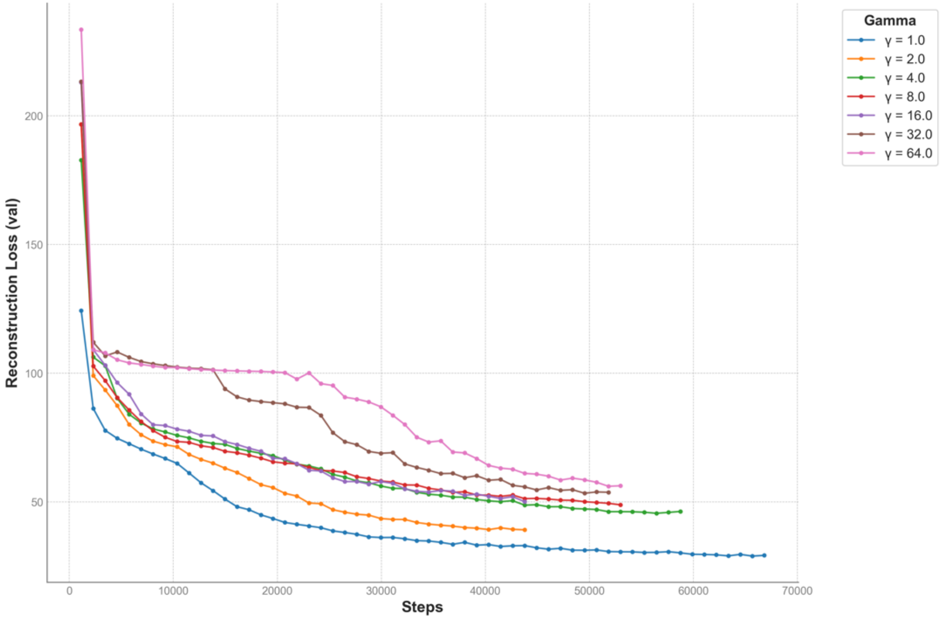


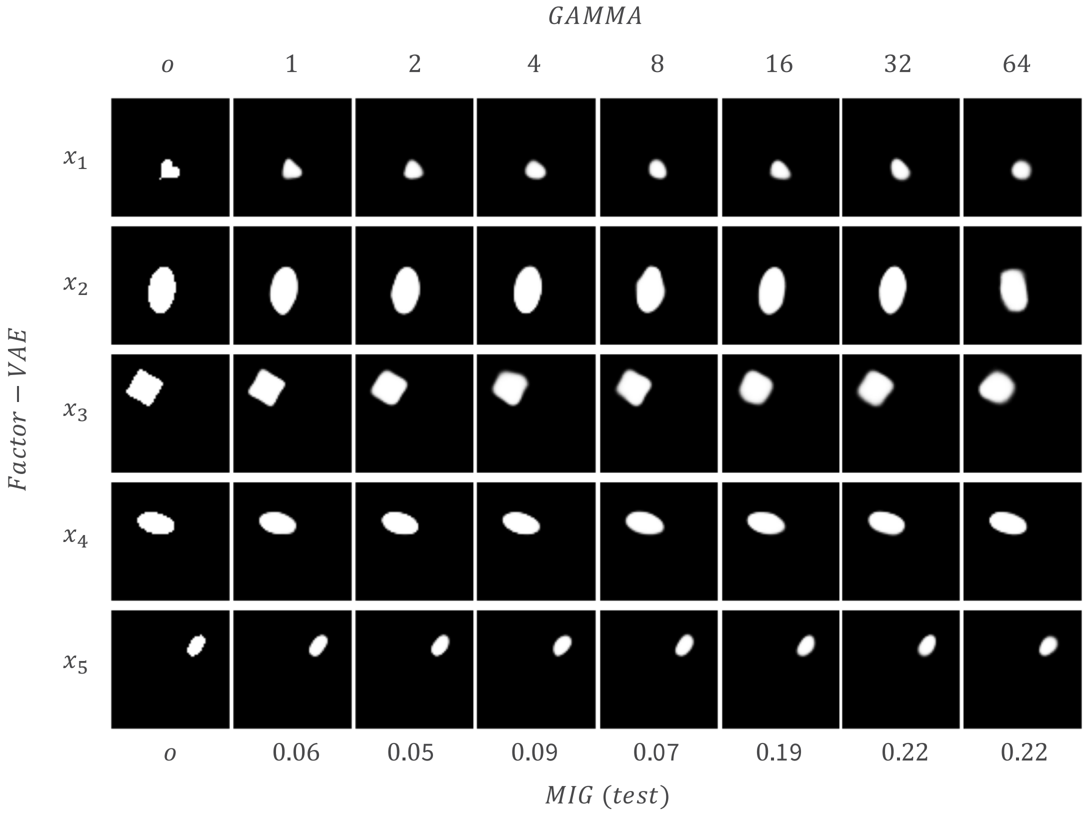


Results – Factor VAE (dSprites)

Gamma	Rec Error	KL	TC	$D(z_{real})$	$D(z_{perm})$
1	29.23	24.32	5.36	0.96	0.04
2	39.17	20.42	2.57	0.85	0.16
4	46.30	17.60	1.12	0.69	0.28
8	48.88	17.86	0.70	0.64	0.33
16	50.11	15.82	0.39	0.58	0.41
32	58.58	16.27	0.08	0.52	0.48
64	56.28	16.29	-0.20	0.45	0.43

