SIAM DATA MINING 2026 · SALT LAKE CITY

Sculpting the Latent Space

Disentanglement with WAE-MMD and Programmable Priors

Quentin Fruytier[1] · Akshay Malhotra[2] · Shahab Hamidi-Rad[2] · Aditya Sant[2] · Aryan Mokhtari[1] · Sujay Sanghavi[1]

[1] The University of Texas at Austin · [2] InterDigital Communications, AI Lab

THE GOAL

Disentanglement asks for two different things

1

Mutually independent latents

No latent unit can be predicted from the others, over the whole dataset.

2

Semantically meaningful latents

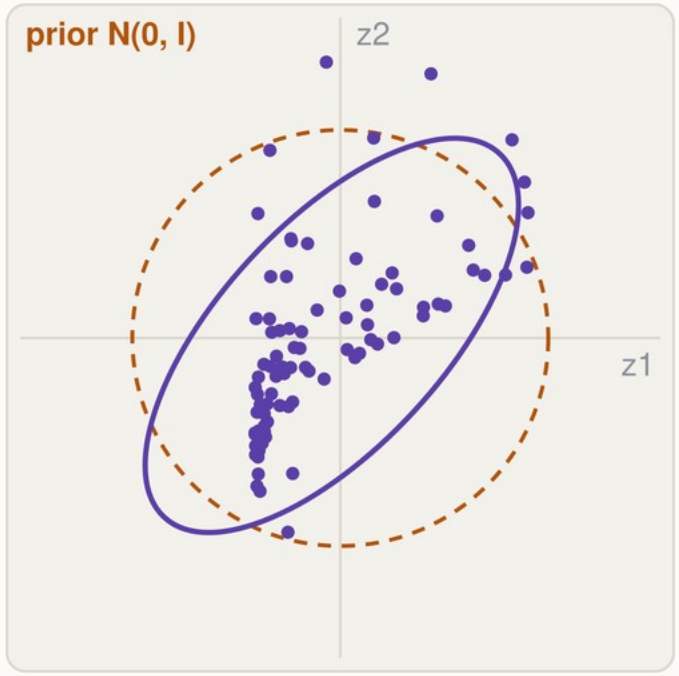
Each unit lines up with a real factor: shape, scale, position, colour.

Without supervision, (2) is not identifiable: many independent representations fit the same data. Unsupervised methods can only aim at (1). We measure (1), enforce it, and then inject (2) through the prior.

