MIND THE GAPS: MULTISCALE DISENTANGLEMENT THROUGH INFOMAX AUTO-ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A primary goal for unsupervised learning is to discover disentangled representations of latent factors. Traditionally, such disentanglement is obtained by enforcing constraints on the prior distribution over latents, but this typically entails a loss of reconstruction quality. Here, we show that such constraints are unnecessary. Leveraging an explicit connection between mutual information and MMSE estimators, we develop a novel objective for information maximization which leads to disentangled representations without reliance on prior-matching constraints. Under specific assumptions on the independence, separability, and scale separation of the generative factors, we prove that this objective, paired with a factorized Gaussian encoder, yields multiscale disentanglement. Numerical experiments on both synthetic and natural image datasets support this theoretical prediction. More generally, our approach shows that the combination of classic information-theoretic principles with the expressive power of modern machine learning offers a fruitful path for the development of new models and algorithms.

1 INTRODUCTION

To evaluate the behavioral relevance of their observations, intelligent agents must be able to decompose complex sensory observations into abstract and semantically meaningful factors of variation, a capability known in machine learning as disentangling (Bengio et al., 2014). Disentangled representations offer several advantages, including improved interpretability, enhanced generalization, and stronger transfer capabilities across related tasks (Higgins et al., 2017). However, learning disentangled representations without explicit supervision is challenging.

Unsupervised learning of disentangled representations has primarily been studied using *variational autoencoders* (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014), which optimize a well-defined evidence lower bound (ELBO), with factorization of the latents encouraged via a KL regularizer. Despite its ubiquity, this solution hides an undesirable tradeoff. While stronger regularization can promote alignment between latent coordinates and underlying factors of variation (Higgins et al., 2017), it also couples the mutual information between data and latents, and a *prior-matching* term, penalizing divergence between the aggregated posterior and the prescribed prior (Hoffman & Johnson, 2016). As a result, increasing β simultaneously encourages compression and closer agreement with the prior, intertwining their effects (Alemi et al., 2018; Mathieu et al., 2019). Substantial subsequent work has focused on the role of a factorized prior in penalizing the total correlation (TC) of the aggregated posterior (Kim & Mnih, 2018; Chen et al., 2018), but the contribution of compression itself to disentanglement remains unclear.

A complementary line of work examines disentanglement through the inductive biases imposed by the encoder. VAE representations are typically parameterized as conditional Gaussians with diagonal covariance matrices, which favors decoders with locally orthogonal Jacobian columns, and provides a geometric account of the alignment of latent coordinates with distinct factors of variation (Rolínek et al., 2019; Reizinger et al., 2022). This mechanism encourages the factorization of the conditional encoder distribution, whereas prior matching constrains the aggregated posterior. But their relative contributions are ambiguous: when does disentanglement depend on a factorized prior, and when can it emerge from encoder factorization and compression? Empirical sensitivity to architecture, dataset, and random seed further underscores the need to understand the interaction of these constraints with the structure of the data (Locatello et al., 2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

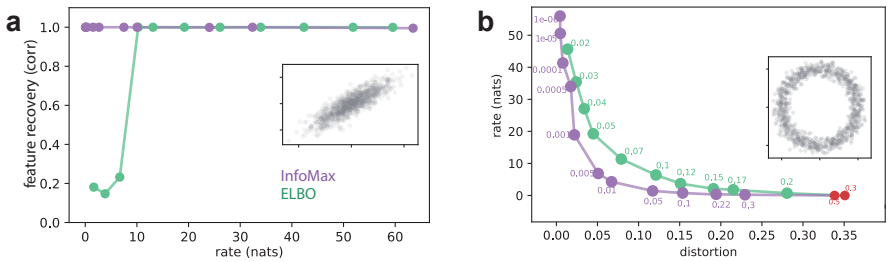

Figure 1: **a.** Disentanglement in models trained on ELBO vs. InfoMax as a function of degree of compression (rate) for two example 2D datasets (inset). Disentanglement is measured as the max correlation between latent axes and factors, and rate as the mutual information between latents and observations. **b.** Rate-distortion tradeoff. The prior matching term in the ELBO induces an increase in distortion for a given rate. Red points indicate posterior collapse (measured MI ≤ 0.01 nats).

Diffusion-based representation learning offers a useful framework for examining these interactions. These models replace the conventional Gaussian decoder with a conditional diffusion model that reconstructs data through denoising across noise levels (Zhang et al., 2022; Preechakul et al., 2022; Yang et al., 2023; Mittal et al., 2023; Wang et al., 2023; Hudson et al., 2024; Kim et al., 2024; Lyo et al., 2025; Proszewska et al., 2025; Jiang & Cai, 2025). Several implementations have reported improved disentanglement and a more favorable trade-off between reconstruction quality and disentanglement, compared to than that obtained with β-VAEs (Preechakul et al., 2022; Kim et al., 2024; Lyo et al., 2025). Here, we identify a mechanism that helps explain these gains: denoising across noise levels exposes differences in the scales of generative factors, which interact with a factorized Gaussian encoder to promote disentangled representations under compression. Under explicit assumptions on factor independence, separability, and scale separation, we show that this interaction can yield disentanglement without requiring the aggregated posterior to match a prescribed prior.