Table 02: validation loss and discriminator prediction

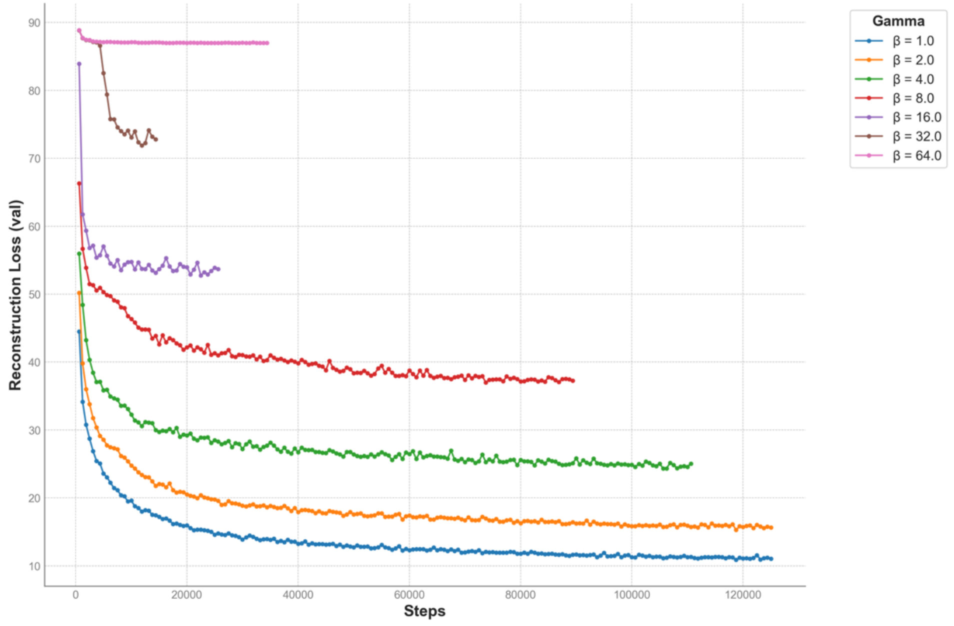


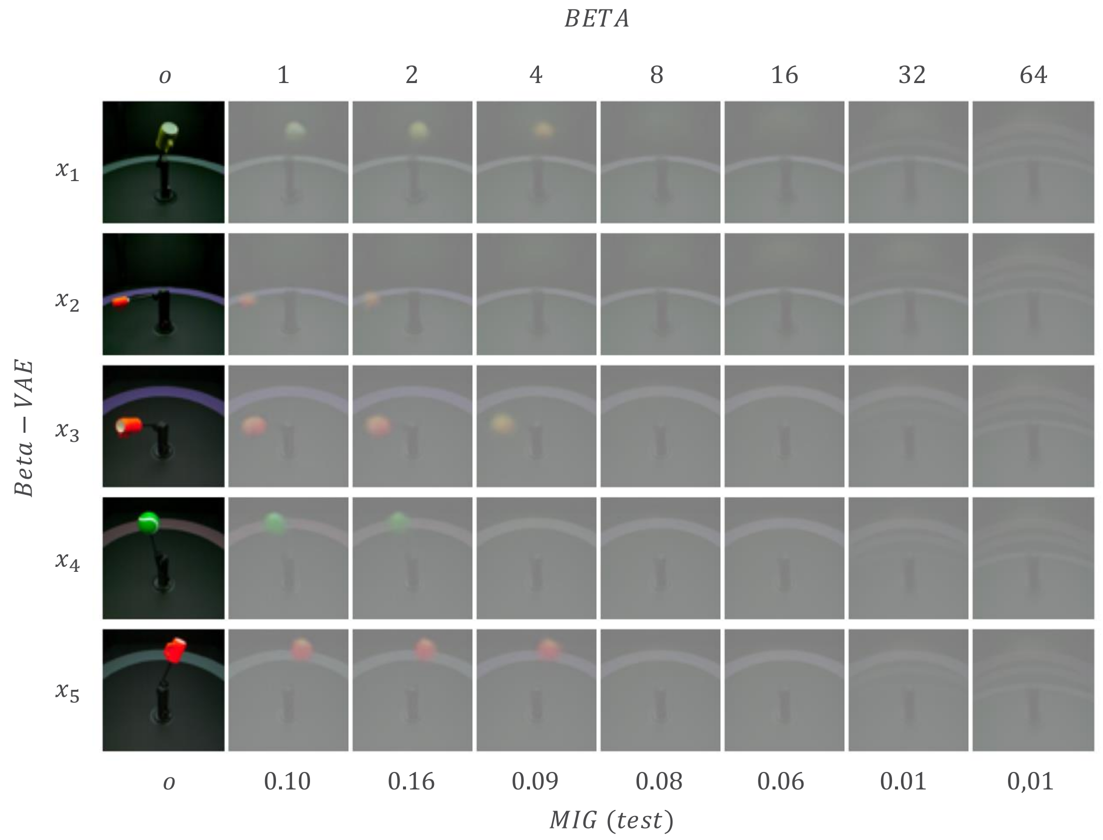


Results – Beta VAE (MPI3D)

Beta	Rec Error	KL
1	11.1	10.34
2	16.65	6.88
4	28.90	3.69
8	76.36	9.64
16	53.72	1.12
32	71.88	0.41
64	87.03	0.01