Hyvärinen & Pajunen 1999; Locatello et al. 2019, 2020; Khemakhem et al. 2020

THE FAILURE

A per-sample KL does not shape the aggregate

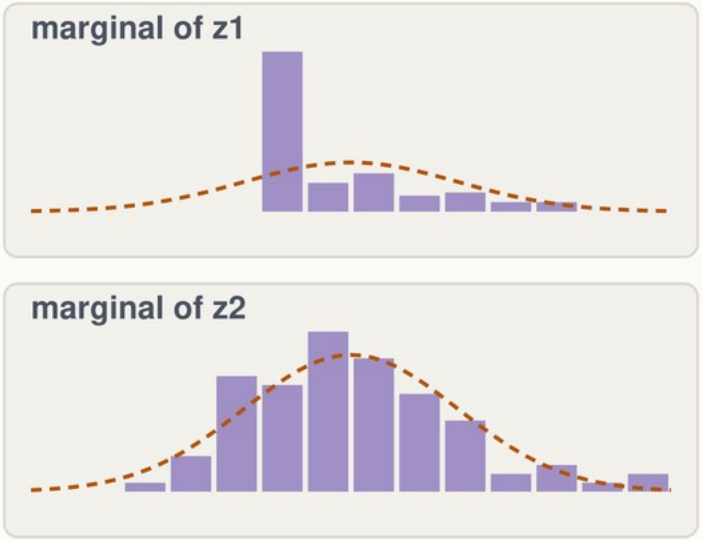


aggregate of all codes
tilted and skewed, unlike the prior

covariance over the data

0.81	0.56
0.56	0.90

target: 1 on the diagonal, 0 off it
correlation 0.66



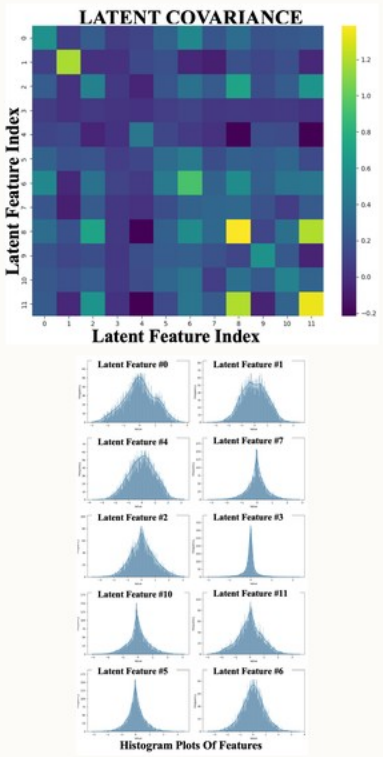
bars: the codes · dashed: the Gaussian the prior asks for

Schematic in two dimensions · the next slide shows the same failure measured on MNIST

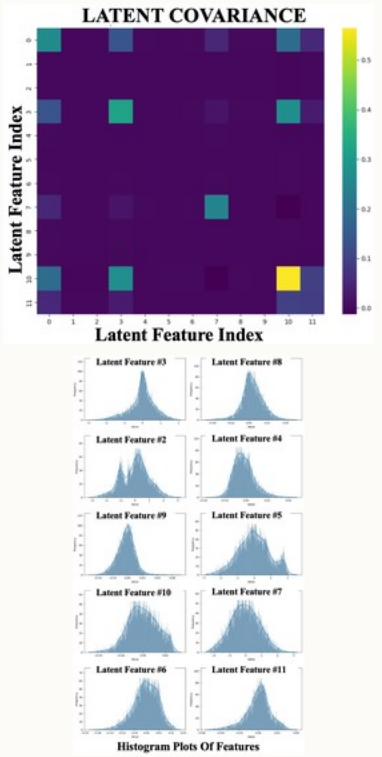
THE EVIDENCE · MNIST, 12 LATENT DIMENSIONS

Only the MMD model matches the prior

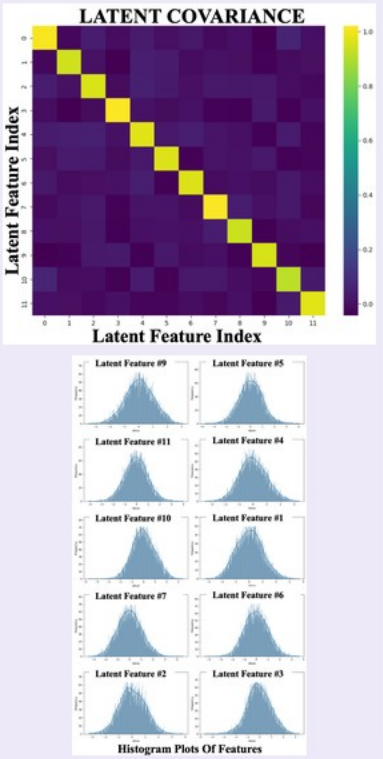
Standard VAE



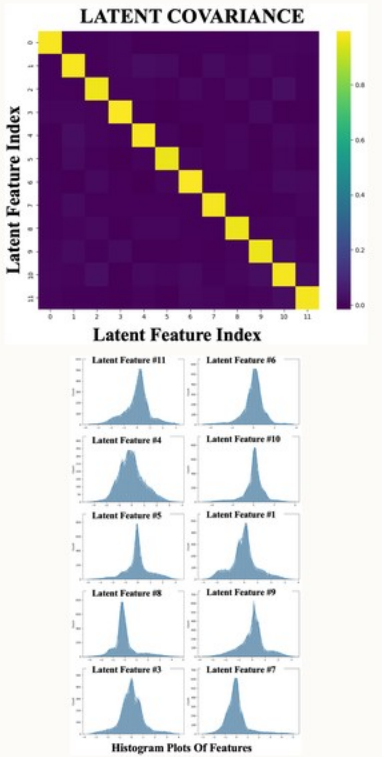
β-VAE (β = 4)



