

Sculpting Latent Spaces With MMD: Disentanglement With Programmable Priors

Quentin Fruytier¹, Akshay Malhotra², Shahab Hamidi-Rad², Aditya Sant², Aryan Mokhtari¹,
and Sujay Sanghavi¹

¹The University of Texas at Austin

²InterDigital Communications, AI Lab

Abstract

Learning disentangled representations, where distinct factors of variation are captured by independent latent variables, is a central goal in machine learning. The dominant approach has been the Variational Autoencoder (VAE) framework, which uses a Kullback-Leibler (KL) divergence penalty to encourage the latent space to match a factorized Gaussian prior. In this work, however, we provide direct evidence that this KL-based regularizer is an unreliable mechanism, consistently failing to enforce the target distribution on the aggregate posterior. We validate this and quantify the resulting entanglement using our novel, unsupervised Latent Predictability Score (LPS). To address this failure, we introduce the Programmable Prior Framework, a method built on the Maximum Mean Discrepancy (MMD). Our framework allows practitioners to explicitly sculpt the latent space, achieving state-of-the-art mutual independence on complex datasets like CIFAR-10 and Tiny ImageNet without the common reconstruction trade-off. Furthermore, we demonstrate how this programmability can be used to engineer sophisticated priors that improve alignment with semantically meaningful features. Ultimately, our work provides a foundational tool for representation engineering, opening new avenues for model identifiability and causal reasoning.

1 Introduction

Learning meaningful representations from raw data is a cornerstone of modern machine learning. An ideal representation should capture the underlying data structure in a way that is useful for downstream tasks. A particularly promising goal is to learn *disentangled* representations, where individual latent units capture distinct, interpretable factors of variation in the data [HMPBGBML17]. Achieving this is considered a critical step towards more robust and generalizable models, ultimately enabling capabilities like compositional generalization, fairness, and causal reasoning [SLBKKB21].

The dominant paradigm, established by a sequence of seminal contributions [HMPBGBML17; KM18; CLGD18; BHPMWDL18; MRST19; LPRSBT20; KKM20; LPRSBT20; KKM20], has been to use a Variational Autoencoder (VAE) framework, which regularizes the latent space using the Kullback-Leibler (KL) divergence. However, this approach has been scrutinized for failing to achieve true disentanglement [LBLRGSB19] and for its architectural rigidity, tying the disentanglement objective to the specific structure of a VAE.

In this work, we decouple the broad objective of disentanglement into two measurable components: learning a feature distribution that is **mutually independent** over the data (as in Nonlinear

Independent Component Analysis or NICA), and ensuring that the learned features are **aligned** to semantically meaningful attributes. To realize this decoupling and address the shortcomings of prior work, we utilize a regularizer based on the architecture-agnostic and non-parametric Maximum Mean Discrepancy (MMD) [GBRSS12] to explicitly **sculpt** the latent space’s aggregate posterior distribution. Unlike parametric and analytical regularizers such as the KL divergence, the sample-based MMD approach only requires access to samples from a target prior. This allows a practitioner to easily engineer a complex and task-specific inductive bias, making the choice of the prior a **programmable** component of model design. Our primary contributions are summarized as follows:

- **Flexible and Programmable Disentanglement.** We utilize the sample based Maximum Mean Discrepancy (MMD) to enable *programmable disentanglement*. Unlike methods reliant on the KL-divergence, our approach can robustly sculpt the latent space to match **any** (mixed, dependent, or independent) target distribution (e.g., Gaussian, Uniform, or even Gaussian Mixture Models). This provides fine-grained control over the latent structure, allowing for direct enforcement of task-specific distributional properties.
- **State-of-the-Art Performance without Trade-offs.** We achieve state-of-the-art disentanglement performance **without sacrificing task-specific efficacy**. Our approach demonstrates superior factorization of latent variables across a wide range of benchmarks while maintaining high-fidelity reconstruction, effectively mitigating the common trade-off between disentanglement and reconstruction quality that plagues many VAE-based methods.
- **A Novel Unsupervised Metric for Disentanglement.** We propose a new unsupervised metric —The Latent Predictability Score (LPS)— for quantifying the mutual independence of latent features. This metric provides a robust way to evaluate the degree of factorization within a learned representation without requiring access to ground-truth factor labels, addressing a key challenge in the field.

To facilitate reproducibility and encourage further research, our complete codebase, including all experiments and our novel metric, is made publicly available.¹

2 Background

This section provides the necessary background on the generative models, theoretical challenges, and statistical tools that form the context for our work. We begin by formalizing the autoencoder and variational autoencoder paradigms before discussing the core challenges of disentanglement and introducing the Maximum Mean Discrepancy as an alternative regularizer.

Latent Variable Models and the Variational Autoencoder. In its simplest form, a latent variable model is defined by a joint distribution $p(x, z)$ of observed data x and latent variables z . In deep learning, this relationship is parameterized by a neural network, often subdivided into an encoder (or inference model) $q_{\theta_1}(z|x)$ and a decoder (or generative model) $q_{\theta_2}(x|z)$. A natural training objective is to maximize the marginal log-likelihood, $\mathbb{E}_{p(x)}[\log q_{\theta}(x)]$. Training these two components end-to-end with this objective is the basis of the classical Autoencoder (AE) paradigm.

Prior work has largely focused on frameworks that provide the ability to specify and constrain the distribution of the latent variable, $p(z)$. Re-writing the log-likelihood as $\mathbb{E}_{p(x)}[\log \mathbb{E}_{p(z)}[q_{\theta}(x|z)]]$

¹Code url: <https://github.com/Quentin-Fruytier/Sculpting-Latent-Spaces-With-MMD>

provides the most obvious objective to maximize, but direct optimization in this form is generally intractable. As a solution, the Variational Autoencoder (VAE) framework provides a popular and effective solution that consists in maximizing the Evidence Lower Bound (ELBO) of the log-likelihood:

$$\mathcal{L}_{\text{ELBO}}(x) := \underbrace{\mathbb{E}_{q_{\theta_1}(z|x)}[\log q_{\theta_2}(x|z)]}_{\text{Reconstruction Loss}} - \underbrace{D_{\text{KL}}(q_{\theta_1}(z|x)||p(z))}_{\text{Regularization Loss}} \quad (1)$$

In a standard VAE, the encoder outputs the parameters of a diagonal Gaussian, a mean vector $\mu(x)$ and a variance vector $\sigma^2(x)$. A latent code is then sampled via the reparameterization trick: $z = \mu(x) + \sigma(x) \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The regularization term becomes the per-sample KL-divergence between the approximate posterior and the prior, which is typically a standard normal distribution: $D_{\text{KL}}(N(\mu(x), \sigma(x))||N(0, I))$.

The Challenge of Unsupervised Disentanglement. In the context of disentanglement, the overarching goal of the VAE framework is to learn a representation where the latent vector z is factorized, meaning its components are statistically independent. The 'Regularization Loss' is thought to encourage this by forcing the aggregate posterior distribution towards a diagonal Gaussian structure. In fact, seminal works like β -VAE [HMPBGBML17], FactorVAE [KM18], and β -TCVAE [CLGD18] all build on this principle through distinct tweaks to the D_{KL} term. Other, but less popular alternative paradigms exist, such as using information-theoretic objectives in GANs [CDHSSA16] or imposing direct geometric constraints like the learning of orthogonal group action features [CT23].

However, achieving interpretable disentanglement where the learned features are semantically meaningful on top of being mutually independent is complicated by deep theoretical and practical challenges. The theoretical issue is the unidentifiability of Nonlinear Independent Component Analysis (NICA) [HP99; KKM20], which posits that without a strong inductive bias, an infinite number of valid independent representations can be learned from the same data [LBLRGSB19; LPRSB20].

The Maximum Mean Discrepancy. An alternative to the KL-divergence for estimating the distance between two distributions is the Maximum Mean Discrepancy (MMD) [GBRSS12]. The core intuition of MMD is to represent distributions as single mean embeddings in a high-dimensional Reproducing Kernel Hilbert Space (RKHS). The MMD between two distributions, P and Q , is then simply the RKHS norm of the difference between their mean embeddings, μ_P and μ_Q :

$$\text{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}} \quad (2)$$

If the mean embeddings are identical, the MMD is zero, and the distributions are identical. This is operationalized via a kernel function, $k(\cdot, \cdot)$, such as a Gaussian RBF kernel.

In practice, finite set of samples $Z = \{z_1, \dots, z_m\} \sim P$ and $Z' = \{z'_1, \dots, z'_n\} \sim Q$ are used to estimate the mean embeddings. The MMD is then the distance between these empirical estimates:

$$\text{MMD}(P, Q) \approx \left\| \frac{1}{m} \sum_{i=1}^m \phi(z_i) - \frac{1}{n} \sum_{j=1}^n \phi(z'_j) \right\|_{\mathcal{H}} \quad (3)$$

where ϕ is the feature map into the RKHS associated with the kernel k . Using the kernel trick, $\langle \phi(a), \phi(b) \rangle_{\mathcal{H}} = k(a, b)$, we can expand this into a computable form. The standard unbiased empirical estimator for the squared MMD is given by the sum of three terms:

$$\text{MMD}^2(P, Q) = \frac{1}{m(m-1)} \sum_{i \neq j} k(z_i, z_j) - \frac{2}{mn} \sum_{i,j} k(z_i, z'_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(z'_i, z'_j) \quad (4)$$

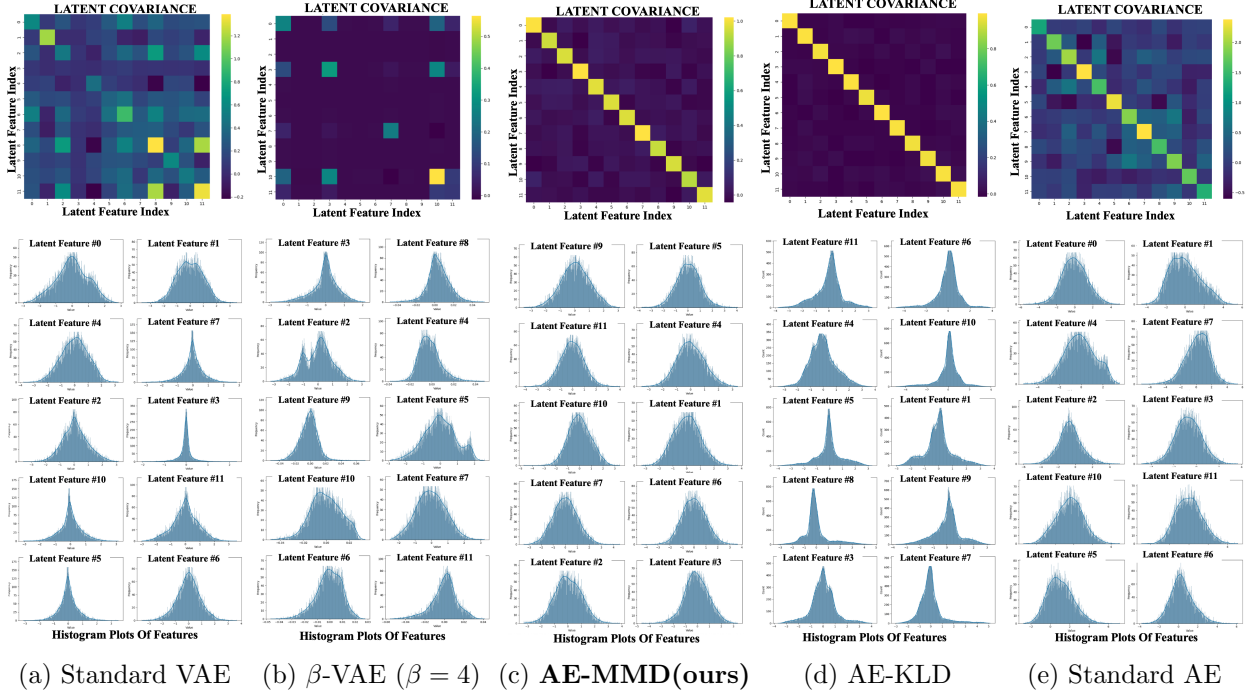


Figure 1: **Visual comparison of learned latent distributions on MNIST for key baselines.** Each column corresponds to a different model, showing the covariance matrix between the latent space features over the dataset (top row) and the histogram plots for the marginal distribution of the latent space features over the whole dataset (bottom row). The VAE-based models (a, b) and the standard AE (e) fail to enforce either independence or the target Gaussian geometry. While a batch-wise KLD regularizer (d) achieves a diagonal covariance, its parametric nature prevents it from correctly shaping the marginal distributions leading to co-dependence (LPS equal to 0.97). Only our MMD-regularized model (c) successfully enforces both properties, producing a representation that is truly independent and matches the target prior.