We develop a constrained information maximization (InfoMax) objective that removes the prior-matching penalty and instead exploits the interaction between encoder factorization and the multi-scale structure of the data. The information-theoretic terms are made tractable through the I-MMSE relationship, which explicitly connects mutual information to minimum mean squared error, allowing optimization of the objective using a denoiser optimized over many noise levels. Experiments on synthetic datasets with known generative factors demonstrate improved reconstruction and disentanglement under strong compression relative to ELBO training, while experiments on naturalistic images reveal rich learned semantic structure. Finally, the noise-dependent information carried by each latent automatically defines the spatial scale of its associated factor.

2 Preliminaries

2.1 Latent Variable Models and Variational Autoencoders

Latent variable models aims to explain observed data using unobserved generative factors. A model $p_d(x, z)$ is specified by the product of a prior $m(z)$ over the latents and a conditional distribution $p_d(x|z;\theta)$ over the observations. The model is fitted by maximizing the marginal likelihood of the data, and the posterior provides a representation of each observation. For expressive high-dimensional models, however, these computations are typically intractable.

Variational autoencoders (VAEs) address these difficulties by introducing an amortized approximate posterior $p_e(z|x;\phi)$ and jointly optimizing the encoder and generative model through the evidence lower bound (ELBO) (Kingma & Welling, 2013; Rezende et al., 2014):

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; x) = \mathbb{E}_{p_e(z|x;\phi)}\left[\log p_d(x|z;\theta)\right] - \text{KL}\left(p_e(z|x;\phi) \| m(z)\right) \leq \log p_d(x). \tag{1}$$

The encoder network predicts the variational parameters for each observation, while a separate decoder network parameterizes the likelihood. The prior is typically chosen to be a simple factorized distribution (e.g., a standard Gaussian).[1]

[1]We use p_e and p_d to denote the encoder and decoder throughout, suppressing parameters when convenient.

2.2 Disentanglement and Prior Regularization

Although likelihood training provides a principled way to fit the data distribution, it does not by itself ensure that the latent representation captures the underlying factors of variation (Alemi et al., 2018). Disentanglement implies that changes in an individual generative factor should correspond to changes along a single latent coordinate (Higgins et al., 2017). Encouraging such structure is achieved in β-VAE by adjusting the strength of the ELBO's KL regularizer (Higgins et al., 2017):

$$\mathcal{L}_\beta = \mathbb{E}_{p(x)p_e(z|x)}\left[\log p_d(x|z)\right] - \beta\,\mathbb{E}_{p(x)}\left[D_{\mathrm{KL}}\left(p_e(z|x)\|m(z)\right)\right], \tag{2}$$

where $p(x)$ denotes the data distribution, and $\beta = 1$ recovers the data-averaged ELBO. This expression can be rewritten as Hoffman & Johnson (2016)

$$\mathbb{E}_{p(x)}\left[D_{\mathrm{KL}}\left(p_e(z|x)\|m(z)\right)\right] = I_e(X;Z) + D_{\mathrm{KL}}\left(p_e(z)\|m(z)\right), \tag{3}$$

where mutual information I_e is evaluated under $p_e(x,z)$, and the subscript indicates the dependence on the encoder. Thus, increasing β simultaneously penalizes the information retained by the representation and its aggregate mismatch with the prescribed prior.

For a factorized prior, the aggregate-matching penalty further decomposes into the total correlation of the latent coordinates and the mismatch of their individual marginals relative to the prior, which motivates methods that specificnally penalize total correlation (Kim & Mnih, 2018; Chen et al., 2018). However, statistical independence alone does not suffice for identifying the underlying generative factors, requiring further assumptions or inductive biases (Locatello et al., 2019). Separating information compression from prior matching allows us to ask which properties of the representation follow from information constraints and which depend on the prescribed latent distribution.

2.3 InfoMax and Rate-Distortion

The InfoMax principle provides a framework for understanding many traditional machine learning, going back to independent component analysis and clustering (Bell & Sejnowski, 1995; Slonim et al., 2005). Infomax seeks encodings that maximizes the mutual information between observations X and representations Z subject to resource constraints(Linsker, 1988). Unlike explicit latent variable models, it does not require a prescribed generative prior. Yet, the constraints are essential: mutual information is invariant to invertible transformations, so it alone cannot specify how latents should be organized across coordinates.

Rate-distortion theory provides a complementary formulation of the same problem: determine the minimum information required to reconstruct a source to a specified fidelity (Shannon, 1948; Cover & Thomas, 2006). For an encoder $p_e(z|x)$ and a reconstruction $\hat{X}$ obtained from Z, the distortion is given by reconstruction quality $D = \mathbb{E}[d(X,\hat{X})]$ with information rate $R = I(X;Z)$. A common choice is squared-error distortion, $d(x,\hat{x}) = \|x - \hat{x}\|^2$. The loss becomes

$$\mathcal{J}_{\mathrm{RD}} = D + \beta R, \tag{4}$$

where β controls the relative cost of retaining information.

This perspective clarifies the relationship between information constraints and VAE training (Alemi et al., 2018). Writing $D_{\mathrm{rec}} = -\mathbb{E}_{p_e(x,z)}[\log p_d(x|z)]$, Eq. 3 gives

$$-\mathcal{L}_\beta = D_{\mathrm{rec}} + \beta I_e(X;Z) + \beta D_{\mathrm{KL}}\left(p_e(z)\|m(z)\right). \tag{5}$$

For a Gaussian decoder with fixed isotropic covariance, D_{rec} is proportional to the expected squared error of the decoder mean, up to an additive constant. No Gaussian assumption on the encoder is needed for this equivalence. In this light, β-VAEs combine this reconstruction–information tradeoff with an additional aggregate-prior-matching penalty. InfoMax and rate–distortion emphasize information retention and compression, respectively, but both argue against the necessity of this additional penalty. Their practical use requires tractable estimates of information theoretic quantities that can furthermore be differentiated with respect to the encoder parameters.