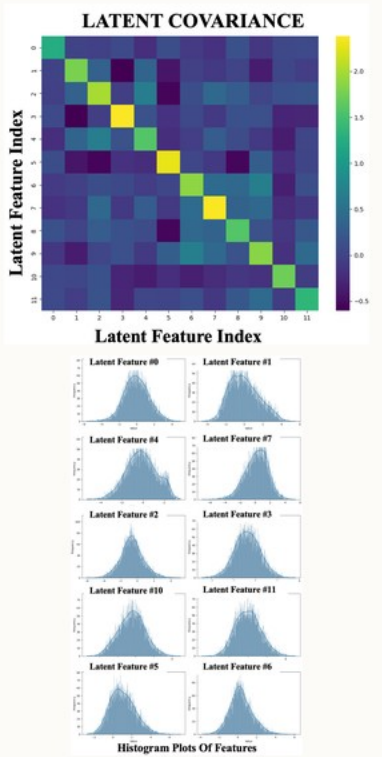
SWAE-MMD (ours)



AE + batch KL



Standard AE



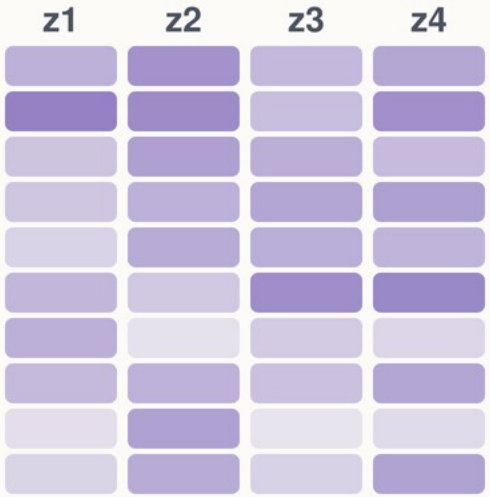
Top: covariance of the latents over the dataset · bottom: their marginals · batch KL fixes two moments, yet its LPS is 0.97

MEASURING IT WITHOUT LABELS

The Latent Predictability Score

model: SWAE-MMD (ours) · CIFAR-10

encoded latents, one row per input

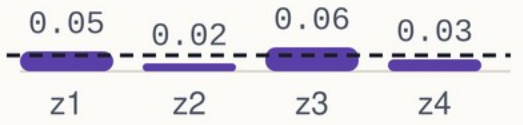


regressor
target from the rest

predicted

actual

R^2 per dimension



LPS = mean R^2 = 0.04

LPS near 0: no latent can be predicted from the others

4 of 64 dimensions shown; per-dimension R^2 illustrative · LPS values are Table 1's CIFAR-10 numbers (LightGBM regressor)

WHY THE SCORE IS SAFE TO TRUST

LPS can only underestimate entanglement

R^2 of any regressor $\leq \eta^2$, the share of a latent's variance the others truly explain

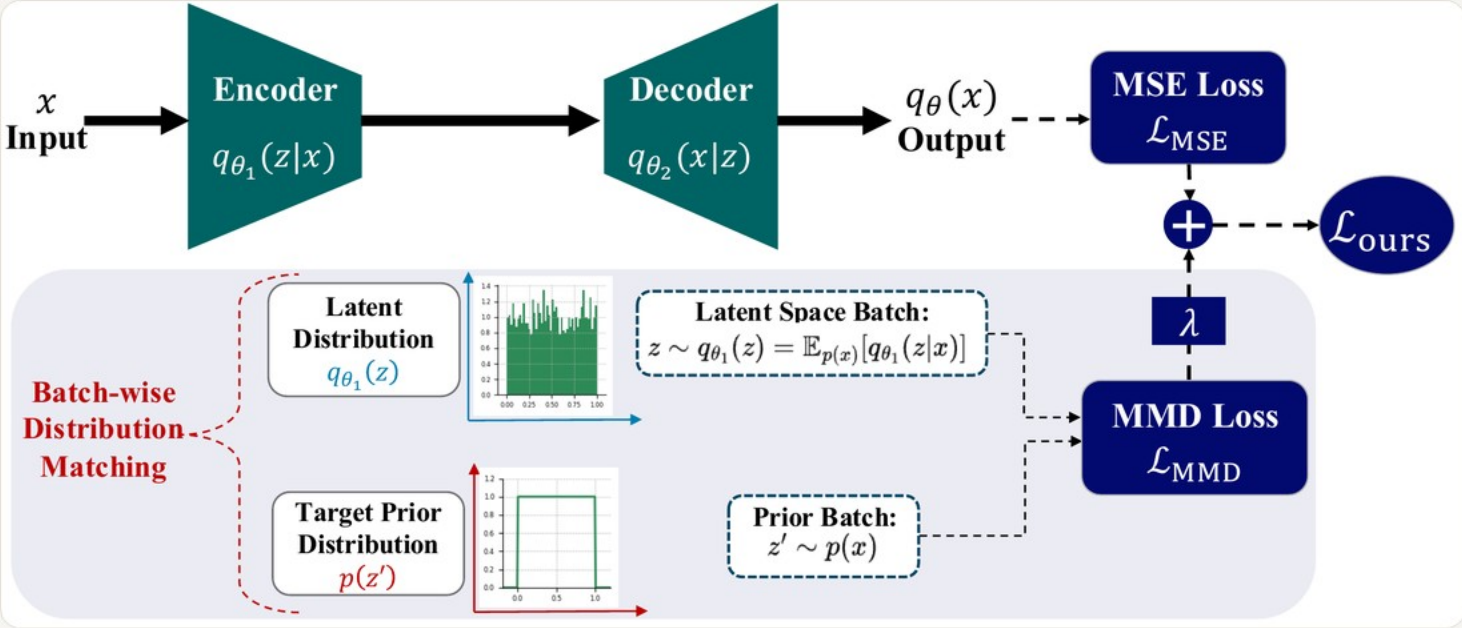
No regressor beats the conditional mean. So an LPS of 0.5 proves that at least half of the latent information is entangled, whatever the ground-truth factors are: a rejection test that needs no labels.