The principle of using an MMD penalty as a more stable substitute for other divergence function in Neural Networks is a well-established concept, primarily explored for generative modeling, domain adaptation, and representational knowledge distillation [TBGS17; LCCYP17; LZWJ17; HW17; ZSE19]. While our work leverages a similar core mechanism as that of the Wasserstein Autoencoder [TBGS17], which explored MMD as a more stable divergence in the context of generative modeling, it is fundamentally distinguished by its objective. In the section that follows, we re-purpose this tool for a different goal entirely: **programmable disentanglement**. Our framework shifts the focus to **representational engineering**, using the MMD regularizer as a precise instrument to inject a wide range of explicit, user-defined inductive biases.

3 The MMD Regularizer for Sculpting the Latent Space

Our core technical contribution is the formulation of an MMD-based framework visualized in Figure 2 that can be integrated into any neural network layer where it is desired for the aggregate posterior, $z \sim q_{\theta_1}(z) = \mathbb{E}_{p(x)}[q_{\theta_1}(z|x)]$, to match a prior distribution $p(z)$. With this in mind, our framework

proposes to maximize the following lower bound on the log-likelihood,

$$\mathcal{L}_{\text{ours}} := \underbrace{\mathbb{E}_{p(x)}[\log q_{\theta}(x)]}_{\text{AE Loss}} - \lambda \cdot \underbrace{\text{MMD}^2(q_{\theta_1}(z), p(z))}_{\text{MMD Regularization}}, \quad (5)$$

where λ is a hyperparameter that controls the strength of the regularization. Unlike the Variational Autoencoder framework where $z|x$ is necessarily probabilistic, our framework does not limit the choice of $z|x$ to be either probabilistic or deterministic.

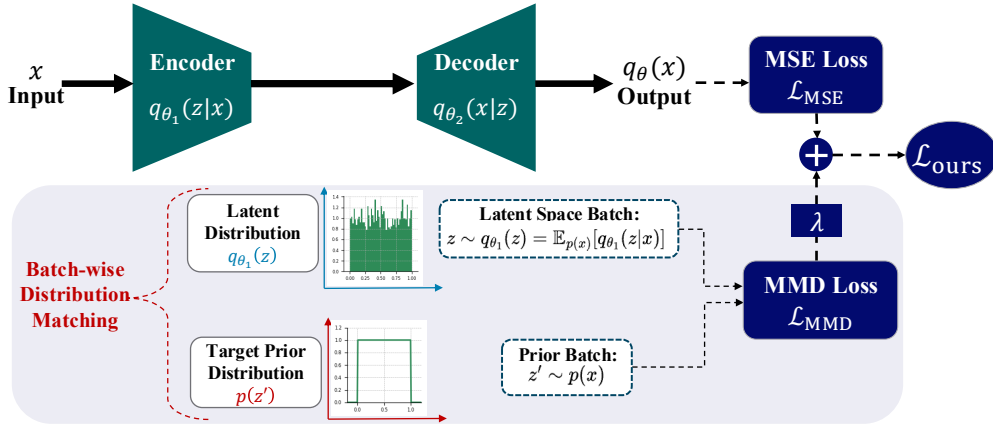


Figure 2: Visual Representation of the Programmable Prior Framework.

3.1 Disentanglement and the Programmable Prior

A central advantage of our framework is the ability to treat the choice of the target prior, $p(z)$, as a powerful and explicit **inductive bias**. This elevates the prior from a fixed, implicit assumption to a flexible, user-defined component for engineering the latent space. By programming the prior, we can sculpt the aggregate posterior’s geometry to have specific, desirable properties—from simple factorized distributions (e.g., Gaussian, Uniform) to priors with complex, known dependencies.

The Failure of KL-based Regularization. The goal of disentanglement requires enforcing a specific structure (e.g., a factorized Gaussian) on the **aggregate posterior** distribution of the latent codes. The VAE framework attempts this with a per-sample KL penalty, $D_{KL}(q_{\theta_1}(z|x)||p(z))$. However, this is an unreliable mechanism that often fails to enforce the desired properties on the aggregate distribution of the representation $\mu(x)$. As shown in Figure 1 (a-b), standard VAE-based models fail to produce representations that are both statistically independent (i.e., have a diagonal covariance matrix) and match the target Gaussian geometry over the data.

A more direct approach is to apply the KL divergence to the batch statistics of the encoded features, using a loss term like $D_{KL}(\mathcal{N}(\mu_z, \Sigma_z)||\mathcal{N}(0, I))$, where μ_z and Σ_z are the empirical mean and covariance. This is equivalent to our framework in (5) in the case that the MMD regularization is replaced with KLD regularization. While this correctly targets the aggregate distribution, its parametric nature is a critical flaw. As shown in Figure 1(d), this batch-wise KL regularizer successfully enforces a diagonal covariance but fails to make the marginals Gaussian leading to consequential entanglement. In fact, we report an LPS score of 0.97 for this method, indicating the features are completely co-dependent over the data. This demonstrates that merely matching the first two moments is insufficient for enforcing true statistical independence.

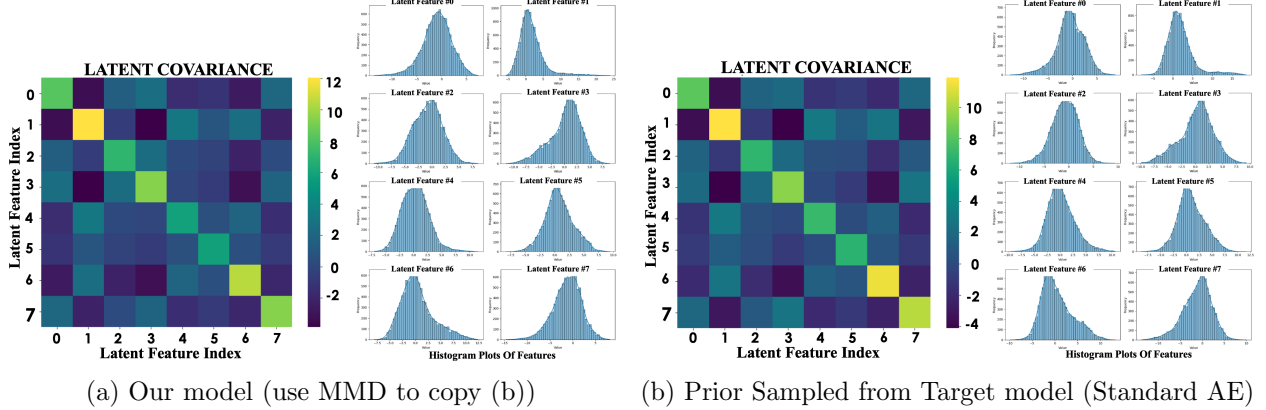


Figure 3: **Visualizing the Latent Space Copying experiment on MNIST.** (b) First, a standard Autoencoder was trained, and its complex, entangled latent distribution was saved to serve as an empirical prior. (a) Then, a new "student" model was trained using our MMD regularizer, tasked with replicating this arbitrary prior. The figure displays the covariance matrix between the latent space features over the whole dataset and the histogram plots of the marginal distribution of the latent space features over the whole dataset. The near-identical covariance matrices and marginal distributions visually confirm our method's ability to enforce a complex, co-dependent latent geometry that would be impossible to specify analytically with KLD.

The MMD Solution For Sculpting The Latent Space. This motivates the need for a non-parametric, sample-based regularizer. Switching from an analytical penalty like KLD to the MMD allows for far more flexible priors. Because MMD operates on samples, the target distribution does not need to be analytically defined; we only need a way to sample from it. This allows us to use fixed priors, like a Gaussian, or even adaptive priors derived from data. As shown in Figure 1(c), our MMD-based approach is the only method that successfully enforces both a perfectly diagonal covariance matrix and perfectly Gaussian marginal distributions, matching the entire distribution, not just its moments.

The true flexibility of our MMD framework is best demonstrated by its ability to enforce priors with complex, structured dependencies. To showcase this, we designed an experiment to push the limits of MMD regularization. First, we trained an unconstrained, standard Autoencoder and saved the dataset of its learned latent vectors. This collection of latent codes—with its complex, co-dependent structure—served as the target prior for a new model trained with our framework. This experiment serves as a demonstration of MMD's power to enforce a prior so intricate it cannot be written analytically for a KLD-based regularizer. The results in Figure 3 visually confirm that our method can convincingly copy this arbitrary, entangled latent geometry with high fidelity, highlighting the precision of MMD as a tool for representational engineering.

4 Experiments and Results

In this section, we first motivate and formulate our new Latent Predictability Score (LPS). We then provide a comprehensive overview of all our quantitative and qualitative experiments and results.

4.1 A Novel Unsupervised Metric for Latent Independence

Motivation. A significant limitation of popular disentanglement metrics like the mutual information gap (MIG) [CLGD18] or the Disentanglement Completeness Information score (DCI) [EW18] is their reliance on access to the ground-truth generative factors for evaluation. Such labels are unavailable for most real-world datasets, creating a critical gap for a metric that can directly quantify the quality of a learned representation in a truly unsupervised manner. Moreover, several works at the interface of NICA and disentanglement have highlighted that semantically meaningful and disentangled representations are not unique which considerably degrades the reliability of any supervised metric [LBLRGSB19; LPRSBT20; KKM20]. To address these challenges, we propose an unsupervised metric that directly measures the level of mutual independence of the learned latent features over the data. While our metric cannot gauge the level of alignment to semantically meaningful features, it enables us to evaluate the extent to which the learned features are distinct from all others.

Introducing the Latent Predictability Score (LPS). The core intuition is simple: if a d dimensional latent space, with latent features $\{z_1, \dots, z_d\}$ are truly independent, then any given feature z_i should not be predictable from the others, $z_{j \neq i}$. The LPS procedure quantifies this intuition as follows:

1. A representative set of latent vectors is obtained by encoding a large sample of data.
2. For each latent dimension i from 1 to d :
 - (a) The feature z_i is treated as a target label.
 - (b) A regression model is trained to predict z_i using the remaining $d - 1$ dimensions, $z_{j \neq i}$.
 - (c) The quality of this prediction is evaluated on a withheld test set using the coefficient of determination (R^2) score.
3. The final LPS is the average R^2 score across all d dimensions.

Interpretation. The final score is a direct measure of latent dependency. A higher score implies that the features are entangled and predictable from one another, while a score approaching zero indicates a high degree of mutual independence. Therefore, a **lower LPS score is better**. See Appendix C for a discussion on the LPS metric and how it compares to more classic supervised metrics.

Variants. To ensure the predictability measure is robust, we employ two powerful and distinct regression models for this task. We therefore report two versions of our metric: **LPS-mlp**, which uses a Multi-Layer Perceptron, and **LPS-lgbm**, which uses the highly effective LightGBM gradient boosting framework [KMF20].

4.2 Experimental Setup

Datasets. We evaluate our method across a diverse suite of benchmarks to test its scalability and effectiveness. We use synthetic datasets where the ground-truth factors are known: **dSprites** [MH17], and the simple generative environments **XY**, **XYC**, and **XYCS** introduced in [CT23], which consist of circles with varying position (XY), color intensity (C), and shape (S). To test performance on complex, real-world data, we use **MNIST** [Den12], **CIFAR-10** [KH09], and **TinyImageNet** [LY15].