How much does the presence of an additional prior regularizer matter in practice? Figure 1 illustrates that, even with a diffusion decoder, ELBO training with an isotropic Gaussian latent prior can impair disentanglement when the underlying factors vary over different spatial scales (see Appendix A.1 for

details). The rotational symmetry of the prior provides no preference for an axis-aligned representation (Mathieu et al., 2019). Our experiments further show that the tension between reconstruction quality and disentanglement becomes especially pronounced when the topology of the underlying factor space (ring) is incompatible with that of the latent space prescribed by the prior (gaussian). From a rate-distortion perspective, these observations further motivate separating the cost of retaining information from the additional constraint of matching a prescribed latent distribution.

2.4 Variational Bounds on Mutual Information

Mutual information is generally difficult to evaluate for high-dimensional distributions known only through samples, motivating a range of estimators and approximations (Paninski, 2003; Kraskov et al., 2004; Poole et al., 2019). For our purposes, we need to optimize mutual information of a representation, rather than merely estimate its value. This motivates differentiable objectives, including variational bounds that can be optimized jointly with the encoder.

The Barber & Agakov (2003) bound uses the conditional distribution induced by an encoder. Let $p_e(x|z)$ denote the conditional associated with $p_e(x, z) = p(x)p_e(z|x)$, and let the decoder $p_d(x|z)$ approximate it. For continuous X with finite differential entropy, (derivation in Appendix A.2)

$$I_e(X; Z) \geq h(X) + \mathbb{E}_{p_e(x,z)} [\log p_d(x|z)] \equiv \widetilde{I}(X; Z). \tag{6}$$

Since $h(X)$ is independent of the encoder and decoder, maximizing this bound amounts to maximizing the expected reconstruction log-likelihood. Its variational gap is

$$I_e(X; Z) - \widetilde{I}(X; Z) = \mathbb{E}_{p_e(z)} [D_{\mathrm{KL}} (p_e(x|z) \| p_d(x|z))] . \tag{7}$$

The bound is tight if and only if the decoder matches the encoder-induced conditional for $p_e(z)$. A restricted decoder family can prevent this equality even with optimal training, limiting how closely the bound tracks mutual information. This decoder approximation error is distinct from the aggregate-prior mismatch in Eq. 3: removing the prior matching term does not eliminate the decoder variational gap. We argue for the need of a tractable formulation of mutual information that accommodates complex conditional distributions. The I-MMSE relation provides such a formulation, computed using flexible regression models.

2.5 Mutual Information Estimation through Denoising

The I-MMSE relation expresses the mutual information in a Gaussian channel in terms of the minimum mean squared error of reconstructing its input (Guo et al., 2005). Consider noisy observations

$$X_\gamma = \sqrt{\gamma}\, X + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{8}$$

where $\gamma \geq 0$ is the channel's SNR parameter and ϵ is independent of (X, Z). The I-MMSE relation states that

$$\frac{d}{d\gamma} I(X; X_\gamma) = \frac{1}{2} \operatorname{mmse}(\gamma), \tag{9}$$

where

$$\operatorname{mmse}(\gamma) = \min_f \mathbb{E} \left[\|X - f(X_\gamma, \gamma)\|^2 \right] . \tag{10}$$

The optimal denoiser is the posterior mean $f^*(x_\gamma, \gamma) = \mathbb{E}[X|X_\gamma = x_\gamma]$. Although this function is generally unavailable analytically, it can be approximated by training a denoising network on pairs of clean and noisy observations, without evaluating or sampling from the posterior explicitly.

Conditioning on the representation gives the corresponding optimal denoiser $f_Z^*(x_\gamma, \gamma, z) = \mathbb{E}[X|X_\gamma = x_\gamma, Z = z]$, with error

$$\operatorname{mmse}(\gamma|Z) = \mathbb{E} \left[\|X - f_Z^*(X_\gamma, \gamma, Z)\|^2 \right] . \tag{11}$$

Here, the expectation includes the distribution of Z. Under finite second moments and finite $I(X; Z)$, applying I-MMSE with and without this conditioning and integrating over SNR yields

$$\begin{aligned} I(X; Z) &= \frac{1}{2} \int_0^\infty [\operatorname{mmse}(\gamma) - \operatorname{mmse}(\gamma|Z)] \, d\gamma \\ &= \frac{1}{2} \int_0^\infty \mathbb{E} \left[\|f_Z^*(X_\gamma, \gamma, Z) - f^*(X_\gamma, \gamma)\|^2 \right] d\gamma. \end{aligned} \tag{12}$$

The second equality follows from the orthogonality principle for conditional expectations (Kong et al., 2024; Laquitaine et al., 2025). Thus one way to view the information carried by Z is the (integrated) improvement in optimal denoising obtained by conditioning on it. An equivalent derivation through marginal and conditional entropies is given in Appendix A.3 (Kong et al., 2023).

These identities are exact for the posterior-mean denoisers. In practice, learned denoisers and numerical integration introduce approximations which may not preserve the equality between the two expressions in Eq. 12; this separates the information-theoretic identity from its implementation.

