

Seminar Thesis

InfoVAE: Balancing Learning and Inference in Variational Autoencoders

Department of Statistics
Ludwig-Maximilians-Universität München

Mert Ekici

Munich, July 25th, 2025



Supervised by M.Sc. Simon Rittel and Dr. Ludwig Bothmann

Abstract

Classical variational autoencoders (VAEs) often suffer from amortized inference failures and the information preference problem, resulting in poor posterior fidelity and uninformative latent codes. The InfoVAE framework extends the standard objective by introducing a posterior-prior trade-off coefficient λ and an information preference weight α , thereby balancing reconstruction accuracy, global distribution alignment, and mutual information retention. This study evaluates InfoVAE through three dimensions: adaptive scheduling of $\{\lambda, \alpha\}$ to stabilize training and accelerate convergence; substitution of the KL divergence with alternative strict divergences to measure their impact on alignment, reconstruction quality, downstream utility, and computational cost; and comparison with lightweight regularization schemes that reduce overall model complexity. Results indicate that dynamic adaptation of $\{\lambda, \alpha\}$ improves stability, kernel and density-ratio-based penalties achieve the strongest posterior alignment and representation utility, and simpler methods approach similar performance at higher latent capacities with markedly lower overhead. These findings clarify the practical trade-offs in VAE design and inform the selection of hyperparameters, divergence and regularization strategies.

Contents

1	Introduction	1
2	Methodology	3
2.1	Balancing Learning and Inference	3
2.1.1	Amortized Inference Failure	3
2.1.2	Information Preference Problem	4
2.1.3	Evidence Lower Bound	5
2.2	InfoVAE Model Family	5
2.2.1	InfoVAE Objective	5
2.2.2	InfoVAE Optimality	6
2.3	Parameters λ and α	6
2.4	Divergence Families	7
2.5	Alternatives for Latent Utility	9
3	Experiments	10
3.1	Evaluation Metrics	10
3.2	Experimental Setup	11
3.3	Experiment 1: Hyperparameter Sensitivity	13
3.4	Experiment 2: Divergence Substitution	15
3.5	Experiment 3: Model Comparison	17
3.6	Contrast with Original InfoVAE Experiments	21
4	Conclusion	22
A	Appendix	V
B	Electronic appendix	VIII

1 Introduction

Variational autoencoders are a type of deep generative model that learn to represent complex data in a lower-dimensional latent space while still allowing for the generation of new samples (Kingma and Welling, 2014). During training, a VAE simultaneously trains an encoder network, which maps each input to a distribution over latent codes, and a decoder network, which reconstructs the original input from samples of these codes. The learning objective balances two goals in a single procedure: reconstructing inputs accurately and shaping the latent space to follow a simple reference distribution, most often a standard normal. Thanks to this combination of reconstruction and regularization, VAEs can both compress high-dimensional data into meaningful features and generate novel, high-quality examples by sampling from the reference distribution and passing samples through the decoder. This blend of representation learning and generative capability has made VAEs a foundational tool in fields such as image synthesis, semi-supervised learning, anomaly detection, and beyond.

Despite their widespread applicability, variational autoencoders exhibit some inherent shortcomings that can limit their effectiveness in practice. In particular, the standard training objective may fail to enforce a faithful match between the learned posterior and the true data distribution, and overly powerful decoders can render the latent code uninformative (Chen et al., 2017). As a result, the model’s uncertainty estimates become unreliable and the latent representations lose their semantic meaning, reducing interpretability. This misalignment further hinders the VAE’s ability to capture key variations in the data and diminishes the practical utility of the learned features.

InfoVAE extends the classical variational autoencoder by adding two new terms to the training objective (Zhao et al., 2019). The first term applies a weighted penalty on the divergence between the overall encoding distribution and the chosen prior, giving fine control over how the latent codes are organized. The second term introduces a mutual information criterion, which explicitly encourages the encoder to capture and preserve information from the input in the latent space. By adjusting the weights of these terms, InfoVAE can balance data reconstruction, posterior regularization, and information retention in a transparent way. In addition, InfoVAE supports a variety of divergence measures for the aggregate penalty, including maximum mean discrepancy, adversarial criteria, and kernel-based methods. This flexibility yields a unified family of models that encompasses earlier variants such as the beta-VAE and the adversarial autoencoder as special cases. As a result, InfoVAE avoids latent collapse and posterior distortion by enforcing both global distribution alignment and local informativeness, providing a principled framework to tailor generative and inference behavior to the needs of diverse applications (Zhao et al., 2019).

$$\mathcal{L}_{\text{InfoVAE}} = -\lambda D_{\text{KL}}(q_{\phi}(z) \| p(z)) - E_{q_{\phi}(z)}[D_{\text{KL}}(q_{\phi}(x | z) \| p_{\theta}(x | z))] + \alpha I_q(x; z)$$

While InfoVAE effectively remedies the core failings of classical VAEs, it also introduces new practical challenges that require careful evaluation. First, the additional terms in the objective depend on hyperparameters whose fixed values can lead to unstable or sub-optimal training, highlighting the need to study adaptive scheduling and robustness to parameter choice (Zhao et al., 2019, Liu and Wang, 2025). Second, the specific divergence used to align the aggregate posterior with the prior has a strong influence on the speed of convergence, the sample fidelity and the quality of the latent space, motivating a systematic comparison as the alternatives such as the maximum mean discrepancy, adversarial criteria, and Stein variational methods (Liu and Wang, 2016, Zhao et al., 2019). Finally, although InfoVAE can achieve excellent performance when well-tuned, simpler modifications such as batch-normalized encoders or contrastive autoencoding often match its latent representation utility with far less complexity and tuning effort (Zhu et al., 2020). In this seminar, we present a concise yet comprehensive critique of InfoVAE through experiments on hyperparameter sensitivity, divergence substitution, and model comparison.

The remainder of this seminar thesis is organized as follows. In Chapter 2, we present the methodological framework of InfoVAE. Section 2.1 examines the balance between learning and inference in standard VAEs, Section 2.2 introduces the InfoVAE model family and its modified objective, Section 2.3 discusses the new parameters governing posterior alignment and information preservation, Section 2.4 surveys the choice of divergence measures, and Section 2.5 explores lighter-weight alternatives for maintaining latent utility. Chapter 3 details our empirical evaluation: Section 3.1 defines the evaluation metrics, Section 3.2 describes the experimental setup, Section 3.3 investigates hyperparameter sensitivity, Section 3.4 compares different divergence substitutions, Section 3.5 provides an overall model comparison with lighter alternatives, and Section 3.6 reviews the original InfoVAE paper experiments. Chapter 4 concludes with a summary of our findings and suggestions for future research.

2 Methodology

Variational autoencoders aim to reconcile two objectives: accurate reconstruction of observed data and a well-structured, regularized latent space. In practice, however, this trade-off can fail, resulting in poor posterior estimates or latent representations that carry little information. InfoVAE addresses these shortcomings by augmenting the VAE objective with additional terms and a mutual information maximization component to restore a reliable balance between learning and inference (Zhao et al., 2019).

2.1 Balancing Learning and Inference

Two principal shortcomings emerge: the encoder can distort posterior estimates by sacrificing inference accuracy for reconstruction, and an expressive decoder can bypass the latent representation entirely, yielding uninformative codes. These problems are detailed in the following sections.

Many applications such as semi-supervised learning depend more critically on accurate inference of latent codes than on generating visually sharp samples. When the encoder overfits to reconstruction, downstream performance can suffer despite high ELBO values. Additionally, because the true data distribution is only known through a finite dataset, driving the encoder to match each observed point too precisely can lead to overfitting and poor generalization to new data.

2.1.1 Amortized Inference Failure

Variational autoencoders rely on an encoder network $q_\phi(z | x)$ to approximate the true posterior $p_\theta(z | x)$, while jointly learning a generative model $p_\theta(x | z)$. In practice, however, optimizing the ELBO objective can incentivize the encoder to distort its output in order to improve reconstruction, rather than to approximate the true posterior accurately. Concretely, the reconstruction term alone pushes the encoder’s variances toward zero and its means toward extreme values, effectively “overfitting” individual data points and driving up the ELBO even as the inferred posterior diverges from $p_\theta(z | x)$ (Zhao et al., 2019). This phenomenon known as *amortized inference failure* leads to unreliable uncertainty estimates and poor downstream performance on tasks that depend on faithful posterior representations.

Proposition 1 (High ELBO Values Do Not Imply Accurate Posterior Inference). *Let the data space \mathcal{X} and latent space \mathcal{Z} both be \mathbb{R} , and consider a dataset $\mathcal{D} = \{-1, 1\}$. Suppose the decoder family $p_\theta(x | z)$ and the encoder family $q_\phi(z | x)$ are sufficiently expressive Gaussian distributions. Then one can choose encoder parameters*

$$\mu_\phi^q(x = 1) \rightarrow +\infty, \quad \mu_\phi^q(x = -1) \rightarrow -\infty, \quad \sigma_\phi^q(x) \rightarrow 0,$$

so that the ELBO grows without bound, while at the same time

$$D_{\text{KL}}(q_\phi(z | x) \parallel p_\theta(z | x)) \rightarrow +\infty,$$

demonstrating that high ELBO values do not guarantee accurate posterior inference (Chen et al., 2017).

Proof. Let the decoder be $p_\theta(x | z) = \mathcal{N}(x; z, \sigma^2)$ for some fixed $\sigma > 0$, and let the encoder be $q_\phi(z | x) = \mathcal{N}(z; \mu_\phi(x), \epsilon^2)$ with $\epsilon > 0$. Take $\mu_\phi(1) = M$, $\mu_\phi(-1) = -M$ and $\epsilon \rightarrow 0$ as $M \rightarrow \infty$.

Reconstruction term. For each $x \in \{-1, 1\}$,

$$E_{q_\phi(z|x)}[\log p_\theta(x | z)] = -\frac{1}{2\sigma^2} E_{q_\phi(z|x)}[(x - z)^2] + \text{const} \approx \frac{M^2}{2\sigma^2} \quad \text{as } M \rightarrow \infty,$$

so the reconstruction term grows on the order of $M^2/(2\sigma^2)$.

KL regularization. Meanwhile,

$$D_{\text{KL}}(q_\phi(z | x) \| p(z)) = \frac{1}{2} \left(\mu_\phi(x)^2 + \epsilon^2 - \log \epsilon^2 - 1 \right) = \frac{M^2}{2} + O(\log M),$$

which grows only as $M^2/2$.