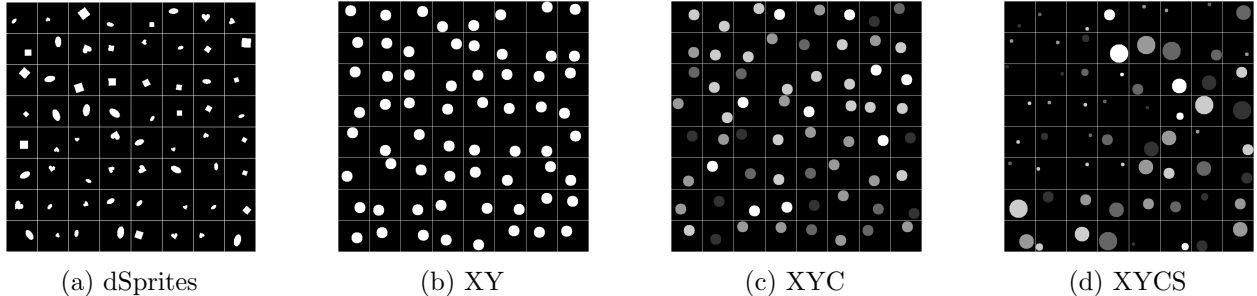


Figure 4: Example images (represented in an 8×8 grid) from the synthetic datasets used in our experiments. From left to right: (a) dSprites, with factors like shape, scale, and orientation; (b-d) The XY family of datasets, with factors for position (XY), color (C), and shape (S).

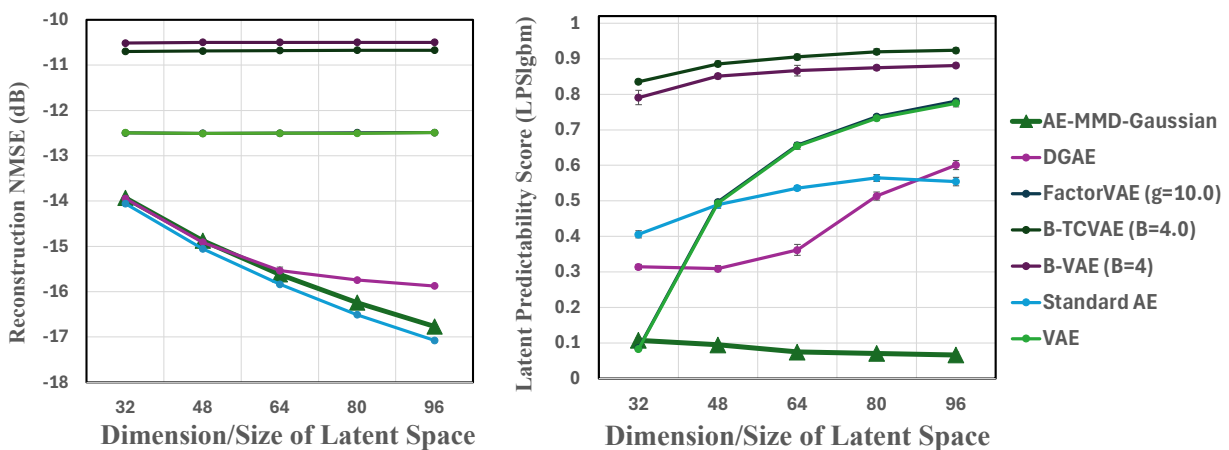


Figure 5: **CIFAR10: Reconstruction and LPS Against Latent Space Size.** Line plot showing (left) the reconstruction NMSE in decibels of all baselines and (right) the mutual independence of the latent space features (estimated with our Latent Predictability Score (LPS)) against the dimension of the latent space.

Baselines. We compare our MMD-based approach against several established and state-of-the-art methods for unsupervised disentanglement. Our primary baselines are β -VAE [HMPBGBML17], β -TCVAE [CLGD18], FactorVAE [KM18], and DGAE [CT23]. We also include a standard **Autoencoder (AE)** with no explicit disentanglement regularizer and standard **VAE** as control variables to establish a performance floor.

While our reported experiments used a batch size of 512 and a regularization coefficient of $\lambda = 0.3$, the MMD regularizer proved robust, yielding consistent results across a wide range of these hyperparameters. The implementation and experimental details can be found in Appendix A and B.

4.3 State-of-the-Art Mutual Independence

Our quantitative analysis reveals that our MMD-based regularizer achieves a new state-of-the-art in enforcing latent statistical independence, a claim supported across a wide range of datasets and model configurations (we provide the complete and comprehensive experiment results and

set-up in Appendix A and Appendix B). The condensed results in Table 1 demonstrate that our method consistently yields the lowest Latent Predictability Scores (LPS), signifying a higher degree of mutual independence, with little to no degradation in reconstruction quality. To contextualize these scores, the LPS measures the proportion of a latent feature’s variance predictable from its peers using the coefficient of determination (R^2). For instance, on dSprites, the popular β -VAE achieves an LPS of 0.52 implying substantial entanglement; in contrast, our score of 0.04 signifies a representation approaching true independence.

Table 1: Main quantitative results comparing our MMD-based Autoencoder against baselines (the latent space dimension is specified next to the dataset name). Each cell presents two key metrics: **LPS-lgbm** (R^2) — *reconstruction* **NMSE** (dB) where lower values are better (\downarrow). The best performance for each metric on each dataset is highlighted in **bold**.

Model	ImageNet ($d = 64$)		CIFAR10 ($d = 64$)		MNIST ($d = 12$)		dsprites ($d = 5$)		XY ($d = 2$)		XYC ($d = 3$)		XYCS ($d = 4$)	
	LPS	NMSE (dB)	LPS	NMSE (dB)	LPS	NMSE (dB)	LPS	NMSE (dB)	LPS	NMSE (dB)	LPS	NMSE (dB)	LPS	NMSE (dB)
AE-MMD (Ours)	0.03	-11.70	0.04	-15.61	0.22	-11.92	0.04	-17.89	-0.05	-15.22	-0.01	-16.31	0.11	-16.76
DGAE	0.77	-10.3	0.69	-15.53	0.63	-11.48	0.56	-14.61	0.25	-6.74	0.50	-10.94	0.59	-15.02
FactorVAE ($\gamma=10.0$)	0.44	-10.75	0.63	-12.49	0.27	-8.62	0.08	-14.56	-0.04	-15.36	0.00	-15.92	0.17	-15.52
β -TCVAE ($\beta=4.0$)	0.79	-9.80	0.89	-10.68	0.84	-5.97	0.22	-8.83	-0.04	-14.68	-0.01	-13.72	0.17	-12.40
β -VAE ($\beta=4$)	0.73	-9.87	0.85	-10.49	0.90	-4.61	0.52	-8.20	-0.05	-14.46	0.00	-13.33	0.28	-10.98
Standard AE	0.46	-11.77	0.51	-15.83	0.54	-12.09	0.19	-18.74	-0.03	-16.90	0.04	-17.81	0.26	-18.42
Standard VAE	0.47	-10.61	0.63	-12.50	0.25	-8.63	0.10	-14.66	-0.05	-15.72	0.02	-16.04	0.18	-15.53

Furthermore, our framework’s ability to enforce this independence is uniquely robust to the dimensionality of the latent space. As shown in Figure 5, while most baseline methods learn increasingly entangled representations as the latent space grows, our MMD regularizer maintains a consistently low LPS. This scalability is a key advantage, demonstrating that our approach provides a reliable mechanism for structuring latent spaces without being constrained by model capacity, a common failure mode for VAE-based regularizers.

4.4 Alignment to Interpretable Features with the Programmable Prior

While achieving mutual independence is a primary objective, the ultimate goal is to align these independent features with interpretable, real-world concepts. Our framework’s “programmable prior” provides a powerful mechanism for injecting the necessary inductive bias to achieve this alignment. By engineering the prior to match the known or hypothesized structure of a dataset’s generative factors, we can guide the model toward learning not just an independent representation, but a semantically meaningful one.

Figure 6 provides a clear demonstration on the XY dataset (we provide additional results for the other synthetic datasets in Appendix D), where the ground-truth factor (position) follows a uniform distribution. When we program our prior to be uniform, our model successfully sculpts the latent marginals to match this target geometry. This is in stark contrast to a standard autoencoder, which defaults to a uni-modal representation that fails to capture the true data structure. The benefit of this correctly structured latent space is evident in the latent traversals: our model’s traversals are clean and disentangled, while the standard AE’s are entangled and less interpretable.

The quantitative results in Table 2 confirm that this improved structure directly translates to better alignment. On the XY dataset, using a standard factorized Gaussian prior achieves excellent statistical independence (LPS of -0.05) but yields a modest alignment score (DCI of 0.199). However, by simply switching to a Uniform prior that mirrors the true generative process, we dramatically improve the DCI and MIG scores, achieving near-perfect alignment without sacrificing independence. This result underscores a key contribution of our work: the prior is a critical tool for representation

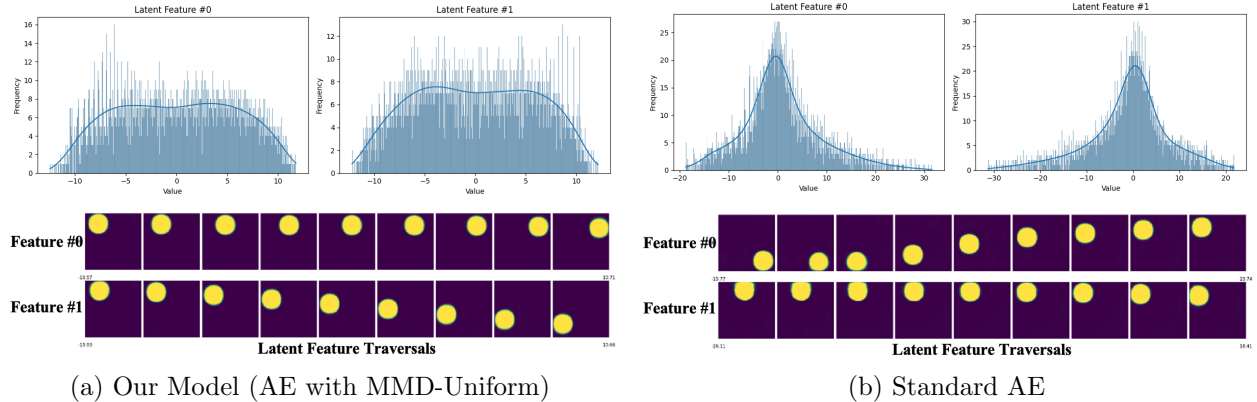


Figure 6: **Qualitative Comparison on the XY Dataset.** Each column compares the learned latent marginals (top) with the corresponding latent traversals (bottom). **(a)** Our MMD-trained model successfully learns the true uniform distribution, which results in clean, disentangled traversals. **(b)** In contrast, the standard AE defaults to a uni-modal distribution, leading to entangled traversals.

engineering, and our framework is unique in its ability to program this bias to achieve interpretable alignment.

Table 2: Results for our MMD-based Autoencoder on the XY dataset with different engineered prior distributions. We find selecting a Uniform target prior significantly improve alignment metrics.

Prior	NMSE (dB)	LPS-lgbm	LPS-mlp	DCI-RF	MIG
AE-Gaussian	-15.2	-0.053	-0.004	0.199	0.751
AE-Uniform	-15.8	-0.064	-0.003	0.469	1.047
AE-GMM (2 component)	-14.9	-0.051	0.000	0.251	0.863

5 Discussion and Conclusion

5.1 Programmable Bias as a New Form of Feature Engineering

The results presented in this paper advocate for a shift in how we approach the construction of learned representations. Our MMD-based framework introduces a new paradigm: **representational feature engineering**. By making the inductive bias of the latent space an explicit, programmable component, we move beyond simply hoping a model learns a good representation and towards directly instructing it on the *kind* of representation it should learn.

This capability allows practitioners to inject high-level domain knowledge directly into the model’s internal geometry. This transforms model design into a more deliberate and controllable process. Practitioners can now explicitly sculpt the latent space to align with theoretical assumptions or empirical observations about the data, paving the way for more robust, interpretable, and tailored models.