Previous work has used denoising-based information estimates to study conditioning in diffusion models, sensory coding, and mutual information estimation (Kong et al., 2024; Laquitaine et al., 2025; Yu et al., 2025). Here, we use these identities as the basis of a learning objective, optimizing the encoder so that its representation provides informative conditioning across noise levels. This supplies the computational foundation for the constrained InfoMax Autoencoder (IMAE).

3 The InfoMax Auto-Encoder

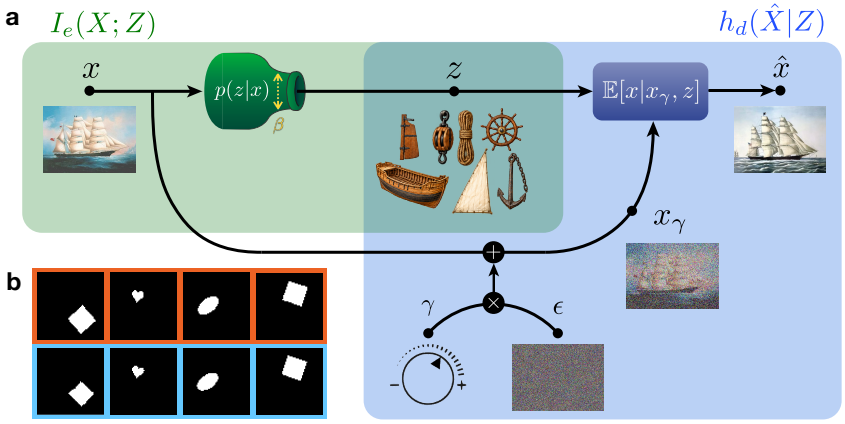

Figure 2: **a.** IMAE is comprised of a probabilistic encoder (green) and diffusion decoder (purple). The encoder maps a clean image to a latent representation, which guides reconstruction from a noisy version of the same image. **b.** Conditional image generation (light blue) based on a target image (orange) from the dSprites dataset.

The InfoMax Autoencoder learns representations through their contribution to denoising across noise levels (Fig. 2a). Given an image $x \in \mathbb{R}^d$, a probabilistic encoder produces a latent sample $z \sim p_e(z \mid x; \phi)$, with $z \in \mathbb{R}^k$ (Fig 2a, green shade). In parallel, a Gaussian channel corrupts the image (purple shade), $x_\gamma = x + \sqrt{\gamma}\,\epsilon,\ \epsilon \sim \mathcal{N}(0, I_d),\ \gamma > 0$, where γ denotes the noise variance. Since the decoder uses both x_γ and z to estimate the clean image, the latent supplies information that cannot be recovered from the noisy image. We aim to optimize the exact rate-distortion Lagrangian:

$$\mathcal{L}(\theta, \phi) = \underbrace{D(\theta, \phi)}_{\text{distortion}} + \beta\ \underbrace{R(\phi)}_{\text{rate}} = h(X|Z) + \beta\, I_e(X; Z), \tag{13}$$

Directly optimizing this information-theoretic objective has historically not been possible without assuming either a simple parametric family for the decoder or involving an explicit prior (Eq. 5). Our contribution is to provide a tractable formulation of the objective via inversion of the I-MMSE (Eq. 9) that can be implemented using only two feedforward neural networks.

Distortion The Barber-Agakov bound (Eq. 6) relies on the parametric decoder family $p_d(x|z)$ being expressive enough to capture the true posterior $p(x|z)$. If we use a simple decoder parametric family, as in β-VAEs, the bound can be loose and lead to a poor estimate of the conditional entropy, in turn leading to a poor representation of data. By contrast, for a fixed variable Z, the I-MMSE provides an exact estimate of conditional entropy (Eq. 12). By *inverting* this relation, we can learn the encoding $z \sim p_e(z|x)$ that best minimizes the conditional entropy.

The I-MMSE formulation of the conditional entropy (Eqs. 9–12) is given as

$$\tilde{h}(X \mid Z) = \frac{1}{2}\int_0^\infty \left[\frac{1}{\gamma^2}\,\mathrm{mmse}_{\theta,\phi}(\gamma \mid Z) - \frac{d}{\gamma(1+\gamma)}\right] d\gamma + \text{const}, \tag{14}$$

where $\mathrm{mmse}_{\theta,\phi}(\gamma|Z)$ is the estimate of Eq. 11 using a Z-conditional denoiser $\hat{x}_{\theta,\phi}(x_\gamma, \gamma, z)$ trained to estimate x from the tuple (x_γ, z). The reference term is independent of the model parameters and makes the entropy expression well defined under usual regularity and integrability conditions.

How do we use this to train the encoder $p_e(z|x)$? We decompose the conditional denoiser into an unconditional denoising estimate from a pretrained unconditional denoiser and a guidance derived from the encoder. For a noisy conditional $p_{e,\gamma}(z \mid x_\gamma; \phi) = \int p_e(z \mid x; \phi)\, p(x \mid x_\gamma)\, dx$ induced by the clean encoder, Tweedie's formula and Bayes' rule give

$$\hat{x}^*(x_\gamma, \gamma, z) = \hat{x}^*(x_\gamma, \gamma) + \gamma\, \nabla_{x_\gamma} \log p_{e,\gamma}(z \mid x_\gamma; \phi), \tag{15}$$