Choosing $\sigma < 1$ makes the coefficient $1/(2\sigma^2)$ exceed $1/2$, so the net ELBO diverges to $+\infty$. However, the true posterior under p_θ is $\mathcal{N}(z; \frac{x}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2})$, and hence

$$D_{\text{KL}}(q_\phi(z | x) \| p_\theta(z | x)) \approx \frac{M^2}{2} (1 + o(1)) \rightarrow +\infty.$$

Thus the ELBO can be driven arbitrarily high even as the true-posterior gap becomes unbounded, proving the claim. \square

2.1.2 Information Preference Problem

A complementary issue arises when the decoder $p_\theta(x | z)$ is sufficiently expressive to model the data distribution without using information from z . In this regime, the optimal solution sets $p_\theta(x | z) = p_D(x)$ for every z , and drives the encoder to match the prior completely ($q_\phi(z | x) = p(z)$), so that the latent code carries no information about the inputs (Chen et al., 2017). Although the ELBO remains high, the model fails to learn meaningful representations, undermining the core motivation for using latent variables in unsupervised learning (Zhao et al., 2019).

More formally, consider the ELBO in its marginal form (Eq. 2). The first divergence term,

$$D_{\text{KL}}(p_D(x) \| p_\theta(x)),$$

is driven to zero if there exists a decoder $p_\theta^*(x | z)$ that reproduces the data marginal $p_D(x)$ for all z . The second term,

$$E_{p_D(x)}[D_{\text{KL}}(q_\phi(z | x) \| p_\theta(z | x))],$$

also vanishes if the encoder collapses to the prior, $q_\phi(z | x) = p(z) = p_\theta(z | x)$. When both terms are zero, the ELBO reaches its global maximum even though the latent variables are ignored. This *information preference* effect was first noted by Chen et al. (2017) via a coding-efficiency argument and formally analyzed in Zhao et al. (2019). In effect, a purely likelihood-based objective offers no incentive to use the latent code when the decoder is sufficiently powerful, leading to latent collapse and trivial representations.

InfoVAE remedies this failure mode by introducing an explicit mutual information term into the objective, which restores a positive gain for encoding input-dependent variation in z (Zhao et al., 2019).

2.1.3 Evidence Lower Bound

To train a VAE, we introduce an approximate posterior $q_\phi(z | x)$ and optimize the evidence lower bound (ELBO) on the marginal log-likelihood (Kingma and Welling, 2014):

$$\mathcal{L}_{\text{ELBO}} = E_{q_\phi(z|x)}[\log p_\theta(x | z)] - D_{\text{KL}}(q_\phi(z | x) \| p(z)) \leq \log p_\theta(x).$$

Equivalently, defining the joint distributions

$$p_\theta(x, z) = p(z) p_\theta(x | z), \quad q_\phi(x, z) = p_D(x) q_\phi(z | x),$$

the ELBO can be written up to an additive constant as

$$\mathcal{L}_{\text{ELBO}} \equiv -D_{\text{KL}}(q_\phi(x, z) \| p_\theta(x, z)) \quad (1)$$

$$= -D_{\text{KL}}(p_D(x) \| p_\theta(x)) - E_{p_D(x)}[D_{\text{KL}}(q_\phi(z | x) \| p_\theta(z | x))] \quad (2)$$

$$= -D_{\text{KL}}(q_\phi(z) \| p(z)) - E_{q_\phi(z)}[D_{\text{KL}}(q_\phi(x | z) \| p_\theta(x | z))]. \quad (3)$$

2.2 InfoVAE Model Family

InfoVAE modifies the standard ELBO by integrating two new components that will learn both the correct model and amortized inference distributions (Zhao et al., 2019). The mutual information maximization term in InfoVAE encourages high mutual information between x and z . This leads the model to use the latent code and avoids the information preference problem.

2.2.1 InfoVAE Objective

The starting point for InfoVAE is the ELBO in its marginal form (Eq. 3), which highlights the tension between matching the aggregate posterior $q_\phi(z)$ to the prior $p(z)$ and reconstructing the data:

$$\mathcal{L}_{\text{ELBO}} = -D_{\text{KL}}(q_\phi(z) \| p(z)) - E_{q_\phi(z)}[D_{\text{KL}}(q_\phi(x | z) \| p_\theta(x | z))].$$

InfoVAE augments this objective with two additional terms to correct the pathologies of classical VAEs (Zhao et al., 2019):

$$\mathcal{L}_{\text{InfoVAE}} = -\lambda D_{\text{KL}}(q_\phi(z) \| p(z)) - E_{q_\phi(z)}[D_{\text{KL}}(q_\phi(x | z) \| p_\theta(x | z))] + \alpha I_q(x; z).$$

In practical form (Eq. 6 of Zhao et al., 2019), this becomes

$$\begin{aligned} \mathcal{L}_{\text{InfoVAE}} = & E_{p_D(x)} E_{q_\phi(z|x)} [\log p_\theta(x | z)] \\ & - (1 - \alpha) E_{p_D(x)} [D_{\text{KL}}(q_\phi(z | x) \| p(z))] \\ & - (\alpha + \lambda - 1) D(q_\phi(z) \| p(z)), \end{aligned}$$

where D may be KL, MMD or any strict divergence. The parameter λ controls the strength of aggregate posterior regularization, while α directly weights the mutual information between x and z , preventing latent collapse. By varying α and λ , InfoVAE interpolates between several known VAE variants: setting $\alpha = 0, \lambda = 1$ recovers the original VAE (Kingma and Welling, 2014), enforcing $\alpha + \lambda - 1 = 0$ yields the β -VAE (Higgins et al., 2017), and choosing $\alpha = 1, \lambda = 1$ with a Jensen–Shannon divergence produces the adversarial autoencoder (Makhzani et al., 2016, Goodfellow et al., 2014).

2.2.2 InfoVAE Optimality

Proposition 2 (Global Optimality of InfoVAE). *Let the data space \mathcal{X} and latent space \mathcal{Z} be continuous, and let the InfoVAE objective be defined with parameters satisfying $\alpha < 1$ and $\lambda > 0$, while holding the mutual information $I_q(x; z)$ at a fixed value I_0 . Then the InfoVAE objective is maximized if and only if the learned generative model exactly matches the data distribution and the encoder recovers the true posterior, that is,*

$$p_\theta(x) = p_{\text{data}}(x) \quad \text{and} \quad q_\phi(z | x) = p_\theta(z | x) \quad \text{for all } x \in \mathcal{X}, z \in \mathcal{Z}.$$

Proof. We outline the two main ingredients of the proof, following (Zhao et al., 2019):

1. **Reconstruction optimality.** The first term in the InfoVAE objective is the expected log-likelihood $E_{p_D(x)} E_{q_\phi(z|x)} [\log p_\theta(x | z)]$. For any fixed encoder q_ϕ , this term is maximized precisely when the decoder satisfies

$$p_\theta(x | z) = q_\phi(x | z) \quad \text{for almost every } (x, z),$$

since matching these conditional distributions yields the highest possible reconstruction log-probability.

2. **Aggregate alignment.** The second key requirement is that the aggregate posterior $q_\phi(z) = \int p_D(x) q_\phi(z | x) dx$ matches the prior $p(z)$ exactly. When this holds and the conditional match from step 1 is in place, the joint distributions coincide:

$$q_\phi(x, z) = p_D(x) q_\phi(z) = p(z) q_\phi(x | z) = p_\theta(x, z).$$

From equality of the joint, it follows that $p_\theta(x) = p_D(x)$ and $q_\phi(z | x) = p_\theta(z | x)$. Under these conditions, all divergence penalties in the InfoVAE objective vanish, and the mutual information term remains at its prescribed value I_0 .

Together, these two conditions perfect reconstruction under the true posterior and exact alignment of the aggregate posterior with the prior are both necessary and sufficient for global maximization of the InfoVAE objective. \square

2.3 Parameters λ and α

In the InfoVAE objective, the two new scalar parameters λ and α serve complementary roles in addressing the failures of standard VAEs. The parameter λ scales the divergence between the aggregated posterior $q_\phi(z)$ and the prior $p(z)$, strengthening the penalty on mismatches and thereby mitigating the tendency of the encoder to overfit individual data points at the expense of a coherent latent distribution. The parameter α weights an explicit mutual-information term between the inputs and latent codes, restoring an incentive for the model to utilize z and avoiding the collapse of the latent space when the decoder is highly expressive (Zhao et al., 2019).

These parameters recover several familiar VAE variants as special cases:

- $\alpha = 0, \lambda = 1$: standard VAE (Kingma and Welling, 2014)
- $\alpha + \lambda - 1 = 0$ (with $\lambda \neq 1$ and $1 - \alpha = \beta$): β -VAE (Higgins et al., 2017)
- $\alpha = 1, \lambda = 1$ (with JS divergence): Adversarial Autoencoder (Goodfellow et al., 2014, Makhzani et al., 2016)

In their experiments, Zhao et al. typically set λ in the range 500 to 1000 and α around 0 to 1, choosing higher λ to strictly enforce posterior–prior alignment and positive α to guarantee non-trivial mutual information. These choices led to markedly improved inference accuracy and representation quality on benchmark datasets, at the cost of introducing additional tuning burden (Zhao et al., 2019).

However, the introduction of λ and α also raises practical concerns. Both parameters require careful tuning to avoid poor-quality optima, and static settings can destabilize training or lead to slow convergence. Techniques such as KL annealing (gradually increasing the weight on the divergence term) have been proposed to ease this sensitivity, drawing on strategies from sequence modeling and language generation (Bowman et al., 2016, Fu et al., 2019). By scheduling λ (or equivalently the KL weight) over the course of training, one can achieve a more stable balance between reconstruction fidelity and latent regularization.

In summary, λ and α provide explicit control over the trade-off between data fit, posterior regularization, and information retention, but they introduce hyperparameter sensitivity that must be managed through careful selection or adaptive scheduling.

2.4 Divergence Families

In the InfoVAE framework, the intractable KL divergence between the aggregated encoder distribution $q_\phi(z)$ and the prior $p(z)$ is replaced by a more general strict divergence $D(q\|p)$ with the property $D(q\|p) = 0$ if and only if $q = p$. This flexibility allows one to trade off statistical efficiency, computational cost, and optimization stability when enforcing latent–prior alignment (Zhao et al., 2019, Genevay et al., 2018). In practice, the choice of divergence substantially shapes training dynamics and final model performance.

Adversarial Training (AAE) Adversarial autoencoders employ a discriminator network to minimize the Jensen–Shannon divergence between samples from $q_\phi(z)$ and $p(z)$ (Makhzani et al., 2016, Goodfellow et al., 2014). This approach can capture complex or multimodal priors and yield expressive latent representations. However, GAN-style training often suffers from instability, mode collapse, and slow convergence when the prior is simple.

Stein Variational Gradient Stein variational methods use functional gradient flows in a reproducing kernel Hilbert space to push a set of particles toward the target distribution, effectively minimizing the KL divergence without an explicit discriminator (Liu and Wang, 2016). This deterministic, particle-based scheme provides principled updates but scales poorly in high dimensions due to its quadratic cost in the number of particles.