Table 03: validation loss



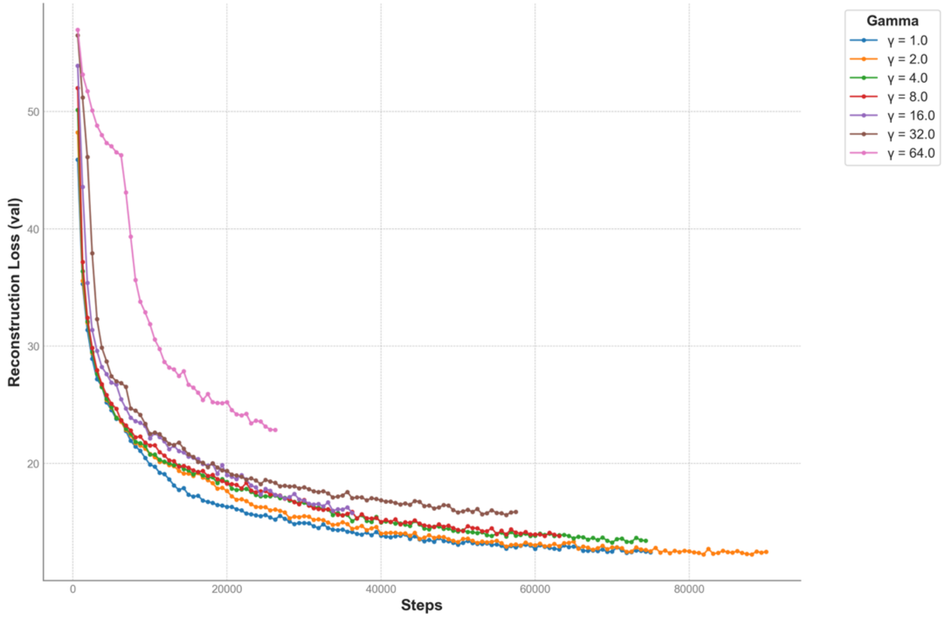


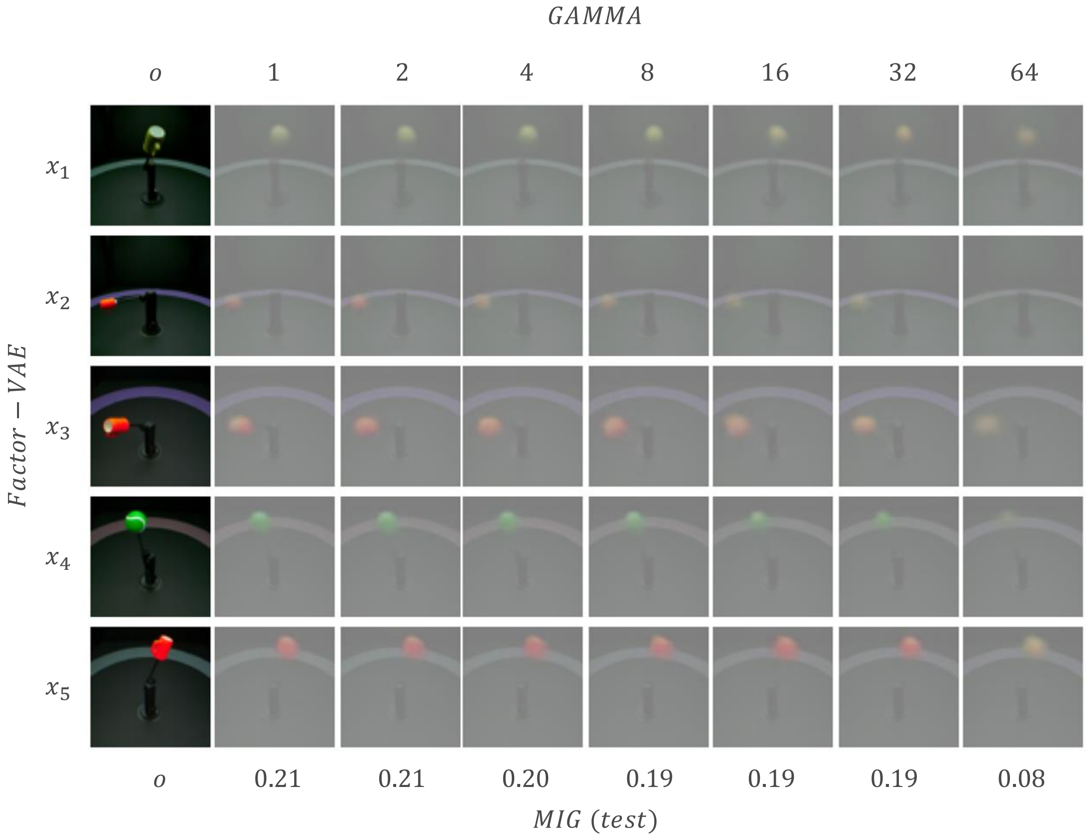
01
02
03
04

Results – Factor VAE (MPI3D)

Gamma	Rec Error	KL	TC	$D(z_{real})$	$D(z_{perm})$
1	14.91	9.12	0.3	0.55	0.43
2	12.94	9.32	0.23	0.54	0.47
4	14.01	9.25	0.23	0.54	0.47
8	13.81	9.21	0.09	0.52	0.49
16	16.68	8.27	0.001	0.48	0.48
32	16.24	8.47	0.03	0.5	0.5
64	28.18	5.92	0.02	0.5	0.5

Table 04: validation loss and discriminator prediction





Conclusion

- Beta VAE shows a clear **reconstruction–disentanglement trade-off**: as beta increases, reconstruction error degrades, while disentanglement improves.
- Factor VAE achieves a better balance by penalizing **Total Correlation (TC)**. It maintains a stable reconstruction and reaches a MIG value up to 0.22 on *dSprites* and 0.21 on *MPI3D*. The increasing confusion of the discriminator confirms improved latent independence with higher γ.
- On the complex *MPI3D* dataset, both models achieved the highest MIG value with low beta or gamma values. This suggests that richer data may inherently support better factor separation, even with weaker regularization.

References

- Chen, T. Q., Li, X., Grosse, R. B., & Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations*
- Kim, H., & Mnih, A. (2018). Disentangling by factorising