where $\hat{x}^*(x_\gamma, \gamma) = \mathbb{E}[X \mid X_\gamma = x_\gamma]$. This is the posterior-mean form of the score decomposition used in classifier-guided diffusion (Dhariwal & Nichol, 2021). The noisy conditional is generally intractable, so we approximate it with a noise-conditioned encoder $p_e(z \mid x_\gamma, \gamma)$, using the same network for clean and noisy inputs and setting $p_e(z \mid x; \phi) = p_e(z \mid x, 0)$. For each clean-image latent sample $z \sim p_e(z \mid x; \phi)$, we compute the *guidance field*

$$a_\phi(x_\gamma, z, \gamma) = \nabla_{x_\gamma} \log p_e(z \mid x_\gamma, \gamma) \tag{16}$$

by automatic differentiation, holding the sampled z fixed in this input derivative. Together with the pretrained unconditional estimate $\hat{x}_\theta$, this yields

$$\hat{x}_{\theta,\phi}(x_\gamma, \gamma, z) = \hat{x}_\theta(x_\gamma, \gamma) + \gamma\, a_\phi(x_\gamma, z, \gamma). \tag{17}$$

Conditioning therefore requires no additional network beyond the encoder and unconditional denoiser. Thus we can minimize the conditional entropy simply by minimizing the error between the guided denoising estimate $\hat{x}_{\theta,\phi}(x_\gamma, \gamma, z)$ and the clean input x, integrated over all values of γ.

Rate The rate measures the information carried by the clean-image representation, $R(\phi) = I_e(X; Z)$. As mentioned before, ELBO regularization includes an extraneous term pulling the aggregated posterior towards an arbitrary prior (Eq. 3). Here we constrain *only* the mutual information, again using the I-MMSE.

Let us first define the instantaneous rate R_γ as the γ-derivative of the rate. Under the endpoint conditions $I_e(X_\gamma; Z) \to I_e(X; Z)$ as $\gamma \downarrow 0$ and $I_e(X_\gamma; Z) \to 0$ as $\gamma \to \infty$, that it is equivalent to the marginal mutual information between the noisy image and the latent:

$$R(\phi) = I_e(X; Z) = \int_0^\infty R_\gamma(\phi)\, d\gamma, \qquad R_\gamma(\phi) \equiv -\frac{d}{d\gamma} I_e(X_\gamma; Z) \tag{18}$$

Intuitively, R_γ measures how quickly information shared by the image and latent is lost as noise increases. We connect this term to the guidance field by using de Bruijn's identity, which provides an exact link between mutual information and Fisher information (Cover & Thomas, 2006):

$$-\frac{d}{d\gamma} I_e(X_\gamma; Z) = \frac{1}{2} J(X_\gamma|Z) = \frac{1}{2}\,\mathbb{E}\Big[\big\|\nabla_{X_\gamma} \log p_{e,\gamma}(Z \mid X_\gamma; \phi)\big\|^2\Big] \geq 0. \tag{19}$$

The derivation is provided in Appx A.4. In practice, we use the encoder-derived guidance field,

$$R(\phi) = \frac{1}{2}\int_0^\infty \mathbb{E}\Big[\|a_\phi(X_\gamma, Z, \gamma)\|^2\Big]\, d\gamma, \tag{20}$$

which recovers the exact mutual information when the learned guidance matches the score of the induced noisy conditional. Intuitively, the penalty limits how strongly the latent can modify denoising across noise levels. We also show in Appx. A.5 that this rate can also be expressed as a constraint on the guidance *direction* rather than on its *norm*.

Combining the definitions of distortion (Eq. 17) and rate (Eq. 20), the joint objective in Eq. 13 can be expressed in terms of the guidance field as

$$\mathcal{L}(\theta, \phi) = \frac{1}{2}\int_0^\infty \mathbb{E}\left[\frac{1}{\gamma^2}\,\|X - \hat{x}_\theta(X_\gamma, \gamma) - \gamma a_\phi(X_\gamma, Z, \gamma)\|^2 + \beta\,\|a_\phi(X_\gamma, Z, \gamma)\|^2\right] d\gamma. \tag{21}$$

In practice, we approximate this integral by sampling γ using discretized intervals and finite endpoints, as in (Kong et al., 2023). The integral needs only two networks: the (pre-trained) unconditional denoiser $\hat{x}_\theta(x_\gamma, \gamma)$ and the encoder $p_e(z|x_\gamma, \gamma)$ that forms the guidance field $a_\phi(x, z, \gamma)$.

Encoder constraints Mutual information is invariant under invertible transformations, so information preservation alone cannot determine the organization of latent coordinates. We introduce an architectural bias by restricting the encoder to a Gaussian with diagonal covariance:

$$p_e(z \mid x_\gamma, \gamma) = \mathcal{N}\left(z; \mu_\phi(x_\gamma, \gamma), \operatorname{diag}\left[\sigma_\phi^2(x_\gamma, \gamma)\right]\right). \tag{22}$$

This restriction makes latent coordinates conditionally independent given the input, without requiring the aggregate latent distribution to match a prescribed prior. Motivated by analyses of the inductive biases associated with diagonal Gaussian encoders (Rolínek et al., 2019; Reizinger et al., 2022), we use this family to constrain how information is organized across coordinates. In the following section, we examine how this constraint interacts with the different noise sensitivities of generative factors to encourage disentanglement of latent factors across scales.