Maximum Mean Discrepancy (MMD) MMD is a kernel-based criterion that measures the distance between all moments of two distributions (Dziugaite et al., 2015). Given a positive-definite kernel k , one has

$$D_{\text{MMD}}(q||p) = E_{z,z' \sim p}[k(z, z')] - 2 E_{z \sim q, z' \sim p}[k(z, z')] + E_{z,z' \sim q}[k(z, z')].$$

MMD is simple to compute using minibatches, fully differentiable, and typically yields stable training when matching a well-specified prior.

Sinkhorn Divergence The Sinkhorn divergence interpolates between the Wasserstein distance and MMD by adding an entropic regularization ε to the optimal transport computation (Genevay et al., 2018). It preserves geometric sensitivity to the data manifold with smoother gradients, but requires iterative matrix-scaling steps whose runtime and numerical stability depend on the regularization strength and batch size (Patrini et al., 2020). As $\varepsilon \rightarrow 0$, the Sinkhorn divergence converges to the true Wasserstein distance, while in the limit $\varepsilon \rightarrow \infty$ it recovers a Maximum Mean Discrepancy (MMD) criterion.

Cramér (Energy) Distance Also called the energy distance, this metric is defined as

$$D_{\text{E}}(q, p) = 2 E_{z \sim q, z' \sim p}[\|z - z'\|] - E_{z, z' \sim q}[\|z - z'\|] - E_{z, z' \sim p}[\|z - z'\|],$$

and constitutes a true probability metric (Bellemare et al., 2017). It avoids kernel bandwidth selection but incurs $\mathcal{O}(B^2)$ cost in batch size and is less widely adopted for high dimensional latent spaces (Zhang, 2023).

χ^2 Divergence via Density-Ratio Estimation The χ^2 divergence,

$$D_{\chi^2}(q||p) = E_{z \sim p}[(q(z)/p(z) - 1)^2],$$

can be minimized by training a critic network to estimate the density ratio $q(z)/p(z)$ (Xiao and Han, 2022). This yields potentially tighter alignment than JS or MMD, but reintroduces adversarial-style training and its accompanying challenges (Kato et al., 2023).

Table 1: Divergence measures in InfoVAE

Divergence	Source	Category
Jensen–Shannon (via AAE)	Used in original InfoVAE paper	Adversarial
Stein Variational Gradient (SVGD)	Used in original InfoVAE paper	Kernel-based
Maximum Mean Discrepancy (MMD)	Used in original InfoVAE paper	Kernel-based
Sinkhorn Divergence	Introduced in this seminar study	Probabilistic
Cramér (Energy) Distance	Introduced in this seminar study	Probabilistic
χ^2 Divergence (Density-Ratio)	Introduced in this seminar study	Density-ratio

Overall, selecting among these divergence families endows InfoVAE with a flexible model family. One may favor MMD for efficient, stable training in low dimensions; choose Sinkhorn when geometric fidelity is paramount; or adopt χ^2 for tighter density control in anomaly-sensitive applications. By tailoring the divergence to the task, practitioners can navigate expressivity, stability, and computational trade-offs in a principled manner.

2.5 Alternatives for Latent Utility

Several lightweight VAE variants have been proposed to mitigate the two key failures of classical VAEs poor amortized inference and latent collapse while avoiding the computational overhead of kernel- or adversarial-based regularizers.

Free-Bits VAE Free-Bits VAE addresses the information preference problem by enforcing a minimum information flow in each latent dimension. Specifically, the per-dimension KL divergence is clamped to a threshold δ :

$$\text{KL}_d \leftarrow \max(\text{KL}(q_\phi(z_d | x) \| p(z_d)), \delta).$$

This guarantee of at least δ nats ($1 \text{ nat} = \log_2 e$ bits) per latent coordinate prevents the encoder from collapsing to the prior, ensuring that each dimension carries nontrivial information about the input (Kingma et al., 2016, Chen et al., 2017). By avoiding any additional sampling or kernel computations, Free-Bits offers a direct and efficient mechanism to approximate the mutual-information maximization of InfoVAE.

Batch-Normalized VAE (BN-VAE) BN-VAE tackles the amortized inference failure by stabilizing the encoder’s output distribution. A BatchNorm layer is applied immediately before the network predicts the latent parameters $(\mu, \log \sigma^2)$, which regularizes the scale of the latent code and prevents extreme encoder variances (Zhu et al., 2020). This architectural tweak yields more reliable uncertainty estimates and avoids posterior collapse with virtually no extra hyperparameters, serving as a lightweight surrogate for InfoVAE’s posterior-prior divergence control.

Contrastive Autoencoder (Contrastive AE) Contrastive AE approximates the mutual information term of InfoVAE by adding an InfoNCE loss on normalized latent codes (Parulekar et al., 2023):

$$\mathcal{L}_{\text{NCE}} = - \sum_i \log \frac{\exp(z_i \cdot z_i / \tau)}{\sum_j \exp(z_i \cdot z_j / \tau)}.$$

This contrastive objective encourages distinct, input-dependent representations by pulling each code away from others in the minibatch, thereby avoiding both latent collapse and poor inference (Menon et al., 2022). It captures much of the benefit of explicit mutual-information maximization while incurring only the cost of a simple dot-product normalization.

Although these alternatives do not provide the full theoretical guarantees of InfoVAE’s joint divergence-information objective, they often achieve comparable latent utility in practice. By selectively enforcing minimal information per dimension, stabilizing encoder outputs, or applying contrastive losses, they offer efficient, practical means to address the same inference and representation failures that InfoVAE is designed to solve.

3 Experiments

3.1 Evaluation Metrics

We assess model performance along four complementary axes, following Zhao et al. (2019):

1. Aggregate-posterior alignment

- **Maximum Mean Discrepancy (MMD)**

$$\text{MMD}^2(q_\phi(z) \parallel p(z)) = E_{p,p}[k(z, z')] - 2 E_{q,p}[k(z, z')] + E_{q,q}[k(z, z')],$$

with an RBF kernel k . We compute this on validation minibatches. MMD is a nonparametric, minibatch-friendly measure that captures differences in all moments, making it a robust proxy for how well the aggregate posterior matches the prior regardless of which divergence was used during training (Dziugaite et al., 2015, Zhao et al., 2019).

2. Reconstruction quality

- **BCE & Expected reconstruction log-likelihood** $E_{p_D(x)} E_{q_\phi(z|x)}[\log p_\theta(x | z)]$. Often reported as binary cross-entropy on a held-out test set; higher log-likelihood (or lower BCE) indicates tighter data reconstruction (Kingma and Welling, 2014).
- **Visual inspection of reconstructions.** Qualitative evaluation of sample reconstructions reveals artifacts and mode failures not captured by aggregate scores.

3. Representation utility

- **Semi-supervised classification error.** A linear SVM trained on frozen latent codes for a small labeled subset (e.g. 1000 MNIST labels - 100 for each class) gauges downstream feature quality; lower error implies more informative representations (Kingma and Welling, 2014).

4. Computational efficiency

- **Training time (s).** Measure wall-clock time to train each model for a fixed number of epochs on the same hardware. This captures the overhead of different divergence computations (e.g. MMD vs. Stein vs. Sinkhorn) and regularization schedules. Adversarial and critic network trainings are included.
- **Convergence speed.** Track the number of gradient-update iterations required to reach a target validation MMD threshold. Faster convergence indicates more efficient enforcement of aggregate-posterior alignment.

Together, these metrics ensure balanced evaluation of inference fidelity, representation utility, reconstruction accuracy, and generative quality, avoiding one-dimensional assessments that may mask critical failure modes.

3.2 Experimental Setup

All experiments were implemented in PyTorch 2.0 on a single NVIDIA A100 (40GB) with CUDA 11.7. To guarantee reproducibility, we fixed all random seeds (NumPy, PyTorch) to 0 and list every architecture detail, data split, optimizer setting and hyperparameter. The results presented in this section are based on a single run for each candidate model or setting due to resource constraints.

Data and Data Loaders

- **Dataset:** MNIST $|\mathcal{D}_{\text{train}}| = 50,000$, $|\mathcal{D}_{\text{val}}| = 5,000$, $|\mathcal{D}_{\text{test}}| = 10,000$.
- **Preprocessing:** Pixel intensities in $\{0, 1\}$.
- **Batch size:** $B = 128$.
- **Loader:** shuffle=True (train), num_workers=2, pin_memory=True.

Network Architectures

- **Encoder:**(Kingma and Welling, 2014)
 - Conv2d: input \rightarrow 32channels, 4×4 kernel, stride 2, padding 1
 - ReLU
 - Conv2d: 32 \rightarrow 64channels, 4×4 kernel, stride 2, padding 1
 - ReLU
 - Flatten
 - Two parallel FC heads: $\mu \in R^z$ and $\log \sigma^2 \in R^z$ (Kingma and Welling, 2014)
- **Decoder (Exp 1–2, transposed conv):**(Kingma and Welling, 2014)
 - FC: $R^z \rightarrow R^{64\times 7\times 7}$
 - ConvTranspose2d: 64 \rightarrow 32 channels, 4×4 kernel, stride 2, padding 1
 - ReLU
 - ConvTranspose2d: 32 \rightarrow C channels, 4×4 kernel, stride 2, padding 1
 - Sigmoid
- **Decoder (Exp 3, autoregressive):**(Van Den Oord et al., 2016)
 - FC: $R^z \rightarrow R^{h\times h\times h}$ and reshape to feature map
 - Repeat twice: Upsample $\times 2$ (nearest) \rightarrow Conv2d (3×3 , padding 1) \rightarrow ReLU
 - Stack of $5 \times$ MaskedConv2d layers (first type-A, then type-B)
 - Conv2d (1×1) \rightarrow Sigmoid
(follows masked-CNN design)
- **Latent dimensions tested:** $z \in \{2, 5, 10, 20, 40, 60, 80\}$.

Variants and Key Hyperparameters

- **ELBO-VAE:** $\beta = 1.0$ (standard) (Kingma and Welling, 2014).
- **ReconAE:** $\beta = 0.0$ (no KLD).
- **Free-Bits VAE:** clamp each per-dim KL at $\delta = \gamma = 0.1$ nats (Chen et al., 2017).
- **BN-VAE:** apply BatchNorm on flattened conv features before FC heads (Zhu et al., 2020).
- **AAE:** discriminator = $2 \times \text{FC}(128)$, ReLU; critic $\text{LR}_D = 10^{-4}$ (Makhzani et al., 2016).
- **ContrastiveAE:** InfoNCE with temperature $\tau = 0.1$, cosine-sim projector, cross-entropy on batch-ID targets (Parulekar et al., 2023, Menon et al., 2022).
- **InfoVAE variants (all use $\lambda = 1000$, $\alpha = 1.0$):**
 - MMD: RBF bandwidth $\sigma = 1.0$ (Dziugaite et al., 2015).
 - Stein: kernelized Stein $\sigma = 1.0$ (Liu and Wang, 2016).
 - Sinkhorn: entropic OT blur $\varepsilon = 0.1$ (Genevay et al., 2018, Patrini et al., 2020).
 - Cramér: energy distance (no kernel bandwidth) (Bellemare et al., 2017, Zhang, 2023).
 - χ^2 : density-ratio critic $T : R^z \rightarrow R$ via $\text{FC}(128)\text{--ReLU--FC}(1)$, $\text{LR}_T = 10^{-4}$ (Xiao and Han, 2022, Kato et al., 2023).