THE FIX · SWAE-MMD

Match the aggregate posterior to prior samples

```
maximise E[ log q(x) ] − λ · MMD²( q(z), p(z) )
```

reconstruct the input, and make the distribution of all codes match the prior, using only samples of the prior



The regularizer acts on the **batch**, not on each sample: it sees the whole aggregate shape.

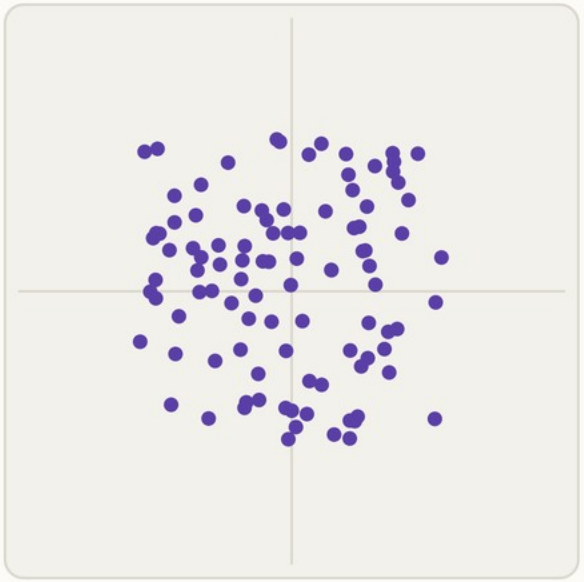
It needs **no density** for p(z), only a way to sample it. That is what makes the prior programmable.

A Wasserstein autoencoder variant (Tolstikhin et al. 2018) used for representation engineering · batch 512, λ = 0.3

SCULPTING, ONE BATCH AT A TIME

MMD pulls the codes onto any prior

loss = reconstruction error + λ · MMD²(latent batch, prior batch)



latent batch z = encoder(x)
what the encoder produces

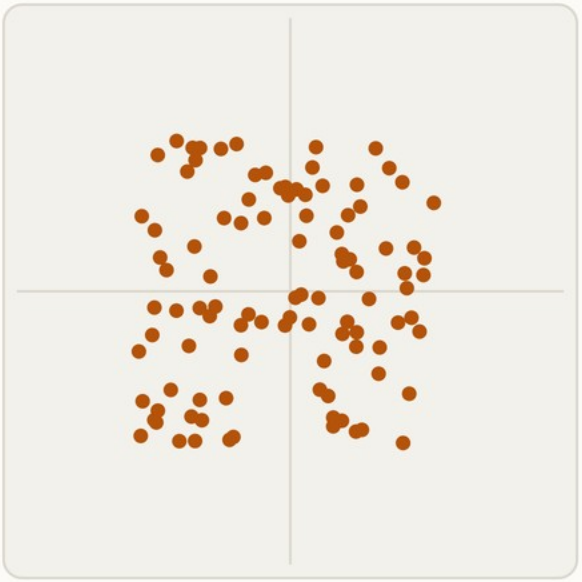
MMD²

0.014

correlation(z1, z2)