3.1 Multiscale disentanglement

Natural images contain structure across a broad range of spatial scales, from the arrangement of objects to their boundaries and surface textures. This organization is reflected in approximately power-law spatial spectra and statistical regularities across resolutions (Field, 1987; Ruderman & Bialek, 1994), and has been linked to the distribution of object sizes in natural scenes (Ruderman, 1997). Information at different scales contributes differently to scene recognition (Schyns & Oliva, 1994), while dependencies across scales influence the appearance of natural textures (Portilla & Simoncelli, 2000). Together, these observations motivate asking whether multiscale structure can help organize learned representations.

We formalize this question in an idealized generative model $x = f(\mathbf{y})$, where the factors $\mathbf{y} = (y_1, \ldots, y_K)$ are independent and standardized to unit variance. We assume that their effects on the image are locally orthogonal and have separated energies. For the Gaussian channel $x_\gamma = \sqrt{\gamma}\, x + \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, these energies determine the local sensitivity of the observation to each factor: larger-energy factors can be estimated at lower signal-to-noise ratios. This gives a noise-based characterization of scale, with each factor possessing a characteristic SNR. While the natural-image evidence motivates this setting, we prove the following theorem under the idealized case of factor independence, orthogonality, and energy separation:

Theorem 3.1 (Disentangled hierarchical representation)**.** *Under assumptions of separability, factor independence, scale separation, and regularity, the maximally informative code formed by a factorized Gaussian encoder assigns one latent coordinate to each factor estimate, $z_i = \hat{y}_{k(i)}$. This assignment is preserved across noise levels. Indexing factors by increasing energy, the latents become informative in the order $y_K, \ldots, y_1$ as the SNR γ increases. The assignment holds up to a permutation and per-coordinate reparametrization.*

The complete proof is provided in Appx. A.6. Intuitively, the argument is based on the observation that different factors survive different amounts of noise. This means that a noisy image can be described by a distribution over plausible factor values. At sufficiently low noise, separability allows the posterior to be approximately Gaussian with diagonal covariance in the factor coordinates. Meanwhile, scale separation gives those coordinates distinct posterior variances, so rotations that mix factors introduce correlations that a factorized Gaussian encoder cannot represent well. This mismatch incurs a cost under the InfoMax objective, thereby favoring alignment with the factor axes. Differences in energy determine the order in which factors become informative as noise decreases; once a factor is assigned to an axis, training across noise levels preserves that assignment.

Altogether, this suggests a possible explanation for the disentanglement improvements reported in diffusion-based representation learning (Lyo et al., 2025): training across noise levels may provide structure that complements the encoder's inductive bias, irrespective the additional regularization.

Next, we investigate if this account extends beyond the idealized setting. On dSprites, known generative factors allow us to quantify the alignment between learned coordinates and ground-truth factors. On CelebA, where the underlying generative factors are unknown but naturally multiscale, we examine the semantic changes and spatial scales associated with individual latent coordinates.

4 NUMERICAL EXPERIMENTS

4.1 DSPRITES

To numerically assess IMAE performance, we first trained the model on the dSprites dataset (Matthey et al., 2017). This is a classic synthetic dataset, suitable for testing multiscale disentanglement as images are generated deterministically from six independent factors that exist at different scales; for example, the x, y positions are orientation coarse and fine scale variables, respectively.

Concretely, the model consists of a three layer convolutional encoder and a UNet denoiser, with $\beta = 1 \times 10^{-7}$. We see in Fig. 2b that conditionally generated images are semantically similar to the target, indicating that the model has learned a useful representation of the data. In terms of disentanglement, the model had a FactorVAE score of 0.73 ± 0.006 and a MIG score of 0.33 ± 0.0007 (Kim & Mnih, 2018; Chen et al., 2018). Chance performance corresponds to 0.2 and 0 respectively. These outperform all VAE-based methods and are consistent with diffusion-based representation learning methods. The full comparison is shown in Appx. A.7.

To directly assess the effect of the objective on compression and disentanglement, we trained the same model on either the InfoMax objective (Eq 21) or the ELBO (Eq. 2). We used the same denoiser-based decoder and Gaussian encoder architecture and training parameters for both models (details in Appx. A.8). We measured the rate, distortion, and disentanglement (FactorVAE and MIG scores). For both objectives, we set the β hyperparameter so as to target a region of the rate-distortion curve that both models could reach while keeping the distortion equivalent.

Rate type	β	Distortion	Rate	FactorVAE	MIG
InfoMax	10^{-8}	4.81 ± 0.22	18.71 ± 4.9	0.653 ± 0.007	0.210 ± 0.001
ELBO	10^{-8}	4.812 ± 0.22	42.57 ± 12	0.429 ± 0.009	0.045 ± 0.001

Table 1: Distortion, rate, and disentanglement for the two training objectives.

From Table 1, we see that the ELBO model requires a far larger rate than the InfoMax model to maintain the same level of distortion, indicating that the InfoMax model learns a more compressed representation. Some of the dSprites factors, such as orientation, has S^1 topology that cannot be well captured by a Gaussian prior, and as we saw in Fig. 1, non-diffeomorphisms result in a larger discrepancy in the rate. We also found that the InfoMax objective leads to better performance on both FactorVAE and MIG disentanglement scores (Chen et al., 2018; Kim et al., 2019), suggesting that despite a lower rate, InfoMax is better at ascribing each factor to its own latent axis. Altogether, this direct comparison indicates that removing the prior matching term in Eq. 5, and relying on the multiscale arguments of 3.1, has a large impact on compression and disentanglement.