5.2 Limitations and Future Work

Despite its successes, our framework has limitations that open promising avenues for future research. The primary challenge shifts from *how* to enforce a prior to *which* prior to enforce. For complex, real-world data, the true generative process is unknown, and selecting an optimal prior remains an open question. While our work demonstrates that even a simple factorized Gaussian is highly effective for achieving statistical independence, engineering a prior for optimal alignment requires domain knowledge that may not always be available. Additionally, while the MMD estimator is robust, fitting complex high-dimensional priors can require careful tuning of the kernel functions.

These limitations point to several exciting future directions:

- **Learning the Prior:** A natural extension of this work is to learn the optimal prior directly from data, perhaps using an auxiliary network or an adaptive mechanism that evolves the prior during training.
- **Identifiability and Causal Representation Learning:** Our demonstration of “latent space copying” (Figure 3) provides a powerful tool for future work in identifiability. Extending our framework to match joint distributions of the aggregate posterior with observable variables of the input offers a practical method for satisfying the theoretical conditions for a unique, disentangled solution as outlined by Khemakhem, Kingma, Monti, and Hyvarinen [KKMH20], a critical step towards causal representation learning [SLBKKG21].
- **Representational Knowledge Distillation:** The ability to precisely copy a latent distribution is a powerful tool for knowledge distillation Huang and Wang [HW17]. However, we report in Appendix E.2 that while the marginal distribution of latents are the same, the learned representations are still fundamentally different, indicating that matching the aggregate posterior only is not enough to learn the same representation.

5.3 Conclusion

In this work, we challenged the foundational assumptions of the dominant VAE-based paradigm for disentanglement. We provided direct visual and quantitative evidence that the KL-divergence is a flawed and unreliable mechanism for enforcing statistical independence. As a solution, we introduced a powerful, architecture-agnostic regularizer based on the Maximum Mean Discrepancy. Our framework replaces the weak, implicit bias of the VAE with a direct enforcement mechanism, enabling the novel concept of a “programmable prior.”

Our experiments demonstrate that this approach achieves a new state-of-the-art in statistical independence on complex, real-world datasets, as measured by our stable and robust Latent Predictability Score (LPS). Furthermore, we showed that by engineering the prior, we can inject the necessary inductive bias to achieve meaningful alignment with ground-truth factors. Ultimately, our work provides a foundational tool for representation engineering, opening new avenues for research in model identifiability, representational knowledge distillation, and the pursuit of causal representation learning.

References

- [BHPMWDL18] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. “Understanding disentangling in β -VAE”. In: *arXiv preprint arXiv:1804.03599* (2018) (page 1).
- [CT23] J. Cha and J. Thiyagalingam. “Orthogonality-enforced latent space in autoencoders: An approach to learning disentangled representations”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 3913–3948 (pages 3, 7, 8, 15).
- [CLGD18] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. “Isolating sources of disentanglement in variational autoencoders”. In: *Advances in neural information processing systems* 31 (2018) (pages 1, 3, 7, 8, 15, 16).
- [CDHSSA16] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets”. In: *Advances in neural information processing systems* 29 (2016) (page 3).
- [Den12] L. Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142 (pages 7, 15).
- [EW18] C. Eastwood and C. K. Williams. “A framework for the quantitative evaluation of disentangled representations”. In: *6th International Conference on Learning Representations*. 2018 (pages 7, 16).
- [GBRSS12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. “A kernel two-sample test”. In: *The journal of machine learning research* 13.1 (2012), pp. 723–773 (pages 2, 3, 17).
- [HMPBGBML17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: *International conference on learning representations*. 2017 (pages 1, 3, 8, 15).
- [HW17] Z. Huang and N. Wang. “Like what you like: Knowledge distill via neuron selectivity transfer”. In: *arXiv preprint arXiv:1707.01219* (2017) (pages 4, 11, 29).
- [HP99] A. Hyvärinen and P. Pajunen. “Nonlinear independent component analysis: Existence and uniqueness results”. In: *Neural Networks* 12.3 (1999), pp. 429–439. ISSN: 0893-6080. DOI: [https://doi.org/10.1016/S0893-6080\(98\)00140-3](https://doi.org/10.1016/S0893-6080(98)00140-3). URL: <https://www.sciencedirect.com/science/article/pii/S0893608098001403> (page 3).
- [KMFWCMYL17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. “Lightgbm: A highly efficient gradient boosting decision tree”. In: *Advances in neural information processing systems* 30 (2017) (page 7).
- [KKMH20] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. “Variational autoencoders and nonlinear ica: A unifying framework”. In: *International conference on artificial intelligence and statistics*. PMLR. 2020, pp. 2207–2217 (pages 1, 3, 7, 11).

- [KM18] H. Kim and A. Mnih. “Disentangling by factorising”. In: *International conference on machine learning*. PMLR. 2018, pp. 2649–2658 (pages 1, 3, 8, 15).
- [KH+09] A. Krizhevsky, G. Hinton, et al. “Learning multiple layers of features from tiny images”. In: (2009) (pages 7, 15).
- [LY15] Y. Le and X. Yang. “Tiny imagenet visual recognition challenge”. In: *CS 231N* 7.7 (2015), p. 3 (pages 7, 15).
- [LCCYP17] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. “Mmd gan: Towards deeper understanding of moment matching network”. In: *Advances in neural information processing systems* 30 (2017) (page 4).
- [LBLRGSB19] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124 (pages 1, 3, 7, 19, 21).
- [LPRSBT20] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. “Weakly-supervised disentanglement without compromises”. In: *International conference on machine learning*. PMLR. 2020, pp. 6348–6359 (pages 1, 3, 7).
- [LZWJ17] M. Long, H. Zhu, J. Wang, and M. I. Jordan. “Deep transfer learning with joint adaptation networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 2208–2217 (page 4).
- [MRST19] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh. “Disentangling disentanglement in variational autoencoders”. In: *International conference on machine learning*. PMLR. 2019, pp. 4402–4412 (page 1).
- [MHHL17] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. *dSprites: Disentanglement testing Sprites dataset*. <https://github.com/deepmind/dsprites-dataset/>. 2017 (pages 7, 15).
- [SLBKKGB21] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. “Toward causal representation learning”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634 (pages 1, 11).
- [TBGS17] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. “Wasserstein autoencoders”. In: *arXiv preprint arXiv:1711.01558* (2017) (pages 4, 27).
- [ZSE19] S. Zhao, J. Song, and S. Ermon. “Infovae: Balancing learning and inference in variational autoencoders”. In: *Proceedings of the aaai conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 5885–5892 (page 4).

Appendix

LLM Use Acknowledgment

We acknowledge the use of Google’s Gemini Pro 2.5 in the preparation of this manuscript. The model was utilized to aid in polishing grammar and clarity, assist with generating formatting for tables and figures, and serve as an assistant for the retrieval of prior work. The authors meticulously reviewed and edited all generated content and take full responsibility for the accuracy and originality of this work.

Table Of Contents

A	Experiment And Implementation Details	15
A.1	Experimental Design	15
A.2	Evaluation Protocol and Metrics	16
A.3	MMD Regularizer Implementation	16
B	Comprehensive Quantitative Experimental Results	18
C	LPS Metric Analysis	19
C.1	The Alignment-Independence Paradox	19
C.2	Metric Instability Across Runs	20
C.3	LPS as a Stable, Intrinsic Measure	21
D	The Power of Programmable Priors: Extended Results	23
D.1	dSprites: Engineering a Mixed-Distribution Prior	23
D.2	XY: Engineering a Uniform Prior	24
D.3	XYC: Engineering a GMM-Heavy Prior	25
D.4	XYCS: Engineering a Mixed-Component Prior	25
E	Additional Experiments	28
E.1	Robustness of MMD to Latent Space Size	28
E.2	MMD for Representational Knowledge Transfer	29

A Experiment and Implementation Details

A.1 Experimental Design

A.1.1 Datasets

Our method was evaluated on a diverse suite of benchmarks to assess its scalability and effectiveness.

- **Synthetic Datasets:** For controlled experiments with known ground-truth factors, we used **dSprites** [MHHL17] and the **XY**, **XYC**, and **XYCS** environments [CT23]. These datasets feature simple shapes with explicit factors of variation, such as position (XY), color (C), and shape (S).
- **Real-World Datasets:** To test performance on complex, high-dimensional data, we used the standard image classification benchmarks **MNIST** [Den12], **CIFAR-10** [KH+09], and **Tiny ImageNet** [LY15].

A.1.2 Baseline Models

We benchmarked our MMD-based approach against established methods for unsupervised disentanglement, including β -VAE [HMPBGBML17], β -TCVAE [CLGD18], **FactorVAE** [KM18], and **DGAE** [CT23]. We also included a standard **Autoencoder (AE)** and a standard **VAE** as controls to establish a performance baseline without explicit disentanglement pressures.

A.1.3 Training Details

All models were trained using the Adam optimizer with a learning rate of 5×10^{-4} and a batch size of 512. Our MMD-based regularizer used a regularization coefficient of $\lambda = 0.3$. We additionally found that our framework was not very sensitive to variation in this hyperparameter as values between 0.1 and 2 resulted in comparable performance. The models are fully convolutional, featuring symmetric encoder-decoder architectures. Specific depths and latent dimensions for each dataset are detailed in Table 3.

Table 3: Dataset-specific hyperparameters. The "Number of Layers" refers to the depth of the encoder and decoder individually (e.g., "6" indicates 6 layers for the encoder and 6 for the decoder).

Dataset	Number of Layers	Latent Dimension (d)	Training Epochs
XY, XYC, XYCS	6, 6, 6	2, 3, 4	500, 250, 125
dSprites	6	5	50
MNIST	4	12	250
CIFAR-10	4	64	250
Tiny ImageNet	6	64	75

A.2 Evaluation Protocol and Metrics

To ensure robust and reproducible results, all reported scores are the mean and standard deviation over 5 independent runs with different random seeds.

Metric Descriptions

NMSE (dB / linear). The Normalized Mean Squared Error measures reconstruction quality, comparing the output of the decoder to the original input image. We report it on both a linear scale and a decibel (dB) scale, which is logarithmic. For the dB scale, more negative values indicate better reconstruction.

LPS (LGBM / MLP). Our proposed Latent Predictability Score measures the mutual independence of the learned latent features. It is based on the average coefficient of determination (R^2) from a regression task that attempts to predict each latent dimension from all others. A score near zero indicates a high degree of independence. Therefore, a **lower LPS score is better**. We report two variants using LightGBM and a Multi-Layer Perceptron as the regressors.

Covariance Ratio. This metric quantifies the degree of *linear* feature independence by measuring how diagonal the latent covariance matrix is. It is calculated as the ratio of the sum of the diagonal elements to the sum of the absolute values of the off-diagonal elements. A **higher ratio indicates a more diagonal covariance matrix** and thus greater **linear** statistical independence.

MIG (Mutual Information Gap). This metric measures the degree to which each latent variable is informative about one, and only one, ground-truth factor. It is calculated as the normalized difference between the highest and second-highest mutual information over all factors [CLGD18]. A **higher MIG score is better**.

DCI (Disentanglement, Completeness, Informativeness) [EW18]. This framework probes the latent space with a Random Forest regressor and reports three scores:

- **Disentanglement:** Measures if each latent dimension captures at most one ground-truth factor. A higher score is better.
- **Completeness:** Measures if each ground-truth factor is captured by a single latent dimension. A higher score is better.
- **Informativeness:** Measures the overall accuracy in predicting ground-truth factors from the latent representation. A higher score is better.

SAP (Separated Attribute Predictability). This metric measures the extent to which each latent dimension is predictive of only a single ground-truth factor. For each factor, it computes the difference in prediction error between a model trained on a single latent dimension and a model trained on all other dimensions. A **higher SAP score is better**.

A.3 MMD Regularizer Implementation

Our MMD regularizer is designed to be robust and adaptive. The following provides specific implementation details.