-0.03

same loss, any prior you can sample from



prior batch z′ ~ p(z)
prior: uniform

Two latent dimensions, schematic training · MMD² and the correlation are computed live from the points shown (Gaussian kernel)

RESULTS · 8 DATASETS

70% less entanglement than the next best method

LPS (mean R^2, lower = more independent); latent size in brackets

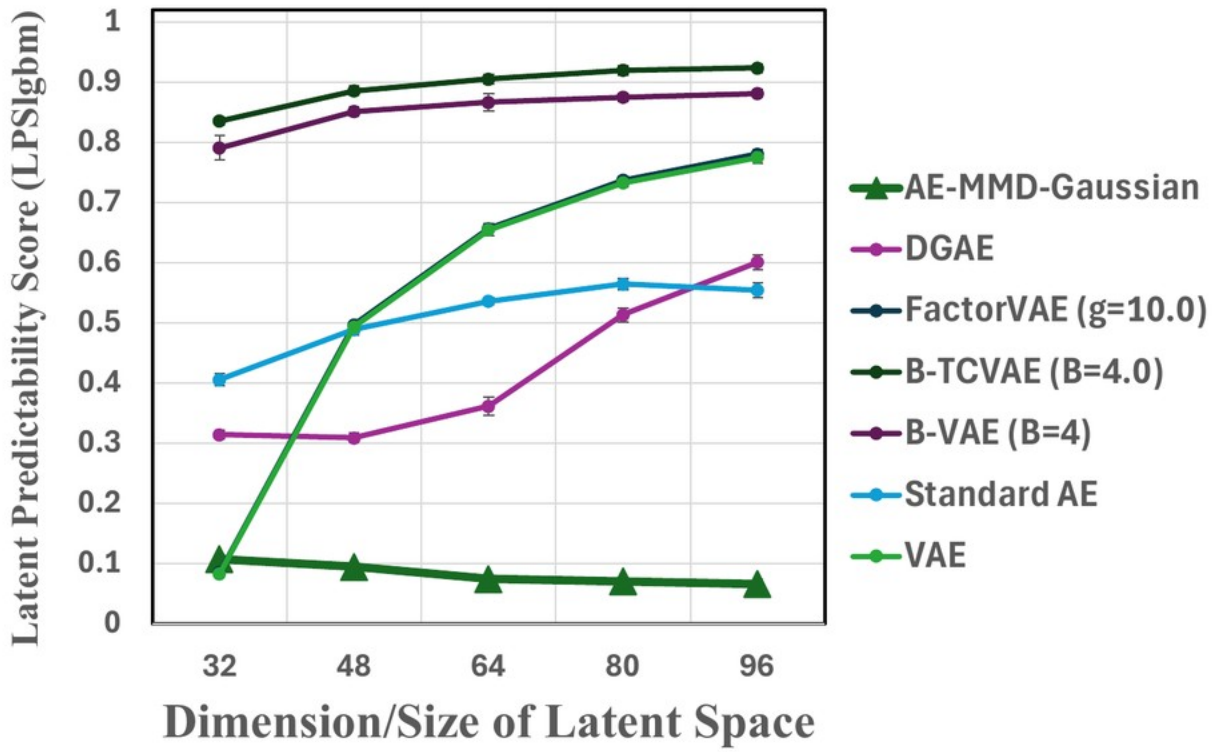
model	ImageNet [64]	CIFAR-10 [64]	MNIST [12]	dSprites [5]	XY [2]	XYC [3]	XYCS [4]	CelebA [32]
SWAE-MMD (ours)	**0.03**	**0.04**	**0.22**	**0.04**	−0.05	**−0.01**	**0.11**	0.08
FactorVAE (γ = 10)	0.44	0.63	0.27	0.08	−0.04	0.00	0.17	**0.07**
NashAE	0.56	0.17	**0.22**	0.10	**−0.07**	0.14	0.37	0.36
β-TCVAE (β = 4)	0.79	0.89	0.84	0.22	−0.04	**−0.01**	0.17	0.59
β-VAE (β = 4)	0.73	0.85	0.90	0.52	−0.05	0.00	0.28	0.53
DGAE	0.77	0.69	0.63	0.56	0.25	0.50	0.59	0.48

Averaged over the 8 datasets, LPS falls by 70% against FactorVAE, with reconstruction as good or better (CIFAR-10: −15.61 dB NMSE vs −12.49 dB).