Optimization and Schedule

- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$).
- **Learning rates:**
 - Exp 1 (hyperparam. sensitivity): 5×10^{-5} .
 - Exp 2 (divergence substitution): 1×10^{-4} .
 - Exp 3 (model comparison): 1×10^{-3} .
- **InfoVAE scaling:** $\lambda = 1000$, $\alpha = 1.0$.
- **Training length:**
 - Exp 1: 5,000 gradient steps, eval every 20.
 - Exp 2–3: 10 epochs.
- **Validation:** fixed 500 minibatches for MMD computation.

Experiments Overview

- **Dataset & Loader:** MNIST, $B = 128$, `shuffle=True`, 2 workers.
- **Optimizer:** Adam (5×10^{-5} , 1×10^{-3} , or 1×10^{-3} depending on experiment).
- **Architectures:** Conv–Conv–FC encoder; Transpose-Conv or PixelCNN decoder.
- **Variants:** ELBO-VAE, ReconAE, Free-Bits, BN-VAE, β -VAE, AAE, ContrastiveAE, InfoVAE (MMD, Stein, Sinkhorn, Cramér, χ^2).
- **Runs:** Exp 1: 5,000 iterations; Exp 2–3: 10 epochs.

Table 2: Experimental Setup Summary Table

Experiment	Schedule	Key Hyperparameters
1. Hparameter Sensitivity	5,000 iter, eval every 20 steps	Learning rate: LR=5e-5
2. Divergence Substitution	10 epochs, val batch = 128	LR=1e-4, $\lambda = 1000$, $\alpha = 1.0$, $\varepsilon = 0.1$
3. Model Comparison	10 epochs, val batch = 128	LR=1e-3, $\lambda = 1000$, $\alpha = 1.0$, $\delta = 0.1$

3.3 Experiment 1: Hyperparameter Sensitivity

In this study, we quantify how the InfoVAE’s performance depends on the two controller-style hyperparameters, λ (posterior–prior trade-off) and α (information preference), by comparing several fixed schedules against three adaptive schemes. We train InfoVAE-MMD (latent dim $z = 20$, RBF bandwidth $\sigma = 1.0$) for 5,000 gradient steps on binarized MNIST using Adam ($\text{LR} = 5 \times 10^{-5}$), and evaluate the validation MMD every 20 iterations on a held-out set of 500 minibatches (batch size 256). Fixed baselines span combinations $\lambda \in \{500, 1000\}$, $\alpha \in \{0, 0.5, 1.0\}$; adaptive variants adjust one or both hyperparameters to track a target MMD of 4×10^{-3} which is obtained after several training runs as models tend to converge to this level of MMD but smaller values (better MMD) can also work.

Adaptive Schedulers

$$e_t = \text{MMD}_t - \text{MMD}^*, \quad \text{MMD}^* = 4 \times 10^{-3},$$

and clip values to $\lambda_{\min} = 500$, $\lambda_{\max} = 1500$, $\alpha_{\min} = 1.0$, $\alpha_{\max} = 3.0$

We define three update rules:

- **Adaptive λ** (keep α fixed):

$$\lambda_{t+1} = \text{clip}(\lambda_t + K_\lambda e_t, [\lambda_{\min}, \lambda_{\max}]), \quad K_\lambda = 5000.0$$

- **Adaptive α** (keep λ fixed):

$$\alpha_{t+1} = \text{clip}(\alpha_t - K_\alpha e_t, [\alpha_{\min}, \alpha_{\max}]), \quad K_\alpha = 0.7$$

- **Adaptive λ & α :** simultaneously apply both update rules above.

Experimental details

- **Validation batches:** pre-sample 500 fixed minibatches ($B = 256$) from the hold-out set at the start of training.
- **MMD evaluation:** unbiased RBF-MMD computed via pairwise distances (see Sec.3.1).
- **Training loop:** for each schedule, run 5000 steps, update (λ, α) per the chosen rule, record validation MMD at iterations 20, 40, \dots , 5000.
- **Baselines:** Fixed $(\lambda, \alpha) \in \{500, 1000\} \times \{0.0, 0.5, 1.0\}$.

Motivation The performance of InfoVAE relies sensitively on the two newly introduced parameters, α (information preference) and λ (posterior–prior trade-off): fixed settings often require extensive tuning and may either under-regularize leading to latent collapse or over-regularize, yielding poor reconstructions and slow or unstable convergence (Zhao et al., 2019). To address these issues and show potential improvements, we introduce simple proportional-control schedulers that adjust α and/or λ to track a target validation MMD level. By increasing λ whenever the aggregate posterior drifts too far from the prior and by modulating α to emphasize mutual information only as needed, adaptive schedules can maintain a balanced trade-off throughout training, accelerate convergence, and reduce manual hyperparameter search.

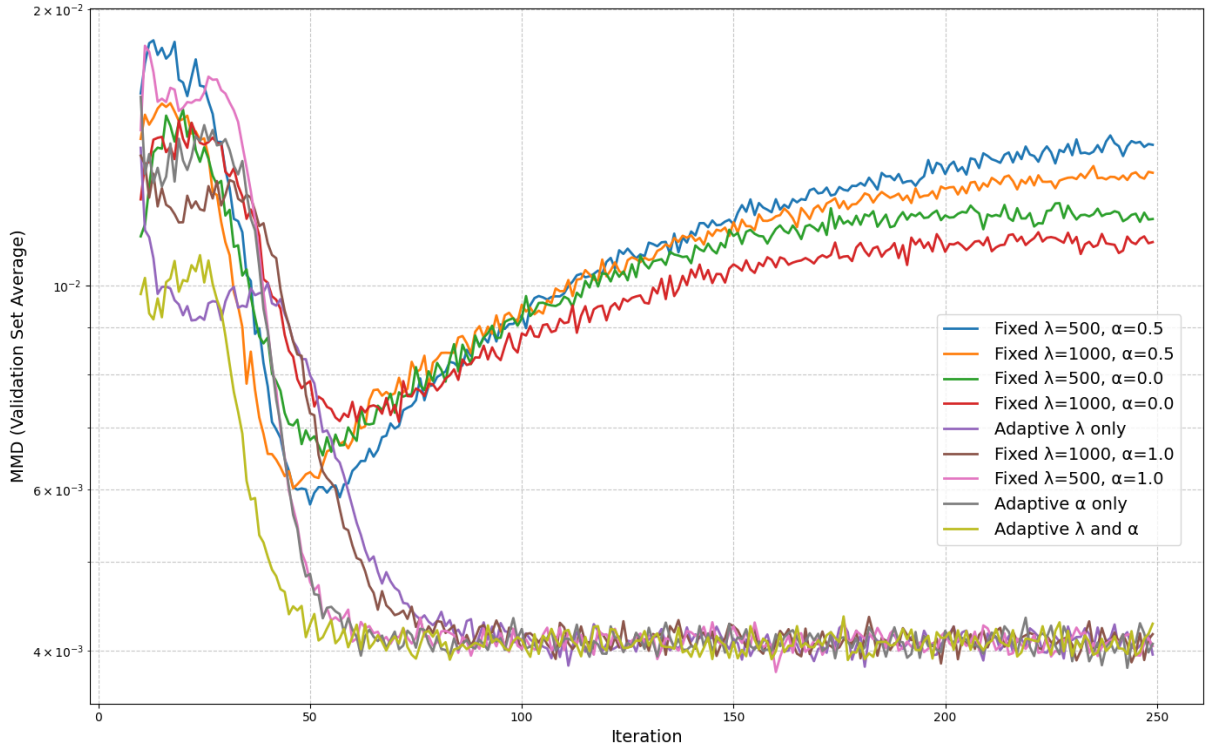


Figure 1: **Validation MMD vs. Training Iterations.** Comparison of fixed α, λ sets with adaptive schedules. Validation MMD at every 20 iteration. Legend ordered by performance, with the best at the bottom and the worst at the top.

In Figure 1, we see that the InfoVAE-MMD models with $\alpha = 0$ and $\alpha = 0.5$ values starts to diverge after approximately 50 iterations. As discussed before, selection of suitable values for these new parameters of InfoVAE is highly significant for training stability and here, we can easily see that even a slight adjustment to a single parameter can cause the training process to fail to converge. The model with combined adaptive schedulers λ and α outperforms others and converges earlier. Adaptive α only scheduler performs on par with the best performer among the fixed parameter settings, $\lambda = 500$ and $\alpha = 1.0$. Fixed $\lambda = 1000$ and $\alpha = 1.0$ and adaptive λ only scheduler converges later than the others.

The results demonstrate that adaptive hyperparameter schedules where λ and α are adjusted dynamically consistently achieve faster convergence and enhanced training stability compared to fixed-parameter baselines. Although the scheduling methods employed here are not yet validated across diverse settings, they yield markedly lower iteration counts to reach the target MMD and reduce oscillatory behavior. This experiment therefore underscores the potential of feedback-driven hyperparameter adaptation to alleviate InfoVAE’s hyperparameter sensitivity and static parameter problems.

3.4 Experiment 2: Divergence Substitution

In this experiment, we investigate how replacing the intractable KL divergence in the InfoVAE objective with alternative strict divergences affects posterior–prior alignment, representation utility, and computational cost. Specifically, we compare five divergence families: RBF-MMD, kernelized Stein discrepancy, entropic Sinkhorn, Cramér energy distance, and the χ^2 divergence (density ratio estimation) via a learned critic under the InfoVAE framework with fixed trade-off parameters ($\alpha = 1.0$, $\lambda = 1000$). For each divergence D , we train InfoVAE $_D$ on binarized MNIST for 10 epochs, sweeping the latent dimension $z \in \{2, 5, 10, 20, 40, 60, 80\}$. We then evaluate: (i) aggregate-posterior mismatch via an unbiased RBF-MMD on 500 held-out minibatches; (ii) downstream feature quality via semi-supervised classification error of a linear SVM trained on 100 labels per class; and (iii) wall-clock training time over the 10 epochs.