4.2 NATURAL IMAGES

To demonstrate the generality of our approach, we trained the IMAE on CelebA, a standard natural image dataset of human faces (model details in Appx. A.9). A first test of the model representation is conditional image generation (Fig. 3a). The algorithm is given in Appx. A.10). We see that the conditionally generated images look visually similar to the target image and differ in semantically similar ways. For example, both the target and generated faces in the first column are wearing glasses, even though they are of different styles. This suggests that the representation has learned something semantically meaningful about the data that it was trained on.

Using a latent prior in the ELBO is known to induce regions of low density termed latent *holes* when the aggregate posterior does not match the assumed prior (Falorsi et al., 2018; Xu et al., 2020; Li et al., 2021; Saha et al., 2024). Since IMAE does not enforce a predefined prior, we expect it to be less prone to developing latent holes in this manner. Exact quantification of latent holes is difficult, especially in high dimensional spaces, but one can gauge their presence by linear interpolation in latent space between the latents of pairs of test images and inspecting the corresponding conditionally generated. For IMAE interpolated images correspond to veridical natural images, indicating that the path does not cross regions of low density in the latent space (Fig. 3b).

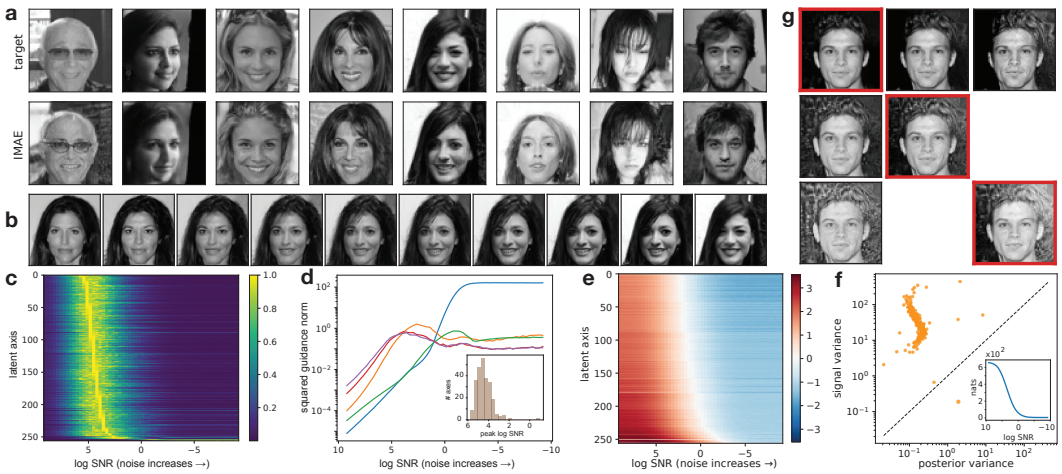

Figure 3: CelebA. **a.** Conditional generation (bottom) based on a target image (top). **b.** Images generated from latents lying along the linear interpolant of two target image latents. **c.** Change in mutual information carried by each latent axis as a function of SNR. We calculate the squared norm of the guidance score (Eq. 20), weighted appropriately, for each latent. Latent axes are sorted by peak sensitivity. **d.** Spectra of guidance norm for the top five most informative latents. Inset: histogram of peak SNR. **e.** An alternative measure of latent informativeness. Each row corresponds to the ratio between the encoder (posterior) variance and the across-dataset variance of the clean image means. Red indicates the usefulness of the latent; blue indicates that the posterior variance has drowned out the signal variance. **f.** Comparison of signal and encoder variances at low noise. Locations above the diagonal signal an informative latent. Inset: the total information of the latents as a function of SNR. **g.** Traversal along two informative axes and their linear combination (red).

Ground-truth generative factors are unavailable for natural images, making disentanglement difficult to assess without inspecting individual latent traversals. Guided by the assumption that semantic factors operate at different scales (Thm. 3.1), we therefore examined whether the learned representation exhibits corresponding structure across noise levels. Our framework allows us to resolve each latent's contribution to the rate across SNRs, revealing that individual latents preferentially encode information within distinct SNR bands (Fig. 3c,d). Encoder uncertainty provides complementary evidence: the ratio of signal variance to encoder variance increases with SNR (Fig. 3e), with progressively more latents becoming informative until nearly all contribute at the lowest noise levels (Fig. 3f). These profiles suggest characteristic SNR *onsets* for individual latents, consistent with the theory (see Appx. A.11 for examples and correspondence to latent traversals). Finally, we found that latents identified as informative by this analysis each control a meaningful semantic feature. Moving within the plane spanned by these latent axes combines their effects in generated images (Fig. 3g), so that the learned representation supports a compositional decomposition of the data.

5 Discussion

InfoMax has long been a guiding principle for unsupervised learning in both neuroscience and machine learning, yet directly optimizing its information-theoretic objective remains challenging. In this work, we address two gaps: the variational gap in estimating mutual information, and the gap between the aggregate posterior and a prescribed latent prior. Using the I-MMSE relation and a denoising decoder, we narrow the first by moving beyond restrictive parametric decoder families. We remove the need to close the second gap altogether by eliminating the prior-matching term in the ELBO. Our theoretical and experimental results show that, under appropriate assumptions on the data's multiscale structure and the encoder, disentanglement can emerge without such a prior constraint. More generally, by separating control over information content from constraints on latent structure, this framework offers a way to investigate the means by whch representations come to reflect the structure of the natural world.