Table 1 of the paper · bold = best in column · R^2 slightly below 0 on held-out data means nothing is predictable

RESULTS · CIFAR-10

Independence holds as the latent space grows



Flat in the latent size

From 32 to 96 dimensions the MMD model stays low, at about 0.1 or below, while most baselines grow more entangled.

Insensitive to its knobs

Batch 64 to 2048, λ from 0.1 to 5, 1 to 5 kernel bandwidths: only small changes in LPS and reconstruction.

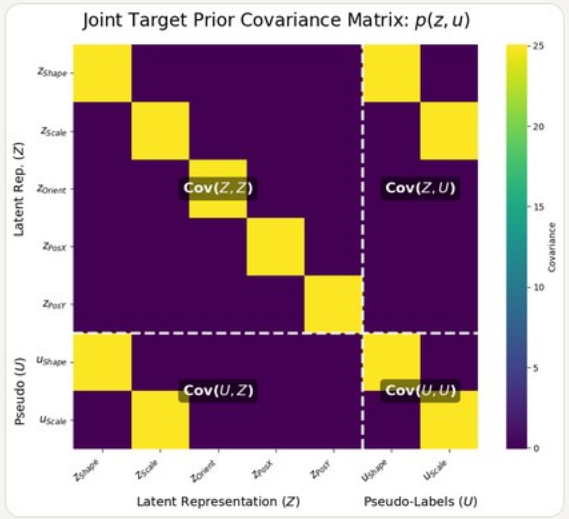
LPS with a LightGBM regressor; lower is better · ablations in the paper's appendix

PROGRAMMABLE PRIOR 1 · SEMANTIC ALIGNMENT

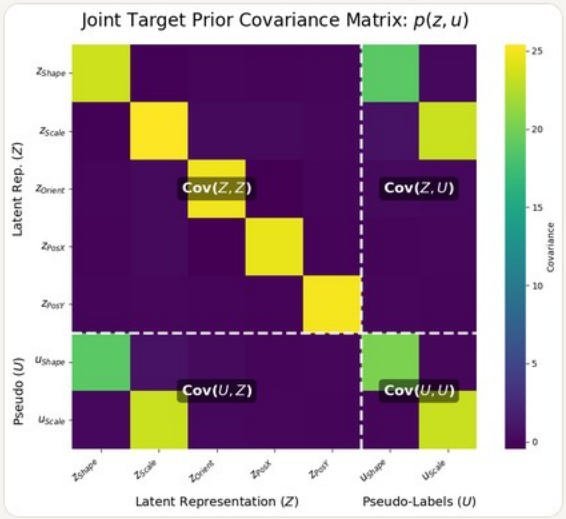
A few pseudo-labels lock latents to real factors

The prior becomes a joint over (z, f(u)) whose first k latents equal the pseudo-labels; MMD on the joint ties those latents to the labels and keeps the rest independent.