All models use Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with learning rate $\text{LR} = 10^{-4}$ and batch size $B = 128$. For RBF-MMD and Stein we set the kernel bandwidth $\sigma = 1.0$, for Sinkhorn the entropic regularization parameter $\epsilon = 0.1$, and for Cramér no bandwidth is required. The χ^2 variant employs a density-ratio critic $T : R^z \rightarrow R$ (two FC layers of 128 units with ReLU) trained jointly with a critic learning rate $\text{LR}_T = 10^{-4}$ as described in Sec. 3.2. Validation batches are fixed once (500 minibatches of size 128), and all other hyperparameters follow the settings in Sec. 3.2.

Motivation The strict divergence term in the InfoVAE objective need not be the KL divergence; any $D = 0 \iff q = p$ measure can enforce aggregate-posterior alignment. Different divergences offer distinct trade-offs between statistical bias, gradient variance and computational overhead. By systematically substituting these divergences and measuring validation MMD, semi-supervised error and training time across latent dimensions, we aim to quantify how divergence choice shapes model stability, latent representation quality and efficiency. This analysis provides practical guidance on selecting a divergence that balances fidelity and scalability in variational autoencoder training.

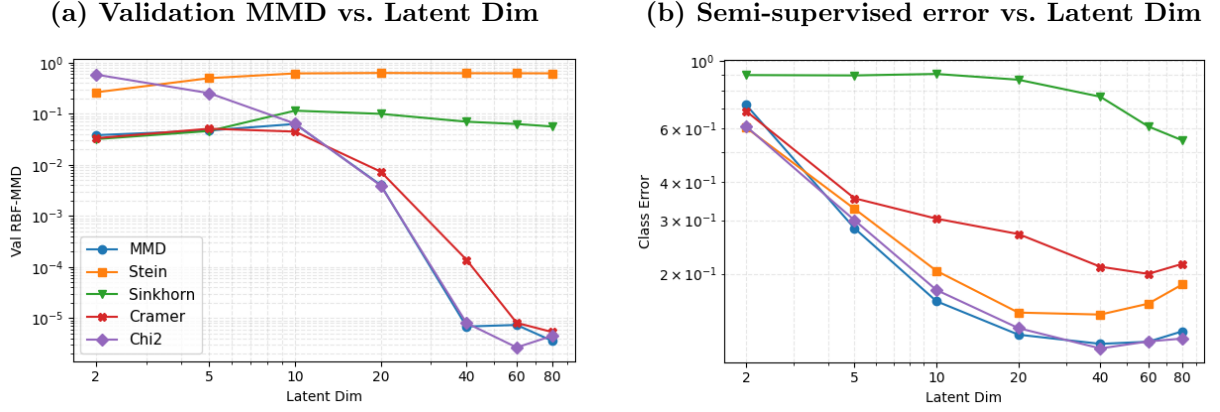


Figure 2: Effect of divergence substitution on (a) aggregate-posterior alignment (MMD) and (b) downstream (semi-supervised) classification error as a function of latent dim.

Figure 2(a) shows that the MMD, Cramér and χ^2 penalties all drive the aggregate posterior $q(z)$ very close to the prior $p(z)$ for latent dimensionalities $z \geq 40$, with MMD² falling below 10^{-5} . In contrast, Stein remains above 10^{-1} and Sinkhorn around 10^{-2} . This alignment similarly translates to latent representation quality: as plotted in Fig. 2(b), the semi-supervised classification error for frozen latent codes falls below 10% for MMD and χ^2 at $z \geq 40$, whereas Stein and Cramér attain moderate performance (10–25% error). Sinkhorn exhibits persistently high error ($\sim 50 - 80\%$) but its convergence behavior becomes noticeably stronger as the latent dimension approaches 80. Comparing Sinkhorn with other methods at even higher latent dimensions would therefore yield more reliable insights. These results confirm a strong coupling between aggregate-posterior alignment and representation quality of latent space which demonstrates that MMD and χ^2 divergences offer superior trade-offs over alternatives such as Cramér, Stein or Sinkhorn in the InfoVAE framework.

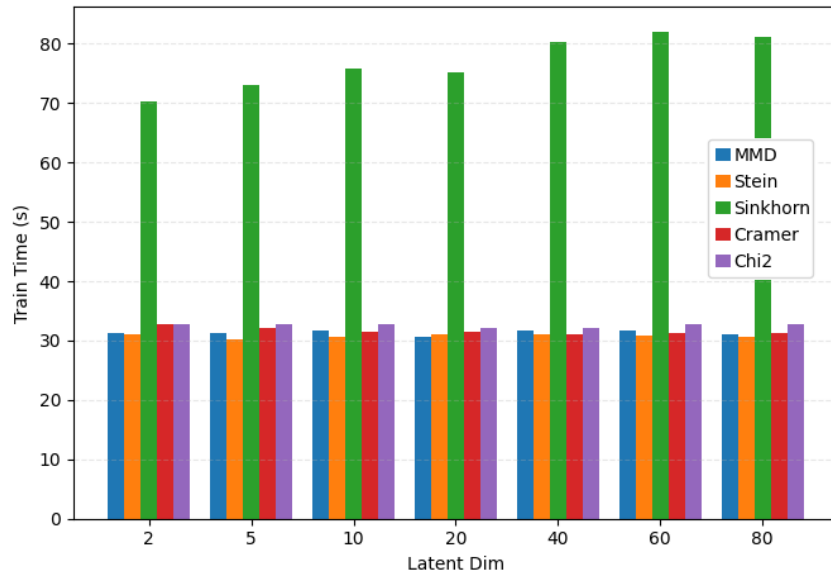


Figure 3: **Training Time vs. Latent Dim.** Total training time (10 epochs) for each divergence variant as a function of latent dimensionality.

Figure 3 compares training times for each divergence penalty across latent dimensions. The Sinkhorn divergence incurs by far the highest computational cost (70–82s for 10 epochs), reflecting its reliance on iterative entropic-optimal transport matrix-scaling (Patrini et al., 2020). In contrast, MMD, Stein and Cramér penalties exhibit nearly identical runtimes (30–32s), since MMD and Cramér distances leverage efficient kernel-based or closed-form moment computations, and Stein’s variational-gradient method scales similarly when implemented via batch-wise kernels (Liu and Wang, 2016). The χ^2 divergence adds only a small overhead (32–33s) owing to its simple critic network architecture (Kato et al., 2023). These results confirm that, while all alternatives enforce strict divergence minimization, only Sinkhorn substantially increases training time, whereas MMD, Stein, Cramér and χ^2 penalties remain practical for large-scale VAE training.

3.5 Experiment 3: Model Comparison

In this experiment, we evaluate and contrast several “lightweight” VAE variants against InfoVAE-MMD to understand the trade-offs between reconstruction fidelity, latent-space alignment, and computational efficiency. Specifically, we consider six models ReconAE (reconstruction loss only ELBO objective), Adversarial Auto Encoder (AAE, see Sec 2.4), InfoVAE-MMD and the new alternatives which are Free-Bits VAE, BN-VAE, and ContrastiveAE (Section 2.5). Each model is trained on binarized MNIST for 10 epochs with the Adam optimizer (learning rate 1×10^{-3}). We vary the latent dimension $z \in \{2, 5, 10, 20, 40, 60, 80\}$ and, for each configuration, record (i) the validation RBF-MMD between $q(z)$ and the prior $p(z)$, (ii) reconstruction error (binary cross-entropy) on the held-out test set, (iii) downstream representation quality via semi-supervised linear SVM classification error (100 labels per class), and (iv) training time for all epochs.

Motivation While InfoVAE-MMD uses an MMD penalty to prevent posterior collapse and control information flow, its kernel-based regularization can be computationally intensive and sensitive to hyperparameters. Alternatively, lighter fixes and recent VAE variants address amortized-inference failures and the information-preference dilemma without introducing additional tuning parameters: Free-Bits VAE clamps each dimension’s KL divergence at a fixed threshold (no kernel computations), BN-VAE applies batch normalization to stabilize encoder outputs (no auxiliary critic or divergence term), AAE employs an adversarial penalty, and ContrastiveAE adds an InfoNCE term. In Experiment 3, we compare these approaches alongside a plain reconstruction AE against InfoVAE-MMD by evaluating aggregate posterior alignment (validation MMD), reconstruction error, downstream semi-supervised accuracy, and training time across latent dimensions. This systematic comparison clarifies which strategies best balance inference fidelity, representation utility, and computational cost in practical VAE training.

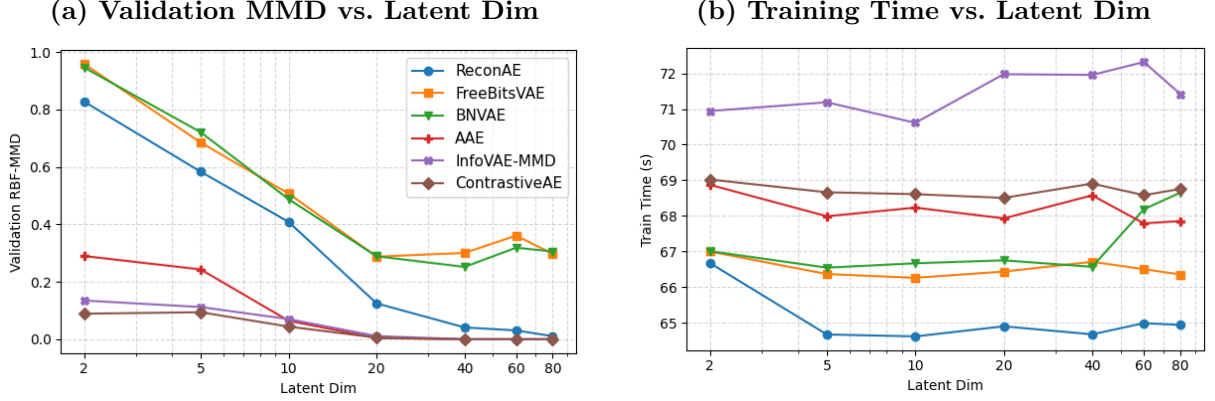


Figure 4: Effect of divergence substitution on (a) aggregate-posterior alignment (MMD) and (b) downstream (semi-supervised) classification error as a function of latent dim.