Kernel Configuration We employ a mixture of Gaussian Radial Basis Function (RBF) kernels for all MMD calculations. The RBF kernel is defined as $k(z_i, z_j) = \exp(-\|z_i - z_j\|^2 / (2\sigma^2))$, where σ is the bandwidth. Using a sum of kernels with different bandwidths (generated from a base bandwidth with multipliers $[0.5, 1.0, 2.0]$) allows the MMD to capture distributional discrepancies across multiple scales, leading to a more robust distance metric.

Adaptive Bandwidth Selection Instead of fixing the kernel bandwidth σ , we use the **median heuristic** [GBRSS12]. At each training step, the base bandwidth σ_{base} is set to the median of all pairwise distances between the points in the combined set of latent samples $z \sim q_{\theta_1}(z)$ and prior samples $z' \sim p(z)$. This adaptive approach automatically scales the kernel to the current batch, eliminating the need for manual hyperparameter tuning.

MMD Estimator At each training step, we draw a batch of samples from the target prior $p(z)$ of the same size as the batch of encoded latent samples from the aggregate posterior $q_{\theta_1}(z)$. The MMD loss is then computed using the standard **biased, quadratic-time estimator** of the squared MMD. Despite its $O(n^2)$ complexity, this estimator is computationally efficient for typical batch sizes and is known to be effective and robust in practice.

B Comprehensive Experimental Results

This appendix provides the complete, unabridged quantitative results of our experiments for all models, and datasets. All reported scores are the mean and standard deviation computed over 5 independent runs with different random seeds.

Table 4: Summary for TINY IMAGENET

Baseline	L=64				
	Covariance Ratio	LPS_LGBM	LPS_MLP	NMSE_dB	NMSE_linear
AE-MMD-Gaussian (Ours)	54.386 \pm 2.220	0.029 \pm 0.002	0.001 \pm 0.004	-11.704 \pm 0.007	0.068 \pm 0.000
FactorVAE ($\gamma=10.0$)	2.338 \pm 0.286	0.437 \pm 0.064	0.483 \pm 0.065	-10.754 \pm 0.239	0.084 \pm 0.005
β -TCVAE ($\beta=4.0$)	4.950 \pm 0.264	0.794 \pm 0.018	0.932 \pm 0.015	-9.798 \pm 0.027	0.105 \pm 0.001
β -VAE ($\beta=4$)	4.934 \pm 0.104	0.726 \pm 0.018	0.872 \pm 0.035	-9.866 \pm 0.009	0.103 \pm 0.000
Standard AE	7.699 \pm 0.207	0.455 \pm 0.006	0.460 \pm 0.007	-11.775 \pm 0.005	0.066 \pm 0.000
Standard VAE	2.409 \pm 0.224	0.470 \pm 0.073	0.518 \pm 0.066	-10.608 \pm 0.361	0.087 \pm 0.008

Table 5: Summary for CIFAR10

Baseline	L=64				
	Covariance Ratio	LPS_LGBM	LPS_MLP	NMSE_dB	NMSE_linear
AE-MMD-Gaussian (Ours)	65.917 \pm 1.788	0.038 \pm 0.004	0.025 \pm 0.008	-15.615 \pm 0.008	0.027 \pm 0.000
FactorVAE ($\gamma=10.0$)	3.347 \pm 0.075	0.628 \pm 0.004	0.788 \pm 0.006	-12.491 \pm 0.012	0.056 \pm 0.000
β -TCVAE ($\beta=4.0$)	10.711 \pm 0.350	0.890 \pm 0.008	0.953 \pm 0.006	-10.684 \pm 0.005	0.085 \pm 0.000
β -VAE ($\beta=4$)	11.506 \pm 0.238	0.846 \pm 0.008	0.925 \pm 0.006	-10.492 \pm 0.006	0.089 \pm 0.000
Standard AE	6.929 \pm 0.192	0.508 \pm 0.005	0.513 \pm 0.005	-15.834 \pm 0.007	0.026 \pm 0.000
Standard VAE	3.405 \pm 0.113	0.632 \pm 0.009	0.793 \pm 0.010	-12.503 \pm 0.009	0.056 \pm 0.000

Table 6: Summary for MNIST

Baseline	L=12				
	Covariance Ratio	LPS_LGBM	LPS_MLP	NMSE_dB	NMSE_linear
AE-MMD-Gaussian (Ours)	46.635 \pm 2.442	0.220 \pm 0.007	0.333 \pm 0.006	-11.922 \pm 0.024	0.064 \pm 0.000
DGAE	6.191 \pm 0.368	0.632 \pm 0.002	0.671 \pm 0.005	-11.478 \pm 0.027	0.071 \pm 0.000
FactorVAE ($\gamma=10.0$)	2.628 \pm 0.094	0.271 \pm 0.066	0.275 \pm 0.098	-8.621 \pm 0.023	0.137 \pm 0.001
β -TCVAE ($\beta=4.0$)	8.071 \pm 2.238	0.838 \pm 0.022	0.893 \pm 0.015	-5.967 \pm 0.008	0.253 \pm 0.000
β -VAE ($\beta=4$)	9.493 \pm 1.000	0.898 \pm 0.009	0.938 \pm 0.007	-4.606 \pm 0.020	0.346 \pm 0.002
Standard AE	8.353 \pm 0.486	0.539 \pm 0.007	0.587 \pm 0.008	-12.089 \pm 0.020	0.062 \pm 0.000
Standard VAE	2.656 \pm 0.267	0.249 \pm 0.053	0.242 \pm 0.070	-8.630 \pm 0.011	0.137 \pm 0.000

Table 7: Summary for DSPRITES

Baseline	L=5									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian (Ours)	102.265 \pm 21.202	0.039 \pm 0.002	0.033 \pm 0.006	0.058 \pm 0.024	-17.888 \pm 0.281	0.016 \pm 0.001	0.000 \pm 0.000	0.026 \pm 0.018	0.025 \pm 0.017	0.950 \pm 0.002
AE-MMD-hybrid-sampler (Ours)	263.218 \pm 87.577	0.105 \pm 0.008	0.087 \pm 0.012	0.247 \pm 0.030	-16.311 \pm 0.261	0.023 \pm 0.001	0.246 \pm 0.027	0.200 \pm 0.038	0.196 \pm 0.038	0.957 \pm 0.001
DGAE	4.065 \pm 1.233	0.557 \pm 0.112	0.573 \pm 0.147	0.261 \pm 0.198	-14.614 \pm 2.173	0.040 \pm 0.026	0.000 \pm 0.000	0.226 \pm 0.104	0.219 \pm 0.101	0.962 \pm 0.004
FactorVAE ($\gamma=10.0$)	8.477 \pm 1.004	0.082 \pm 0.010	0.053 \pm 0.017	0.226 \pm 0.079	-14.564 \pm 0.110	0.035 \pm 0.001	0.000 \pm 0.000	0.081 \pm 0.033	0.079 \pm 0.032	0.964 \pm 0.004
β -TCVAE ($\beta=4.0$)	6.469 \pm 4.162	0.223 \pm 0.165	0.180 \pm 0.182	0.408 \pm 0.215	-8.831 \pm 0.085	0.131 \pm 0.003	0.000 \pm 0.000	0.212 \pm 0.119	0.195 \pm 0.105	0.960 \pm 0.005
β -VAE ($\beta=4$)	9.614 \pm 5.771	0.525 \pm 0.182	0.523 \pm 0.214	0.529 \pm 0.232	-8.200 \pm 0.057	0.151 \pm 0.002	0.000 \pm 0.000	0.236 \pm 0.114	0.217 \pm 0.100	0.956 \pm 0.005
Standard AE	12.669 \pm 2.560	0.187 \pm 0.039	0.181 \pm 0.042	0.078 \pm 0.032	-18.739 \pm 0.290	0.013 \pm 0.001	0.082 \pm 0.058	0.041 \pm 0.018	0.041 \pm 0.018	0.957 \pm 0.001
Standard VAE	5.528 \pm 1.291	0.100 \pm 0.016	0.047 \pm 0.023	0.436 \pm 0.365	-14.661 \pm 0.225	0.034 \pm 0.002	0.000 \pm 0.000	0.231 \pm 0.194	0.225 \pm 0.189	0.969 \pm 0.011

Table 8: Summary for XY

Baseline	L=2									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian (Ours)	240.970 \pm 243.891	-0.053 \pm 0.007	-0.004 \pm 0.003	0.751 \pm 0.554	-15.222 \pm 0.403	0.030 \pm 0.003	0.423 \pm 0.316	0.199 \pm 0.244	0.199 \pm 0.243	1.000 \pm 0.000
AE-MMD-hybrid-sampler (Ours)	380.827 \pm 408.352	-0.065 \pm 0.020	-0.003 \pm 0.005	1.047 \pm 0.414	-15.825 \pm 0.154	0.026 \pm 0.001	0.713 \pm 0.139	0.471 \pm 0.230	0.469 \pm 0.233	1.000 \pm 0.000
DGAE	16.631 \pm 19.417	0.248 \pm 0.380	0.241 \pm 0.385	1.580 \pm 1.414	-6.742 \pm 2.522	0.262 \pm 0.206	0.000 \pm 0.000	0.521 \pm 0.435	0.521 \pm 0.435	0.999 \pm 0.001
FactorVAE ($\gamma=10.0$)	2.022 \pm 1.392	-0.040 \pm 0.011	0.015 \pm 0.013	0.693 \pm 0.545	-15.364 \pm 0.359	0.029 \pm 0.002	0.537 \pm 0.268	0.337 \pm 0.255	0.336 \pm 0.255	1.000 \pm 0.000
β -TCVAE ($\beta=4.0$)	1.597 \pm 0.551	-0.042 \pm 0.015	0.009 \pm 0.006	1.149 \pm 1.093	-14.684 \pm 0.278	0.034 \pm 0.002	0.502 \pm 0.320	0.372 \pm 0.345	0.367 \pm 0.344	1.000 \pm 0.000
β -VAE ($\beta=4$)	2.424 \pm 1.371	-0.053 \pm 0.013	0.016 \pm 0.015	0.688 \pm 0.643	-14.457 \pm 0.137	0.036 \pm 0.001	0.522 \pm 0.323	0.358 \pm 0.339	0.358 \pm 0.339	1.000 \pm 0.000
Standard AE	503.878 \pm 536.454	-0.028 \pm 0.037	0.024 \pm 0.014	0.964 \pm 0.837	-16.904 \pm 0.081	0.020 \pm 0.000	0.470 \pm 0.348	0.362 \pm 0.369	0.362 \pm 0.369	1.000 \pm 0.000
Standard VAE	1.284 \pm 0.134	-0.055 \pm 0.011	0.015 \pm 0.007	0.195 \pm 0.098	-15.720 \pm 0.273	0.027 \pm 0.002	0.309 \pm 0.057	0.103 \pm 0.030	0.102 \pm 0.030	1.000 \pm 0.000