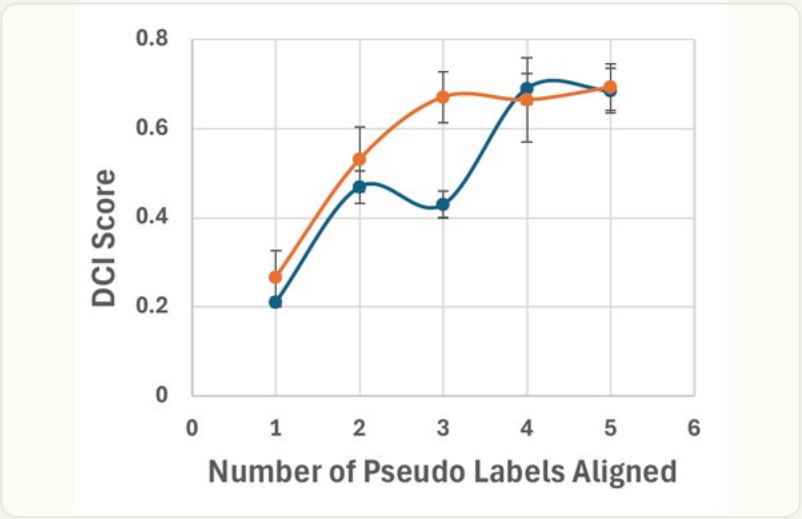
target joint



learned joint



DCI disentanglement vs pseudo-labels aligned



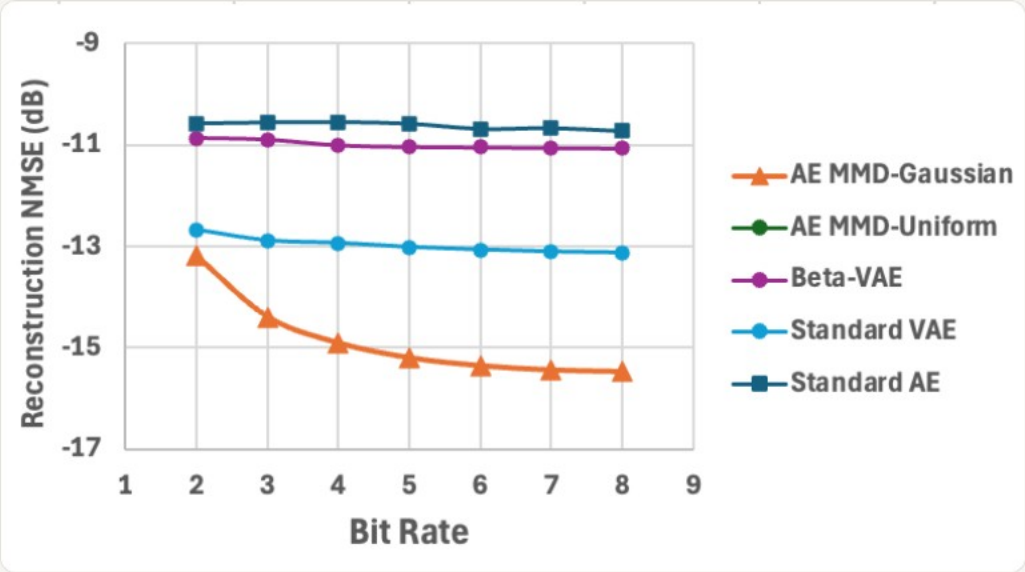
dSprites · DCI peaks at 3 of 5 factors (uniform prior, orange) and 4 of 5 (Gaussian, blue); reported Ada-GVAE and Ada-ML-VAE stay below 0.5

PROGRAMMABLE PRIOR 2 · COMPRESSION

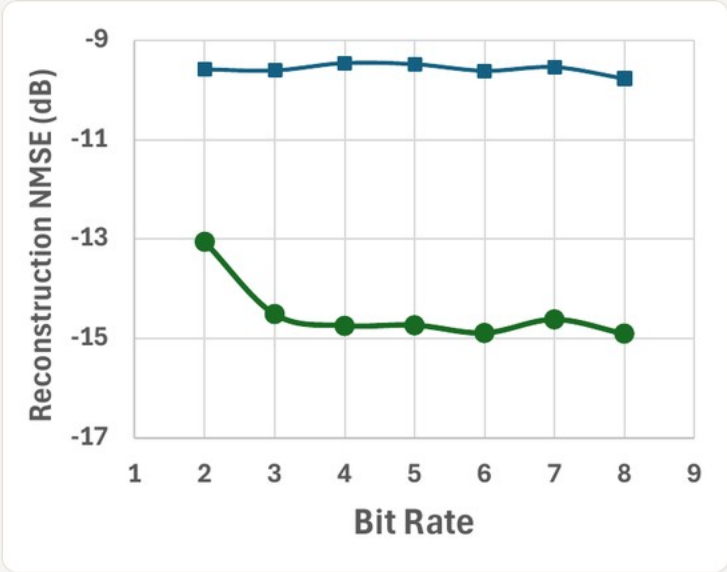
Shape the codes to fit the quantizer

Train with the quantizer in the loop and a prior equal to the distribution the quantizer assumes, Gaussian or uniform.