In Figure 4(a), both InfoVAE-MMD and ContrastiveAE attain near-zero validation MMD for all latent dimensions $z \geq 10$, demonstrating excellent aggregate-posterior alignment $q(z) \approx p(z)$. ReconAE also achieves low MMD at larger z , but exhibits substantially higher discrepancy when $z \leq 10$ due to the absence of any divergence penalty. Free-Bits VAE and BN-VAE reach $\text{MMD} \approx 0.30$ until $z \geq 20$, while AAE yields intermediate alignment for the dimensions 2 and 5 but then it catches InfoVAE and ContrastiveAE converging to zero after $z \geq 20$). These trends confirm that kernel-based (InfoVAE-MMD) and contrastive (InfoNCE) regularizers most effectively enforce $D_{\text{MMD}}(q||p) = 0$ across a wide range of latent capacities, whereas lighter fixes require sufficiently large z to approximate the prior. Moreover, the robust performance of InfoVAE-MMD and ContrastiveAE even at lower dimensionalities underscores their ability to impose strong inductive biases under constrained capacity. In contrast, the fact that ReconAE, Free-Bits VAE and BN-VAE only achieve comparable alignment at higher z highlights the conditional effectiveness of simpler penalty replacements. Overall, these results suggest that kernel-based and contrastive divergences provide more reliable posterior-prior matching, while lighter fixes may demand larger latent spaces or additional tuning to attain similar fidelity.

In Figure 4(b), we compare wall-clock training times over 10 epochs. ReconAE is the fastest (≈ 65 s) since it omits any divergence term. Free-Bits VAE and BN-VAE exhibit near-identical, low training times (≈ 66 – 67 s for 10 epochs), since both variants introduce only lightweight modifications to the base model per-dimension KL clamping in Free-Bits VAE and a single batch-normalization layer before the latent heads in BN-VAE. The slightly higher runtime of BN-VAE can be attributed to the extra forward and backward pass overhead of computing batch statistics (means and variances) and performing the normalization step in each mini-batch, which marginally increases its per-iteration computational cost compared to Free-Bits VAE. AAE and ContrastiveAE, which rely on an adversarial critic or contrastive loss, run slightly slower (≈ 68 – 69 s). InfoVAE-MMD is the slowest (≈ 71 – 72 s) due to repeated kernel evaluations in the MMD penalty. This highlights a clear trade-off: the most precise alignment methods demand approximately 5–10% more training time compared to the lightest variants, a consideration that guides the choice of divergence mechanism in large-scale VAE training.

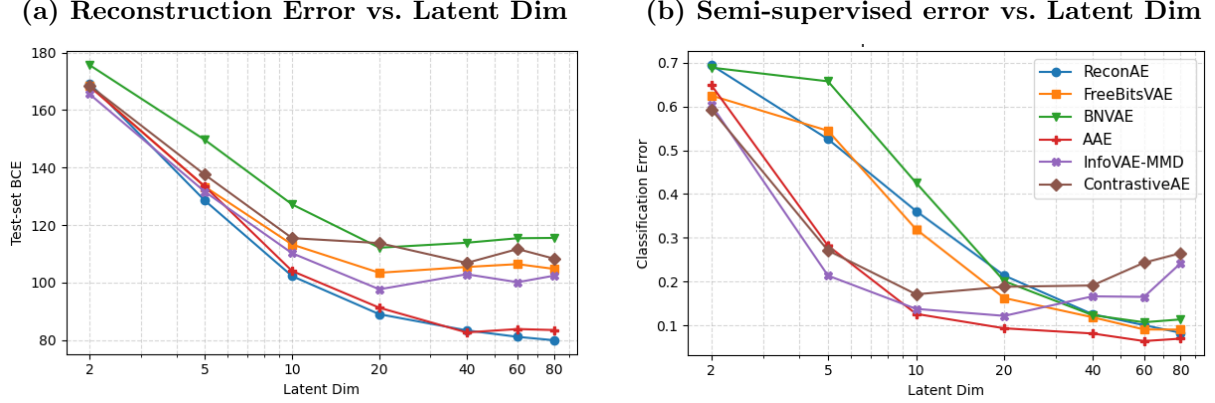


Figure 5: Effect of divergence substitution on (a) aggregate-posterior alignment (MMD) and (b) downstream (semi-supervised) classification error as a function of latent dim.

In Figure 5(a), all models exhibit monotonically decreasing test-set binary cross-entropy as the latent dimension z increases. ReconAE and AAE achieve the lowest reconstruction error across all z , with ReconAE marginally outperforming due to its unconstrained optimization of reconstruction loss. Free-Bits VAE and ContrastiveAE follow InfoVAE closely, indicating that simple KL clamping and InfoNCE regularization introduce only minor reconstruction overhead. InfoVAE-MMD incurs slightly higher error (compared to the best performers) particularly for $z \leq 20$ reflecting its stronger aggregate-posterior matching (MMD ≈ 0 in Figure 4(a)) at the expense of some reconstruction tightness. BN-VAE consistently shows the highest error, suggesting that batch-norm stabilization alone is insufficient to fully reconcile reconstruction and posterior alignment. These reconstruction trends, combined with the training-time results in Figure 4(b), highlight the trade-off between generative fidelity and computational burden: simpler variants (ReconAE, Free-Bits, ContrastiveAE) achieve higher error with minimal overhead (≈ 65 – 67 s), whereas InfoVAE-MMD requires (≈ 71 – 72 s) to enforce stricter alignment.

In Figure 5(b), semi-supervised classification error decreases with increasing latent dimension for all models, but the ordering differs from reconstruction performance. At low dimensions ($z \leq 10$), InfoVAE-MMD, AAE and ContrastiveAE achieve substantially lower error (≈ 0.12 – 0.20) than BN-VAE and Free-Bits VAE (≈ 0.30 – 0.45), indicating that explicit divergence or contrastive penalties yield more informative codes when capacity is limited. As z grows beyond 20, AAE attains the lowest error (≈ 0.10 at $z = 20$ and ≈ 0.06 for $z \geq 60$), closely followed by BN-VAE, Free-Bits VAE and ReconAE (≈ 0.10 – 0.12 for $z \geq 40$). InfoVAE-MMD also achieves low error at the earlier stages but its performance degrades at $z = 80$ (≈ 0.25), suggesting sensitivity to overly large latent spaces.

InfoVAE-MMD and ContrastiveAE achieve the strongest aggregate posterior alignment but incur the highest training cost, whereas ReconAE minimizes reconstruction error with minimal overhead yet fails to align the latent and prior distributions under low capacity. Free-Bits VAE and BN-VAE occupy a middle ground, delivering moderate alignment and efficient training but only matching other models’ reconstruction and downstream performance once the latent space is sufficiently large. AAE combines reliable posterior matching and informative representations under constrained capacity with moderate runtime.

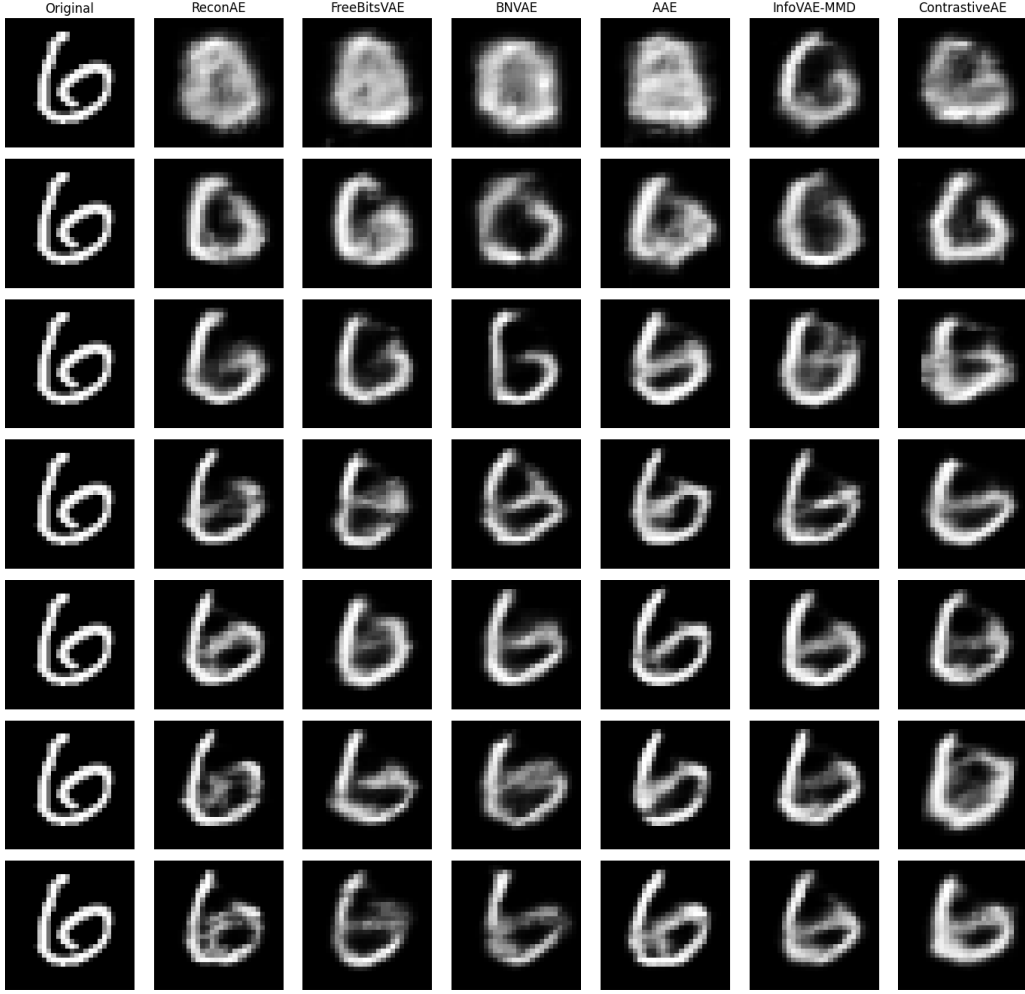


Figure 6: **Reconstructed images (6) per model for different latent dimensions:** Each row shows reconstructions of the digit “6” at $z \in \{2, 5, 10, 20, 40, 60, 80\}$, and each column corresponds to a specific model variant. All models were trained for 10 epochs on binarized MNIST.

Reconstruction Procedure All reconstructions are generated by first sampling $z \sim q_\phi(z | x)$ via the reparameterization trick (i.e. $z = \mu + \sigma \odot \epsilon$) and then passing this sampled z through the decoder. We do not use the posterior mean μ alone for visualization, but rather a single draw from the approximate posterior for each input.

In the reconstruction grid (Figure 6), the perceptual quality of each model’s outputs closely mirrors its quantitative BCE profile (Figure 5(a)). ReconAE produces sharp, digit-faithful reconstructions for $z \geq 10$, consistent with its lowest test-set BCE; this result is expected since ReconAE optimizes reconstruction loss without any regularization. AAE produces reconstructions almost as good as those of ReconAE, reflecting its comparably low BCE error across all latent dimensions. InfoVAE appears slightly blurred at small z , but achieves comparable clarity for $z \geq 20$, in line with its mid-range BCE reduction. BN-VAE remains visibly smeared across all dimensions, matching its persistently higher BCE. ContrastiveAE’s outputs also follow its BCE curve.