Table 9: Summary for XYC

Baseline	L=3									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian (Ours)	49.341 \pm 12.168	-0.014 \pm 0.016	0.016 \pm 0.007	0.256 \pm 0.118	-16.311 \pm 0.245	0.023 \pm 0.001	0.237 \pm 0.077	0.115 \pm 0.039	0.113 \pm 0.038	0.999 \pm 0.000
AE-MMD-hybrid-sampler (Ours)	234.644 \pm 274.142	0.168 \pm 0.030	0.117 \pm 0.035	0.465 \pm 0.336	-15.962 \pm 0.205	0.025 \pm 0.001	0.266 \pm 0.240	0.228 \pm 0.224	0.227 \pm 0.224	1.000 \pm 0.000
DGAE	4.802 \pm 2.926	0.498 \pm 0.417	0.503 \pm 0.414	0.340 \pm 0.181	-10.936 \pm 3.421	0.110 \pm 0.083	0.000 \pm 0.000	0.184 \pm 0.105	0.175 \pm 0.098	0.996 \pm 0.005
FactorVAE ($\gamma=10.0$)	1.581 \pm 0.076	-0.000 \pm 0.006	0.032 \pm 0.005	0.140 \pm 0.059	-15.919 \pm 0.219	0.026 \pm 0.001	0.183 \pm 0.157	0.079 \pm 0.069	0.076 \pm 0.065	1.000 \pm 0.000
β -TCVAE ($\beta=4.0$)	2.900 \pm 0.258	-0.010 \pm 0.009	0.023 \pm 0.007	0.188 \pm 0.105	-13.720 \pm 0.129	0.042 \pm 0.001	0.198 \pm 0.120	0.093 \pm 0.078	0.090 \pm 0.075	1.000 \pm 0.000
β -VAE ($\beta=4$)	1.861 \pm 0.119	0.003 \pm 0.010	0.027 \pm 0.009	0.266 \pm 0.149	-13.333 \pm 0.089	0.046 \pm 0.001	0.278 \pm 0.123	0.143 \pm 0.107	0.138 \pm 0.100	1.000 \pm 0.000
Standard AE	9.459 \pm 4.389	0.042 \pm 0.019	0.072 \pm 0.019	0.200 \pm 0.082	-17.811 \pm 0.134	0.017 \pm 0.001	0.227 \pm 0.090	0.087 \pm 0.052	0.085 \pm 0.049	1.000 \pm 0.000
Standard VAE	1.606 \pm 0.125	0.016 \pm 0.012	0.048 \pm 0.008	0.176 \pm 0.082	-16.043 \pm 0.147	0.025 \pm 0.001	0.197 \pm 0.056	0.086 \pm 0.043	0.085 \pm 0.043	0.999 \pm 0.000

Table 10: Summary for XYCS

Baseline	L=4									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian (Ours)	42.117 \pm 2.445	0.111 \pm 0.011	0.106 \pm 0.020	0.092 \pm 0.036	-16.757 \pm 0.143	0.021 \pm 0.001	0.110 \pm 0.032	0.048 \pm 0.016	0.048 \pm 0.016	0.997 \pm 0.000
AE-MMD-hybrid-sampler (Ours)	133.148 \pm 52.617	0.203 \pm 0.021	0.151 \pm 0.025	0.235 \pm 0.070	-16.101 \pm 0.283	0.025 \pm 0.002	0.278 \pm 0.060	0.149 \pm 0.051	0.146 \pm 0.050	0.998 \pm 0.000
DGAE	7.087 \pm 3.365	0.586 \pm 0.235	0.583 \pm 0.233	0.275 \pm 0.159	-15.023 \pm 0.701	0.032 \pm 0.005	0.000 \pm 0.000	0.190 \pm 0.076	0.187 \pm 0.074	0.997 \pm 0.001
FactorVAE ($\gamma=10.0$)	1.609 \pm 0.127	0.167 \pm 0.020	0.165 \pm 0.032	0.176 \pm 0.114	-15.517 \pm 0.366	0.028 \pm 0.002	0.192 \pm 0.090	0.149 \pm 0.098	0.148 \pm 0.097	0.997 \pm 0.001
β -TCVAE ($\beta=4.0$)	1.770 \pm 0.411	0.171 \pm 0.024	0.148 \pm 0.028	0.277 \pm 0.115	-12.402 \pm 0.257	0.058 \pm 0.003	0.227 \pm 0.109	0.246 \pm 0.108	0.245 \pm 0.107	0.999 \pm 0.000
β -VAE ($\beta=4$)	1.224 \pm 0.050	0.276 \pm 0.063	0.258 \pm 0.073	0.396 \pm 0.065	-10.983 \pm 0.672	0.081 \pm 0.012	0.238 \pm 0.053	0.264 \pm 0.031	0.262 \pm 0.031	0.988 \pm 0.006
Standard AE	12.609 \pm 4.281	0.265 \pm 0.028	0.284 \pm 0.031	0.099 \pm 0.052	-18.422 \pm 0.279	0.014 \pm 0.001	0.130 \pm 0.081	0.069 \pm 0.029	0.068 \pm 0.028	0.998 \pm 0.000
Standard VAE	1.630 \pm 0.175	0.179 \pm 0.042	0.165 \pm 0.051	0.163 \pm 0.077	-15.531 \pm 0.432	0.028 \pm 0.003	0.184 \pm 0.079	0.106 \pm 0.062	0.105 \pm 0.061	0.996 \pm 0.001

C Metric Analysis

This section provides a deeper analysis of the properties of common disentanglement metrics, justifying the introduction and use of our Latent Predictability Score (LPS). While alignment-based metrics like DCI and MIG are popular, our experiments reveal that they can exhibit paradoxical behavior and high instability, potentially leading to misleading conclusions about the quality of a learned representation.

C.1 The Alignment-Independence Paradox

A core assumption is that a higher alignment score (like DCI) corresponds to a better, more disentangled representation. Our findings challenge this notion, revealing a troubling paradox: optimizing for alignment can actively work against the goal of learning statistically independent factors.

This effect is visualized in the scatter plot in Figure 7. The ideal model would occupy the bottom-right quadrant, signifying both high alignment (high DCI) and high independence (low LPS). However, we observe that baseline models frequently land in the bottom-left or top-left "paradox zone." A clear example is DGAE on the XY dataset, which achieves the highest DCI score (0.520) but does so with a highly entangled representation (LPS of 0.25) and poor reconstruction. This suggests that DCI may be rewarding the model for learning spurious correlations, an effect conjectured in past work which noted that many benchmark datasets contain significant correlations between their ground-truth factors [LBLRGSB19]. For instance, in dSprites and XYCS, an object's size

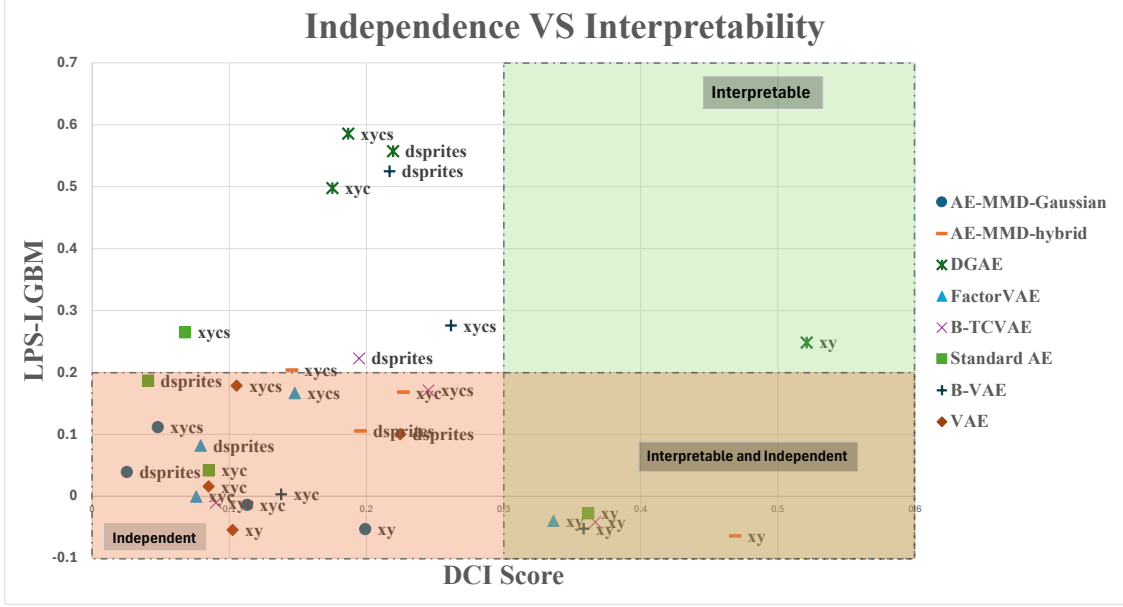


Figure 7: **The Paradox of Independence vs. Alignment.** Each point represents a model. The ideal model would be in the bottom-right (high alignment, high independence). We observe that models like DGAE occupy the unfavorable top-left quadrant, achieving a high DCI score only by sacrificing independence. Our method (with a Gaussian prior) occupies the bottom-left, prioritizing true independence over potentially flawed alignment metrics.

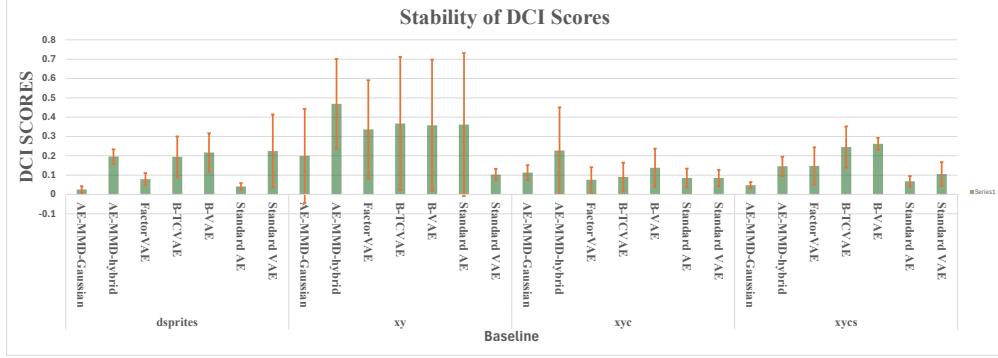
and position are co-dependent (a large shape cannot be centered near an edge). A model that correctly learns this co-dependence may score well on DCI but has fundamentally failed to learn truly independent factors. This highlights a critical flaw in relying solely on alignment-based metrics as a proxy for disentanglement.

Table 11: Comparison of Disentanglement (DCI) and Latent Predictability Score (LPS) across models and synthetic datasets. For Disentanglement, higher is better (\uparrow), and for LPS, lower is better (\downarrow).

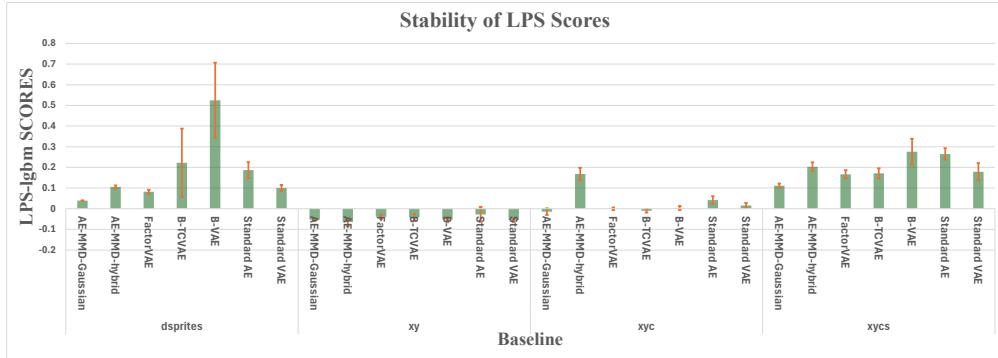
Baseline	dsprites		xy		xyc		xycs	
	Disent. (\uparrow)	LPS (\downarrow)	Disent. (\uparrow)	LPS (\downarrow)	Disent. (\uparrow)	LPS (\downarrow)	Disent. (\uparrow)	LPS (\downarrow)
AE-MMD-Gaussian	0.025	0.039	0.199	-0.053	0.113	-0.014	0.048	0.111
AE-MMD-hybrid	0.196	0.105	0.469	-0.065	0.227	0.168	0.146	0.203
DGAE	0.219	0.557	0.521	0.248	0.175	0.498	0.187	0.596
FactorVAE	0.079	0.082	0.336	-0.040	0.076	-0.000	0.148	0.187
β -TCVAE	0.195	0.222	0.368	-0.042	0.090	-0.010	0.245	0.171
β -VAE	0.217	0.525	0.358	-0.053	0.138	0.003	0.262	0.276
Standard AE	0.041	0.187	0.362	-0.028	0.085	0.042	0.068	0.265
Standard VAE	0.225	0.100	0.102	-0.055	0.085	0.016	0.105	0.180

C.2 Metric Instability Across Runs

Beyond the paradox, we find that standard alignment metrics are highly unstable, exhibiting extreme variance across different random seeds. This is quantified and visualized in Figure 8. As shown in



(a) DCI Score Instability



(b) LPS Score Stability

Figure 8: **Comparison of metric stability across 5 random seeds.** The error bars represent one standard deviation. **(a)** The DCI scores exhibit extremely high variance, with the standard deviation often being as large as the mean itself, highlighting the metric’s unreliability. **(b)** In stark contrast, our proposed LPS metric shows minimal variance, demonstrating its stability and reliability as an intrinsic measure of representation quality.