Gaussian quantizer · CIFAR-10



Uniform quantizer · CIFAR-10



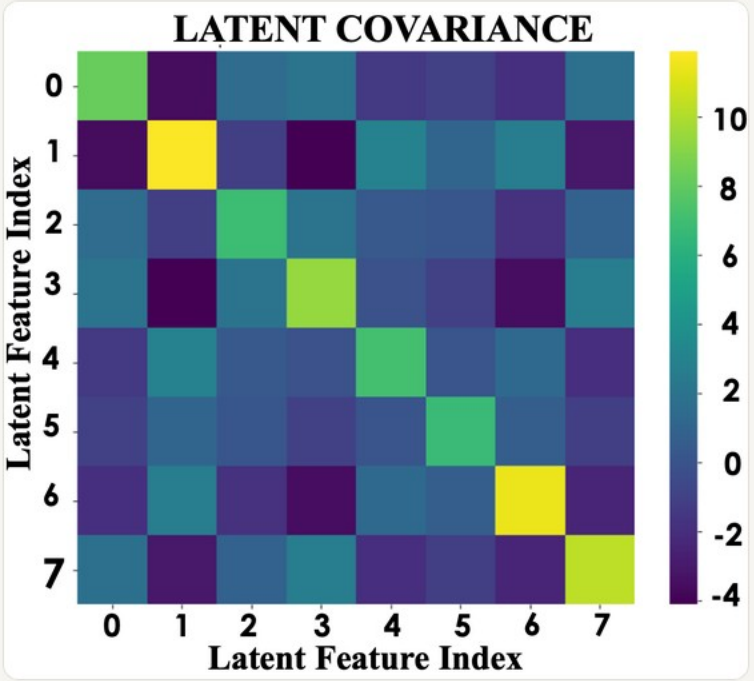
2–7 dB lower reconstruction error at 2 to 8 bits per dimension, with no post-training vector quantization or entropy coding.

Reconstruction NMSE in dB, lower is better · green: the MMD model matched to the quantizer

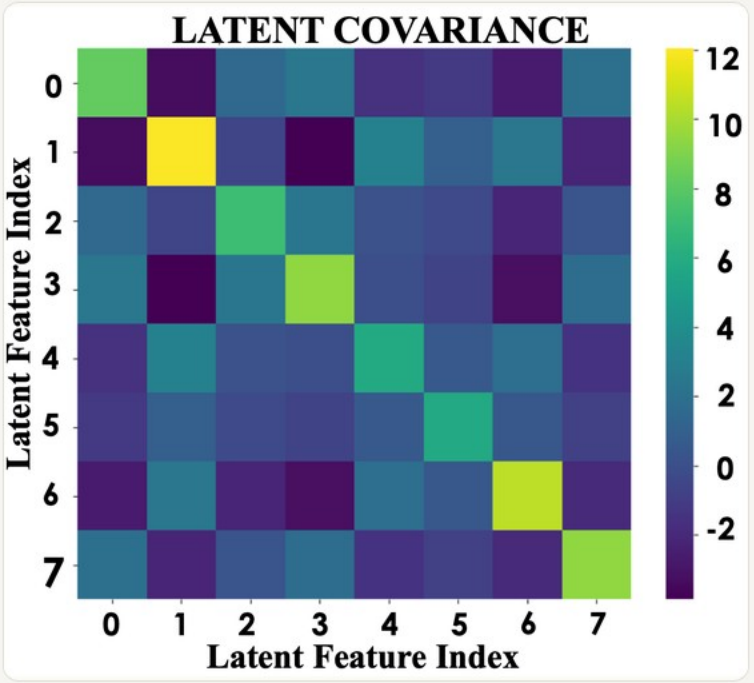
PROGRAMMABLE PRIOR 3 · ANY DISTRIBUTION YOU CAN SAMPLE

MMD can copy another model's latent space

teacher: a standard AE's codes, used as the prior



student: trained with MMD to match them



An entangled, co-dependent geometry that no closed-form KL target could express.

Covariance and marginals come out nearly identical.

MNIST, latent covariance over the dataset · paper appendix, latent copying experiment

TAKEAWAYS

The prior is an engineering tool, not a fixed assumption

A label-free test

LPS lower-bounds entanglement and exposes the VAE-style failure.

A regularizer that works

MMD on the aggregate posterior: 70% less entanglement, reconstruction kept.

One loss, many priors

Semantic alignment, quantizer matching and latent copying.

JOINT WORK WITH



Akshay Malhotra
InterDigital



Shahab Hamidi-Rad
InterDigital



Aditya Sant
InterDigital



Aryan Mokhtari
UT Austin



Sujay Sanghavi
UT Austin

arxiv.org/abs/2510.11953 · github.com/Quentin-Fruytier/Sculpting-Latent-Spaces-With-MMD

Work done at InterDigital AI Lab with the University of Texas at Austin · thank you