Notably, Free-Bits VAE, BN-VAE, InfoVAE and ContrastiveAE exhibit minimal visual improvement and even slight degradation in reconstruction sharpness for $z \geq 40$, paralleling the plateau or slight increase in their BCE errors beyond that point. These observations confirm that lower binary cross-entropy reliably predicts visually sharper reconstructions up to a saturation point determined by latent capacity and regularization strategy. Additional reconstruction grids of all digits for latent dimensions $z = 5, 20$, and 60 are provided in the Appendix, enabling direct visual comparison with the quantitative reconstruction errors reported in Figure 5(a).

Collectively, the results of Experiment 3 demonstrate that no single variant uniformly dominates across all evaluation axes. Kernel-based penalties (InfoVAE–MMD) and contrastive regularization (ContrastiveAE) achieve the strongest aggregate-posterior alignment even at low latent dimensionality, at the cost of modestly increased training time. Adversarial regularization (AAE) and unconstrained reconstruction (ReconAE) excel in reconstruction fidelity and downstream classification, with AAE providing a favorable balance of alignment and utility. Lightweight fixes (Free-Bits VAE, BN-VAE) incur minimal overhead but require larger latent spaces to match the performance of more sophisticated penalties. These trade-offs underscore that the choice of divergence or regularizer should be guided by the specific priorities of a given application whether one values strict posterior matching, reconstruction accuracy, feature usefulness, or computational efficiency.

3.6 Contrast with Original InfoVAE Experiments

In their seminal work, Zhao et al. (2019) evaluated a family of VAEs on binarized MNIST, including the standard ELBO (β -VAE with $\beta = 1$), β -VAE, adversarial autoencoders (AAE), Stein-VAE (kernelized Stein discrepancy) and InfoVAE–MMD. They measured aggregate-posterior alignment via RBF-MMD and covariance log-det metrics, posterior collapse via class-distribution mismatch, semi-supervised classification error, and held-out log-likelihood. Their results consistently favored InfoVAE–MMD and Stein-VAE: both achieved near-zero MMD distance and matched the true prior in latent space, recovered informative latent codes (low classification error) and yielded higher log-likelihood estimates than ELBO or AAE, thus validating the benefit of using the InfoVAE framework.

Building on the original InfoVAE study, we not only compare MMD, Stein and adversarial penalties but also introduce the χ^2 density-ratio divergence, Sinkhorn and Cramér distances, as well as lightweight variants (Free-Bits VAE, BN-VAE, ContrastiveAE). We further investigate InfoVAE’s hyperparameter sensitivity by deploying simple proportional controllers on λ and α , which consistently stabilize training and accelerate convergence compared to static schedules. Across the same alignment (reconstruction and downstream metrics with the added dimension of wall-clock cost) InfoVAE–MMD and χ^2 yield the strongest posterior–prior matching and classification accuracy, and adversarial/contrastive schemes offer competitive utility at lower runtime. Sinkhorn maintains high alignment but with substantial overhead, whereas the lightweight fixes require larger latent capacity to match performance, thus sharpening the original conclusions on the trade-offs between alignment fidelity, representation utility and computational efficiency.

4 Conclusion

The InfoVAE framework directly addresses known limitations of the standard ELBO, which can either under-regularize (yielding inaccurate posteriors) or over-regularize and induce latent collapse. By introducing a posterior–prior trade-off coefficient λ and an information preference weight α , InfoVAE successfully remedies both failure modes. Moreover, InfoVAE unifies β -VAE and adversarial autoencoders under a single objective, enabling principled comparisons and the design of hybrid models. Its flexibility to substitute any strict divergence $D(q||p)$ allows practitioners to tailor regularization to computational constraints or application needs. Extensive experiments in the original work demonstrate that InfoVAE with an MMD penalty matches or outperforms prior methods in terms of reconstruction accuracy, aggregate-posterior alignment, and downstream representation utility (Zhao et al., 2019).

Our extended evaluation uncovers key considerations and alternative paths. InfoVAE’s performance is highly sensitive to the choice of $\{\lambda, \alpha\}$; simple adaptive schedulers on these hyperparameters markedly stabilize training and accelerate convergence compared to fixed schedules. This finding emphasizes that automated hyperparameter control can be as impactful as model architecture in practical applications. The divergence term itself strongly influences stability and quality, with MMD regularization remaining consistently effective and the χ^2 density-ratio penalty also achieving competitive alignment and utility with modest overhead. Hence, selecting the appropriate divergence becomes a central design decision. Finally, lightweight variants such as ContrastiveAE and BN-VAE deliver comparable latent-space fidelity with fewer hyperparameters and substantially lower computational cost, suggesting practical substitutes when strict divergence minimization is computationally prohibitive. These efficient alternatives broaden the reach of InfoVAE-style models to resource constrained or rapidly prototyped settings without sacrificing representation quality.

Future Research Although this study sheds light on the interplay between divergence choice, hyperparameter adaptation and computational cost in VAEs, several directions could further strengthen and extend these insights. One promising area is the design of more sophisticated, data-driven schedulers for λ and α , potentially learned alongside the model or via reinforcement learning. It would also be valuable to assess an even broader range of divergences, including higher-order kernels and transport-based costs, particularly at larger latent capacities where their trade-offs may differ. Moreover, replicating these experiments on more complex domains such as natural images, text or multimodal data will test the generality of our conclusions. Finally, investigating multimodal VAE architectures within the InfoVAE framework could reveal how divergence mechanisms and adaptive hyperparameters interact in richer, heterogeneous representation learning settings.

A Appendix

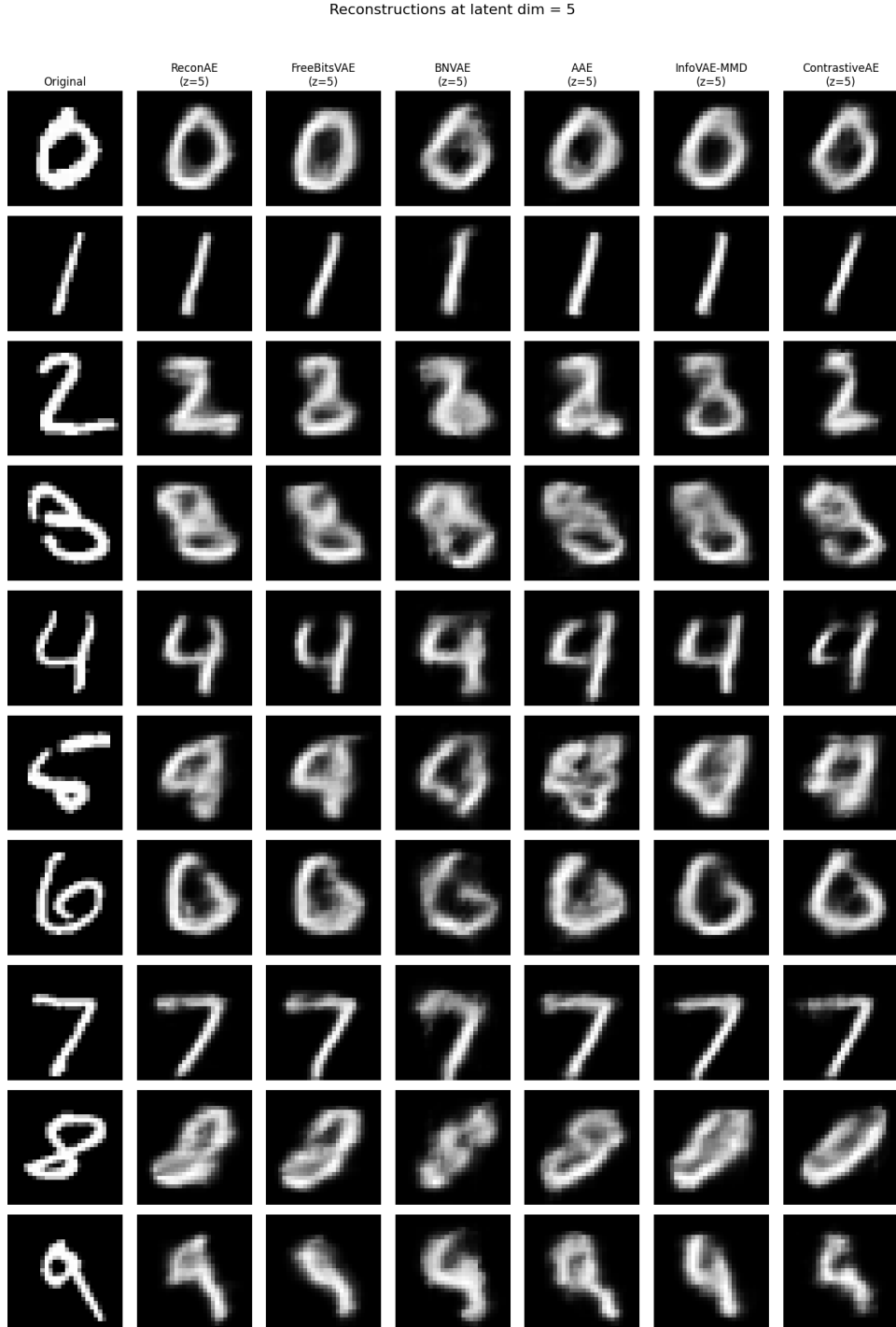


Figure 7: Reconstructions of all ten digits (rows) produced by each model variant (columns) at latent dimension $z = 5$. From left to right: Original inputs, ReconAE, Free-Bits VAE, BN-VAE, AAE, InfoVAE-MMD, ContrastiveAE.