Figure 8(a), the standard deviation of the DCI score for baseline methods is often as large as, or even larger than, the mean itself. This instability, a known symptom of the NICA identifiability problem [BLRGSB19], makes scores from single runs untrustworthy and complicates reliable model comparison.

In stark contrast, Figure 8(b) shows that our LPS metric is highly stable, with consistently small error bars across all models and datasets. This stability stems from the fact that LPS measures an intrinsic, unsupervised property of the representation itself (mutual independence) rather than its alignment to a specific, and potentially arbitrary, set of ground-truth labels.

C.3 LPS as a Stable, Intrinsic Measure

The paradoxical behavior and instability of alignment-based metrics underscore the need for a more fundamental and reliable measure of representation quality. We argue that the LPS fills this critical gap. By focusing on statistical independence, LPS provides a stable, intrinsic, and fully unsupervised signal. It directly answers the question, “Are the learned features statistically independent?” without being confounded by the potential issues of the ground-truth labels or the inherent instability of the alignment problem. While alignment is a desirable downstream goal,

achieving a verifiably independent representation is a more foundational first step, and LPS provides the tool to measure it robustly.

D The Power of Programmable Priors: Extended Results

This section provides a deeper dive into the concept of "programmable priors," showcasing (1) the flexibility of our MMD framework and (2) how that can be leveraged to significantly improve the alignment of learned representations with ground-truth factors. We present results for our **AE-MMD-hybrid-sampler** model, where the latent space was sculpted to match a custom set of per-dimension priors, reflecting a better inductive bias for each dataset. The following visual and quantitative data will demonstrate that this prior engineering significantly improves ground-truth alignment metrics over the baseline **AE-MMD-Gaussian** model.

D.1 dSprites: Engineering a Mixed-Distribution Prior

For the 5-dimensional latent space of dSprites, we designed a hybrid prior to test the model's ability to learn a mix of simple and multi-modal distributions. The visual results in Figure 9 confirm that the model's learned marginals consistently lead to improved recovery of desired latent features.

Analysis. The figure shows a direct correspondence between our specified priors and the learned representations:

- **Dimension 1 (Uniform):** We specified a **Uniform**(-5, 5) prior. The model chose to ignore this instruction during training and instead adopted a Gaussian distribution with the same mean and variance.
- **Dimension 2 (Gaussian):** The target was a wide **Gaussian**(0, 2), and the model learned the corresponding bell curve.
- **Dimensions 3 (Bi-modal GMMs):** For this dimensions, we used **auto-gmm** priors with two modes centered at -5 and 5, and with standard deviation equal to 5. The histograms clearly show that the model learned to partition the data into a more uniform distribution. We have found that specifying such a GMM structure would generally help the model adhere to a more uniform representation. while specifying the uniform distribution directly as in Dimension 1 would sometimes be 'ignored'.
- **Dimensions 4-5 (Bi-modal GMMs):** For the remaining dimensions, we used **auto-gmm** priors with two modes located at -10 and 10 respectively with standard deviation of 1 (using a small standard deviation of 1 is used to ensure the model adopts a bi-modal representation and not a uniform distribution as observed in Dimension 3) . The histograms clearly show that the model learned to partition the data into these distinct modes for each respective dimension.

Quantitative Results. As shown in Table 12, this engineered prior leads to a dramatic improvement in alignment metrics. The **AE-MMD-hybrid-sampler** achieves a MIG score of 0.247, an over **4x improvement** compared to the Gaussian prior's score of 0.058. Similar substantial gains are seen in the SAP score (0.246 vs. 0.000) and the DCI disentanglement score (0.196 vs. 0.025), demonstrating that providing a better inductive bias directly translates to a more interpretable and aligned representation.

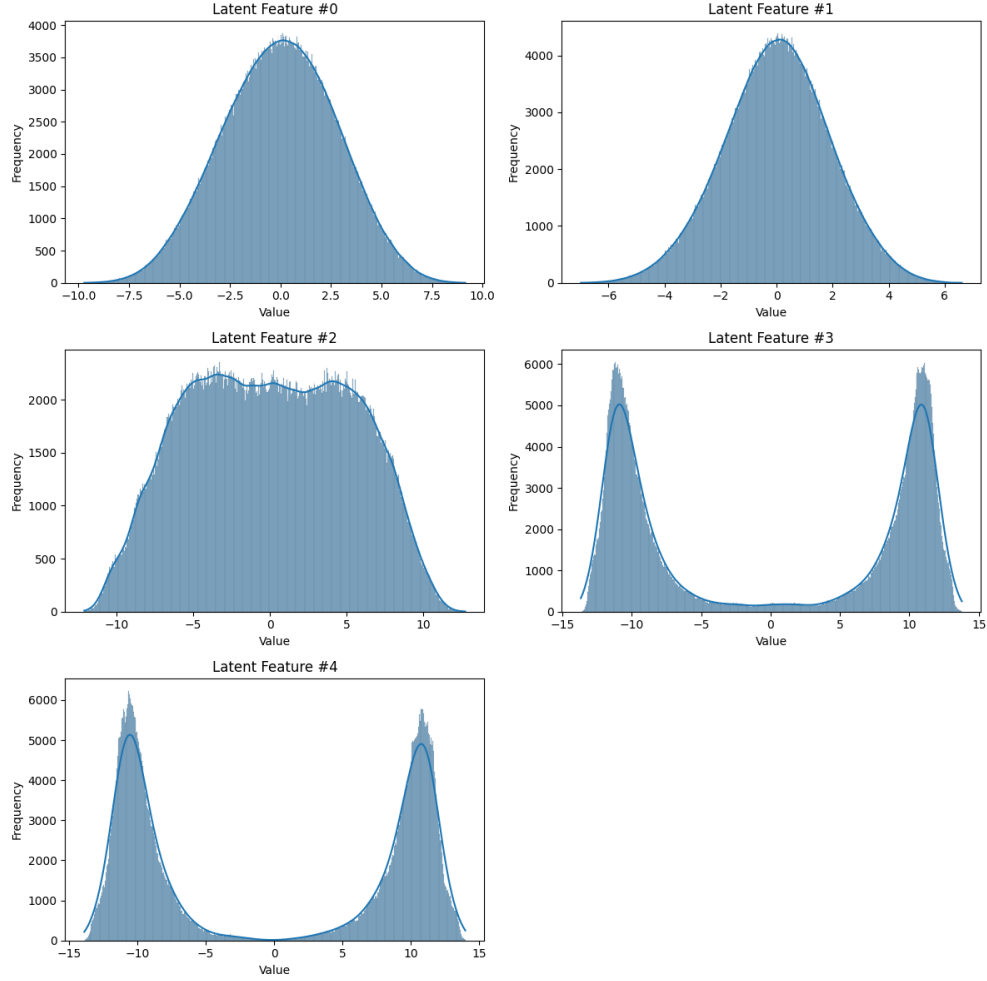


Figure 9: Learned marginal distributions for the **AE-MMD-hybrid-sampler** on dSprites. The histograms provide direct visual proof that the model successfully learned to match the specified target distributions for each latent dimension.

Table 12: Summary for DSPRITES

Baseline	L=5									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian	102.265 \pm 21.202	0.039 \pm 0.002	0.033 \pm 0.006	0.058 \pm 0.024	-17.888 \pm 0.281	0.016 \pm 0.001	0.000 \pm 0.000	0.026 \pm 0.018	0.025 \pm 0.017	0.950 \pm 0.002
AE-MMD-hybrid-sampler	263.218 \pm 87.577	0.105 \pm 0.008	0.087 \pm 0.012	0.247 \pm 0.030	-16.311 \pm 0.261	0.023 \pm 0.001	0.246 \pm 0.027	0.200 \pm 0.038	0.196 \pm 0.038	0.957 \pm 0.001

D.2 XY: Engineering a Uniform Prior

We provide the full table of results for the XY experiment in the main paper.

Table 13: Summary for XY

Baseline	L=2									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian	240.970 \pm 243.891	-0.053 \pm 0.007	-0.004 \pm 0.003	0.751 \pm 0.554	-15.222 \pm 0.403	0.030 \pm 0.003	0.423 \pm 0.316	0.199 \pm 0.244	0.199 \pm 0.243	1.000 \pm 0.000
AE-MMD-hybrid-sampler	380.827 \pm 408.352	-0.065 \pm 0.020	-0.003 \pm 0.005	1.047 \pm 0.414	-15.825 \pm 0.154	0.026 \pm 0.001	0.713 \pm 0.139	0.471 \pm 0.230	0.469 \pm 0.233	1.000 \pm 0.000

D.3 XYZ: Engineering a GMM-Heavy Prior

On the XYZ dataset, we used a prior dominated by bi-modal distributions to encourage the model to discover clustered structures. Figure 10 visualizes the successful enforcement of this prior.

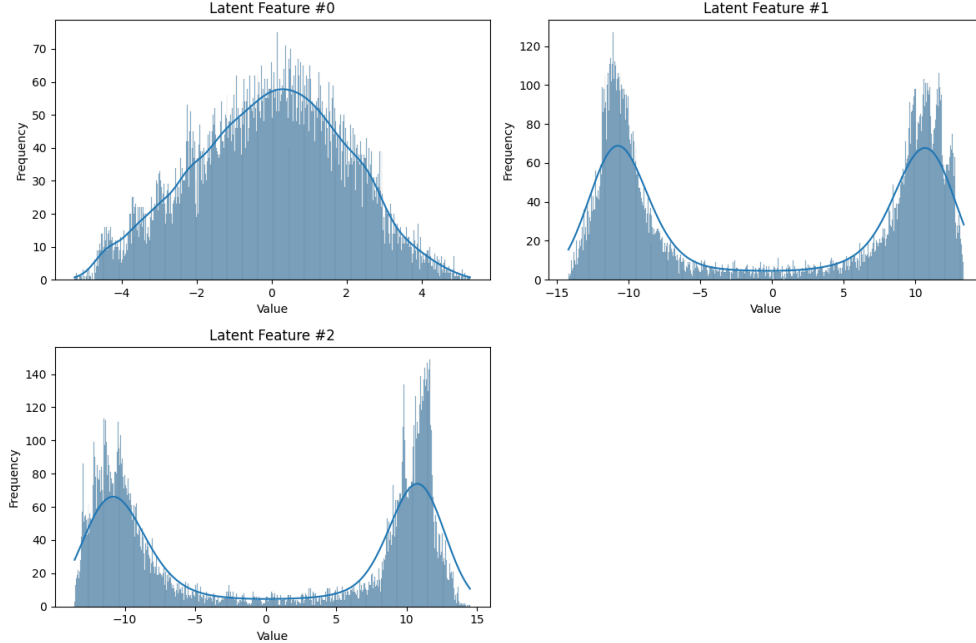


Figure 10: Learned marginal distributions for the AE-MMD-hybrid-sampler on XYZ.

Analysis. The learned marginals again match the specified inductive bias:

- **Dimension 1 (Gaussian):** The model correctly learns a $\text{Gaussian}(0, 2)$ distribution.
- **Dimensions 2 & 3 (Bi-modal GMMs):** We specified two identical `auto-gmm` priors with modes at -10 and 10. The resulting histograms clearly show two distinct clusters centered at these values, demonstrating the model’s ability to learn multi-modal representations.

Quantitative Results. The quantitative results in Table 14 confirm the benefit of this custom prior. The hybrid-sampler model nearly doubles the MIG score (0.465 vs. 0.256) and the DCI disentanglement score (0.227 vs. 0.113) compared to the model trained with a simple Gaussian prior.

Table 14: Summary for XYZ

Baseline	L=3									
	Covariance Ratio	LPS.LGBM	LPS.MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian	49.341 \pm 12.168	-0.014 \pm 0.016	0.016 \pm 0.007	0.256 \pm 0.118	-16.311 \pm 0.245	0.023 \pm 0.001	0.237 \pm 0.077	0.115 \pm 0.039	0.113 \pm 0.038	0.999 \pm 0.000
AE-MMD-hybrid-sampler	324.644 \pm 274.142	0.168 \pm 0.030	0.117 \pm 0.035	0.465 \pm 0.336	-15.962 \pm 0.205	0.025 \pm 0.001	0.266 \pm 0.240	0.228 \pm 0.224	0.227 \pm 0.224	1.000 \pm 0.000

D.4 XYCS: Engineering a Mixed-Component Prior

For the 4-dimensional XYCS dataset, we again specified a mixed set of priors. The results in Figure 11 show that the model faithfully learns this complex target geometry.

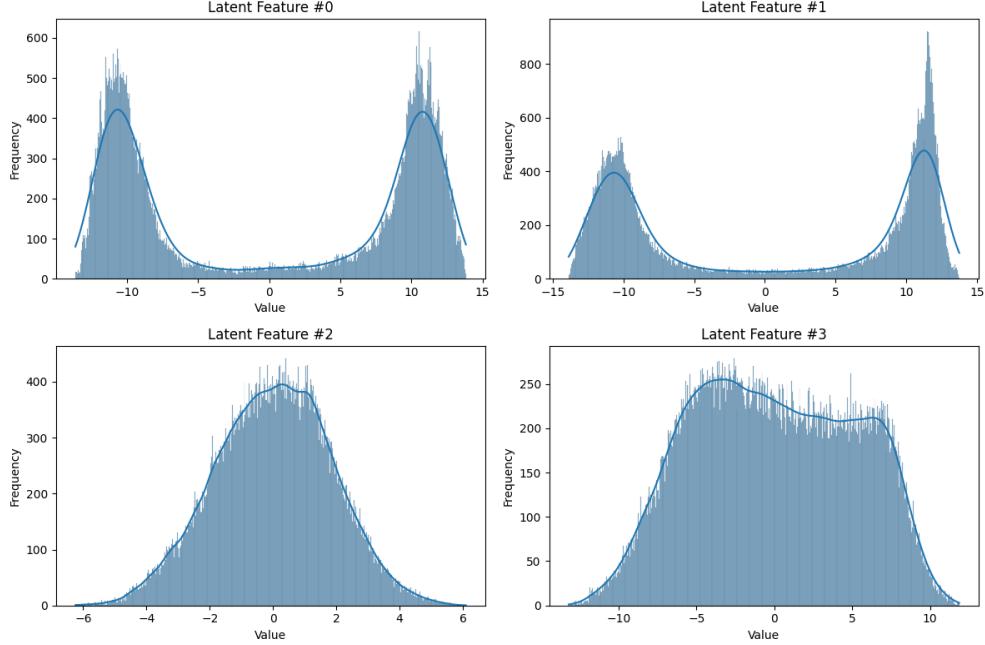


Figure 11: Learned marginal distributions for the AE-MMD-hybrid-sampler on XYCS.

Analysis. The learned distributions follow the specified configuration: two bi-modal GMMs with modes at ± 10 , a standard Gaussian, and a narrower bi-modal GMM with modes at ± 5 . Each histogram in the figure visually confirms that the intended structure was successfully imposed on the corresponding latent dimension.

Quantitative Results. As shown in Table 15, injecting this more informed inductive bias again leads to superior alignment. The hybrid-sampler model achieves a MIG score of 0.235, more than doubling the 0.092 score from the Gaussian model. Likewise, the SAP and DCI scores show significant improvements of over 2.5x and 3x, respectively.

Table 15: Summary for XYCS

Baseline	L=4									
	Covariance Ratio	LPS_LGBM	LPS_MLP	MIG_mig_score	NMSE_dB	NMSE_linear	SAP_sap_score	completeness	disentanglement	informativeness
AE-MMD-Gaussian	42.117 \pm 2.445	0.111 \pm 0.011	0.106 \pm 0.020	0.092 \pm 0.036	-16.757 \pm 0.143	0.021 \pm 0.001	0.110 \pm 0.032	0.048 \pm 0.016	0.048 \pm 0.016	0.997 \pm 0.000
AE-MMD-hybrid-sampler	133.148 \pm 52.617	0.203 \pm 0.021	0.151 \pm 0.025	0.235 \pm 0.070	-16.101 \pm 0.283	0.025 \pm 0.002	0.278 \pm 0.060	0.149 \pm 0.051	0.146 \pm 0.050	0.998 \pm 0.000

Summary

The results in this section highlight a key takeaway: while a factorized Gaussian prior is highly effective at achieving statistical independence, it is not always the optimal choice for interpretability and alignment with ground-truth factors. The true power of the MMD framework is its flexibility, which allows practitioners to inject domain knowledge by engineering specific latent geometries. This capability leads to representations that are not only independent (or have a desired dependence structure) but are also more meaningfully aligned with the true factors of variation in the data.

However, our experiments with programmable priors also revealed a nuance in the MMD framework’s behavior. While the regularizer consistently learned unimodal, Gaussian-like distributions with

high fidelity, enforcing more specific geometries, such as uniform or multi-modal distributions, could sometimes present challenges. In these cases, the model occasionally defaulted to a unimodal distribution that correctly matched the target’s mean and variance but not its overall shape.

We hypothesize this is primarily due to the inductive bias of the RBF kernel. As the RBF kernel is itself a Gaussian function, it is naturally more adept at matching distributions defined by their low-order moments [TBGS17]. A secondary factor may be architectural constraints; the purely convolutional nature of the encoder might limit its capacity to shape the latent space into more complex geometries. We believe that future work could explore alternative kernels or incorporate fully connected layers to enhance the encoder’s ability to model a wider variety of prior distributions.

E Additional Experiments

E.1 Robustness of MMD to Latent Space Size

While the main text demonstrates the robustness of our MMD regularizer, this section provides the corresponding plots for both the ‘LPS-lgbm’ and ‘LPS-mlp’ variants compared to the reconstruction NMSE in decibels for completeness. The experiment, conducted on CIFAR-10, was designed to test how well different methods maintain statistical independence as the dimensionality (and thus capacity) of the latent space increases.

As shown in Figure 12, our MMD-based framework consistently enforces a high degree of mutual independence across all tested latent dimensions. The LPS scores for our model remain low and stable, indicating that the learned features stay disentangled regardless of the bottleneck size. In contrast, most baseline methods exhibit a clear degradation in performance; as the latent space grows, their LPS scores rise, suggesting that their representations become progressively more entangled. This highlights a key advantage of our direct, non-parametric enforcement mechanism: it provides a reliable and scalable method for structuring latent spaces, a property not shared by many VAE-based regularizers.

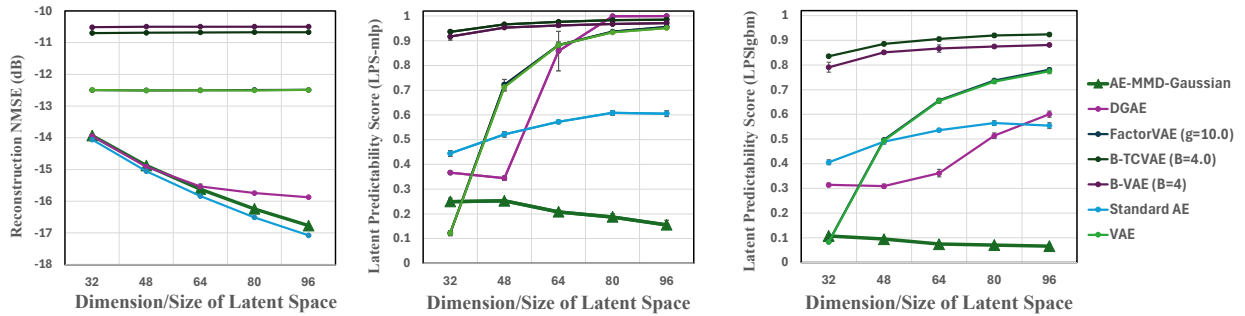


Figure 12: **Latent Independence vs. Latent Space Dimension on CIFAR-10.** The plots show (left) the reconstruction NMSE in decibels, (middle-right) the Latent Predictability Score (LPS) with mlp and lgbm regression respectively for various models as the size of the latent bottleneck increases. Lower scores are better. Our method (AE-MMD-Gauss, green line) maintains a high degree of independence across all dimensions, while most baselines learn increasingly entangled representations. Moreover, we do so at no significant cost to the reconstruction error which is only beaten by the unregularized standard AE model.

E.2 MMD for Representational Knowledge Transfer

In the main text, we demonstrated our framework’s ability to ”copy” the latent distribution of a target model. Here, we re-examine this experiment from the perspective of student-teacher knowledge transfer to probe a deeper question: is matching the aggregate posterior sufficient to transfer a learned representation? We note that Huang and Wang [HW17] has also previously utilized MMD regularization for representational knowledge distillation.

In this experiment, a ”student” model was trained with our MMD regularizer to replicate the entire aggregate posterior distribution of a pre-trained, unregularized ”teacher” autoencoder. As shown in the left panel of Figure 13, the student was highly successful—the learned marginals and covariance matrix are nearly identical to those of the teacher. However, the right panel reveals that this statistical alignment does not translate to functional equivalence. Latent traversals, which reveal the semantic meaning encoded in each dimension, produce wildly different outputs for the same input image. While a traversal in the teacher model might alter an attribute like the slant of a digit, the corresponding traversal in the student model results in a completely different transformation.

This finding is crucial: **matching the aggregate posterior distribution is not enough to transfer a model’s learned representation**. Even with identical latent distributions, the two models have learned fundamentally different mappings from the input data to the latent space. This suggests that for true representational knowledge transfer, one must enforce a more complex constraint, such as matching the *joint* distribution of the latent representation with observable variables from the input data.

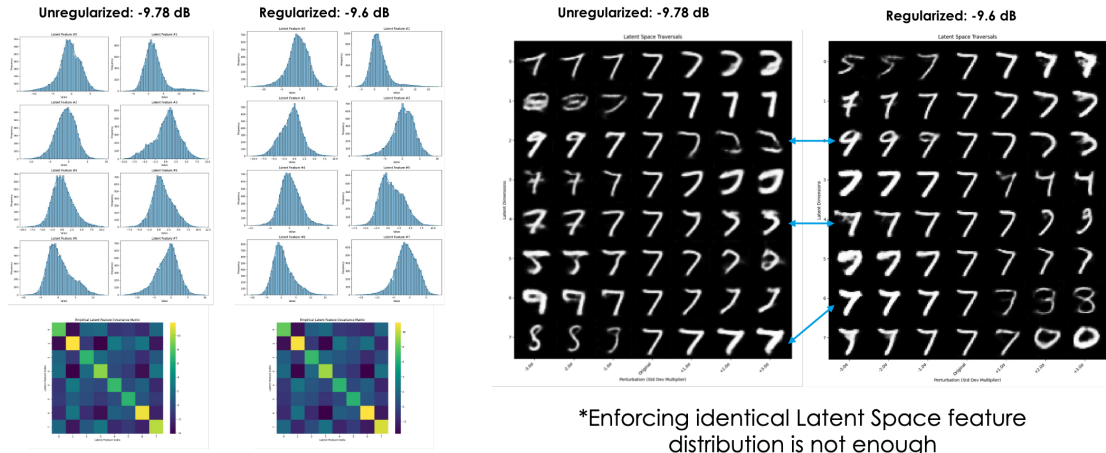


Figure 13: **Failure of Knowledge Transfer via Aggregate Posterior Matching.** (Left) Our MMD regularizer successfully forces a ”student” model to replicate the latent distribution of a ”teacher” model. (Right) Despite this, latent traversals show that the semantic meaning of the latent dimensions is completely different. For the same inputs (connected by lines), traversing a latent feature produces different visual changes, demonstrating that the underlying representations are not the same.