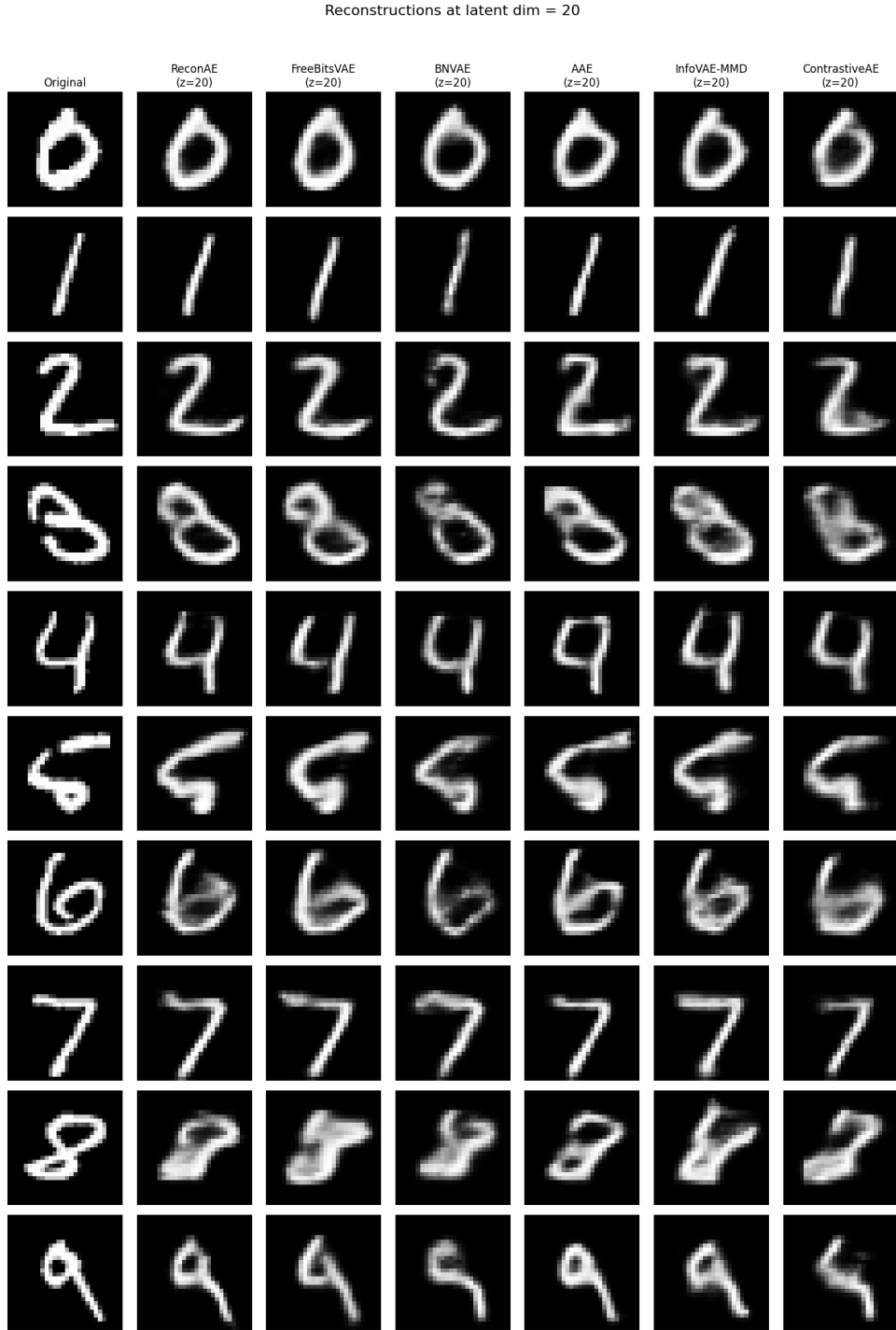


Figure 8: Reconstructions of all ten digits (rows) produced by each model variant (columns) at latent dimension $z = 20$. From left to right: Original inputs, ReconAE, Free-Bits VAE, BN-VAE, AAE, InfoVAE-MMD, ContrastiveAE.

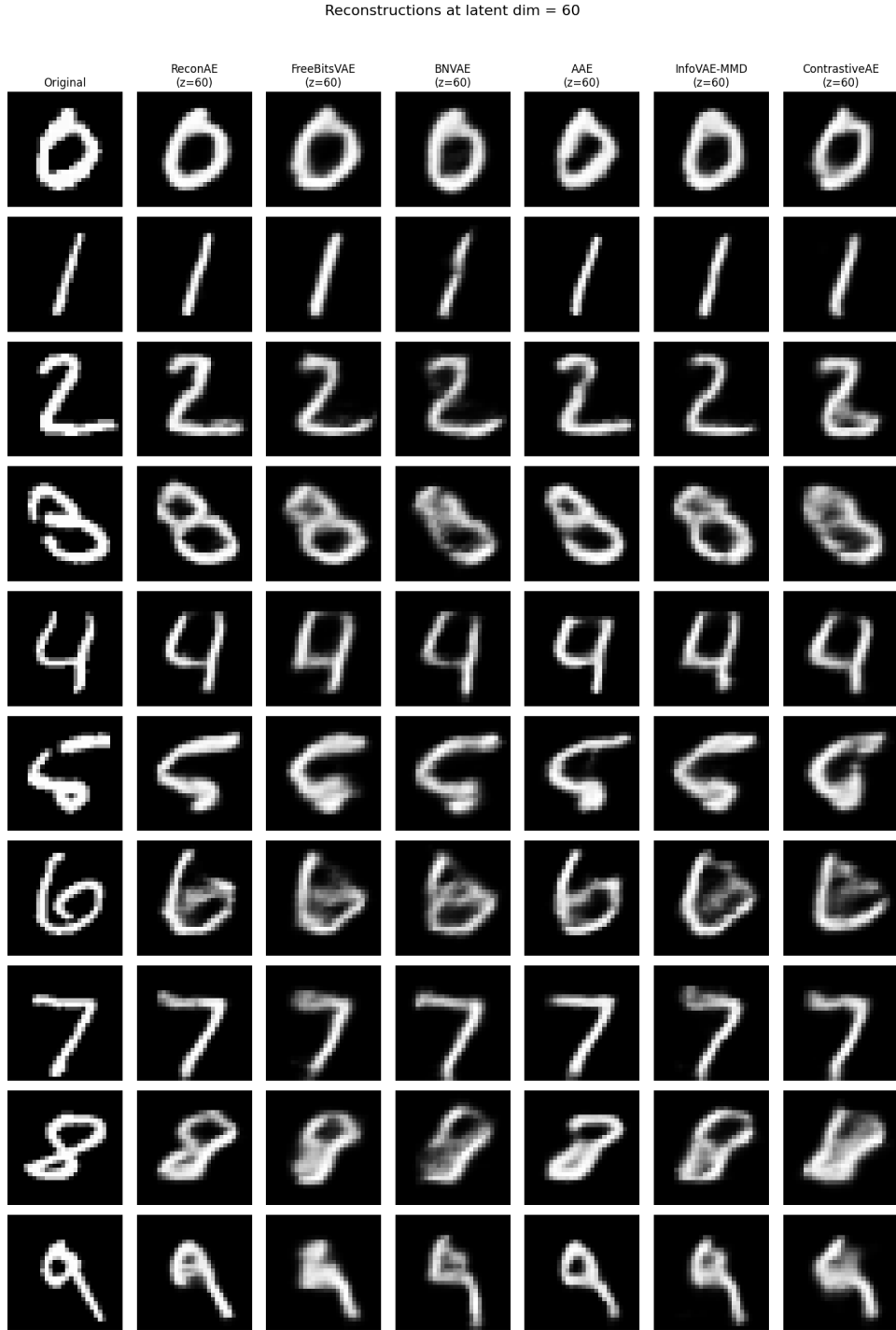


Figure 9: Reconstructions of all ten digits (rows) produced by each model variant (columns) at latent dimension $z = 60$. From left to right: Original inputs, ReconAE, Free-Bits VAE, BN-VAE, AAE, InfoVAE-MMD, ContrastiveAE.

B Electronic appendix

The code and supplementary materials for this study are available at:

<https://github.com/ekcmert/infovae-seminar>

Data, code and figures are provided in electronic form.

References


- Bellemare, M., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S. and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients.
- Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R. and Bengio, S. (2016). Generating sentences from a continuous space, *CoNLL 2016* p. 10.
- Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I. and Abbeel, P. (2017). Variational lossy autoencoder, *International Conference on Learning Representations*.
- Dziugaite, G. K., Roy, D. M. and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization, *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 258–267.
- Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A. and Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating KL vanishing, in J. Burstein, C. Doran and T. Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 240–250.
URL: <https://aclanthology.org/N19-1021/>
- Genevay, A., Peyre, G. and Cuturi, M. (2018). Learning generative models with sinkhorn divergences, in A. Storkey and F. Perez-Cruz (eds), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Vol. 84 of *Proceedings of Machine Learning Research*, PMLR, pp. 1608–1617.
URL: <https://proceedings.mlr.press/v84/genevay18a.html>
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative Adversarial Nets, *Advances in Neural Information Processing Systems* **27**.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A. (2017). β -VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, *International Conference on Learning Representations*.
- Kato, M., Imaizumi, M. and Minami, K. (2023). Unified perspective on probability divergence via maximum likelihood density ratio estimation: Bridging kl-divergence and integral probability metrics, *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*.
- Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I. and Welling, M. (2016). Improved variational inference with inverse autoregressive flow, *Advances in neural information processing systems* **29**.
- Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Liu, C. and Wang, X. (2025). Doubly robust conditional VAE via decoder calibration: An implicit KL annealing approach, *Transactions on Machine Learning Research*.
URL: <https://openreview.net/forum?id=VIkycTWDWo>
- Liu, Q. and Wang, D. (2016). Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm, *Advances in Neural Information Processing Systems* **29**.
- Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. and Frey, B. (2016). Adversarial Autoencoders, *International Conference on Learning Representations*.

- Menon, S., Blei, D. and Vondrick, C. (2022). Forget-me-not! contrastive critics for mitigating posterior collapse, in J. Cussens and K. Zhang (eds), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, Vol. 180 of *Proceedings of Machine Learning Research*, PMLR, pp. 1360–1370.
URL: <https://proceedings.mlr.press/v180/menon22a.html>
- Parulekar, A., Collins, L., Shanmugam, K., Mokhtari, A. and Shakkottai, S. (2023). Infonce loss provably learns cluster-preserving representations, *The Thirty Sixth Annual Conference on Learning Theory*, PMLR, pp. 1914–1961.
- Patrini, G., van den Berg, R., Forré, P., Carioni, M., Bhargav, S., Welling, M., Genewein, T. and Nielsen, F. (2020). Sinkhorn autoencoders, in R. P. Adams and V. Gogate (eds), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, Vol. 115 of *Proceedings of Machine Learning Research*, PMLR, pp. 733–743.
URL: <https://proceedings.mlr.press/v115/patrini20a.html>
- Van Den Oord, A., Kalchbrenner, N. and Kavukcuoglu, K. (2016). Pixel recurrent neural networks, *International conference on machine learning*, PMLR, pp. 1747–1756.
- Xiao, Z. and Han, T. (2022). Adaptive multi-stage density ratio estimation for learning latent space energy-based model, *Advances in Neural Information Processing Systems* **35**: 21590–21601.
- Zhang, R. (2023). Cramer type distances for learning gaussian mixture models by gradient descent.
URL: <https://arxiv.org/abs/2307.06753>
- Zhao, S., Song, J. and Ermon, S. (2019). InfoVAE: Balancing learning and inference in variational autoencoders, *AAAI Conference on Artificial Intelligence* **33**(01): 5885–5892.
- Zhu, Q., Bi, W., Liu, X., Ma, X., Li, X. and Wu, D. (2020). A batch normalized inference network keeps the KL vanishing away, in D. Jurafsky, J. Chai, N. Schluter and J. Tetreault (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 2636–2649.
URL: <https://aclanthology.org/2020.acl-main.235/>

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, 25.07.2025



Mert Ekici