AI USE STATEMENT

In this work, we used generative AI tools for writing python code for generating figures, generating the image of boat parts in Fig. 2, proofreading the initial paper draft, and double checking the derivations and proofs, and assist in editing the proof into a more readable format. We have not used generative AI tools to develop the conceptual idea, formulate the objective, prove the multiscale disentanglement theorem, write the software to train the model, generate synthetic datasets, implement results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to perform literature search. We have reviewed all AI-assisted work. The final proof was checked by external readers and by the first author, and the text was carefully checked by all authors for potential plagiarism and for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our paper raises no questions regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

We have placed clear explanations of our assumptions, derivations, and a complete proof of our main theoretical claim in the appendix. The model training and data processing steps are also provided in the appendix. Our (anonymous) code is available to view and download here: https://anonymous.4open.science/r/infomaxAE/

REFERENCES

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 159–168. PMLR, July 2018.

David Barber and Felix Agakov. The IM Algorithm : A variational approach to Information Maximization. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, November 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.6.1129.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives, April 2014.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, April 2018.

T. M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, N.J, 2nd ed edition, 2006. ISBN 978-0-471-24195-9.

Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Luca Falorsi, Pim de Haan, Tim R. Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S. Cohen. Explorations in Homeomorphic Variational Auto-Encoding, July 2018.

D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and Image Science*, 4(12):2379–2394, December 1987. ISSN 0740-3232. doi: 10.1364/josaa.4.002379.

Dongning Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, April 2005. ISSN 1557-9654. doi: 10.1109/TIT.2005.844072.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, February 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Matthew D Hoffman and Matthew J Johnson. ELBO surgery: Yet another way to carve up the variational evidence lower bound. 2016.

Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. SODA: Bottleneck Diffusion Models for Representation Learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23115–23127, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02181.

Limai Jiang and Yunpeng Cai. Automated Learning of Semantic Embedding Representations for Diffusion Models. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 2025. doi: 10.1137/1.9781611978520.1.

Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2649–2658. PMLR, July 2018.

Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance Factor VAE: Learning and Identifying Disentangled Factors, February 2019.

Yeongmin Kim, Dongjun Kim, HyeonMin Lee, and Il-chul Moon. Unsupervised controllable generation with score-based diffusion models: Disentangled latent code guidance. In *NeurIPS Workshop on Score-Based Methods*, 2022. URL https://openreview.net/pdf/42edd28be60b6cd7a3759bd64227c043ea43959d.pdf.

Yeongmin Kim, Kwanghyeon Lee, Minsang Park, Byeonghu Na, and Il-chul Moon. Diffusion Bridge AutoEncoders for Unsupervised Representation Learning. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, 2013.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-Theoretic Diffusion, February 2023.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable Diffusion via Information Decomposition, May 2024.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004. doi: 10.1103/PhysRevE.69.066138.

Steeve Laquitaine, Simone Azeglio, Carlo Paris, Ulisse Ferrari, and Matthew Chalk. Decomposing stimulus-specific sensory neural information via diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

Ruizhe Li, Xutan Peng, and Chenghua Lin. On the Latent Holes of VAEs for Text Generation, October 2021.

Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. InfoGAN-CR and ModelCentrality: Self-supervised model training and selection for disentangling GANs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6127–6139. PMLR, 2020. URL https://proceedings.mlr.press/v119/lin20e.html.

R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. doi: 10.1109/2.36.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4114–4124. PMLR, May 2019.

Benjamin S. H. Lyo, Eero P. Simoncelli, and Cristina Savin. Disentangled representations via score-based variational autoencoders, December 2025.

Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling Disentanglement in Variational Autoencoders, June 2019.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion Based Representation Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 24963–24982. PMLR, July 2023.

Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, June 2003. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976603321780272.

Vaishnavi Patil, Matthew Evanusa, and Joseph JaJa. DOT-VAE: Disentangling one factor at a time, 2022. URL https://arxiv.org/abs/2210.10920.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On Variational Bounds of Mutual Information. https://arxiv.org/abs/1905.06922v1, May 2019.

Javier Portilla and Eero P. Simoncelli. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, 40(1):49–70, October 2000. ISSN 1573-1405. doi: 10.1023/A:1026553619983. URL https://doi.org/10.1023/A:1026553619983.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Magdalena Proszewska, Nikolay Malkin, and N. Siddharth. On Designing Diffusion Autoencoders for Efficient Generation and Representation Learning, May 2025.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius Von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the Gap: VAEs Perform Independent Mechanism Analysis. In *Advances in Neural Information Processing Systems*, October 2022.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. https://arxiv.org/abs/1401.4082v3, January 2014.

Michal Rolínek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12398–12407, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.01269.

D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, December 1997. ISSN 0042-6989. doi: 10.1016/s0042-6989(97)00008-4.

Daniel L. Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, August 1994. doi: 10.1103/PhysRevLett.73.814.

Surojit Saha, Sarang Joshi, and Ross Whitaker. Matching aggregate posteriors in the variational autoencoder, October 2024.

Philippe G. Schyns and Aude Oliva. From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. *Psychological Science*, 5(4):195–200, July 1994. ISSN 0956-7976. doi: 10.1111/j.1467-9280.1994.tb00500.x.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379–423, 1948. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Noam Slonim, Gurinder Singh Atwal, Gasper Tkacik, and William Bialek. Information-based clustering. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (51):18297–18302, December 2005. ISSN 0027-8424. doi: 10.1073/pnas.0507432102.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models, June 2023.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. On Variational Learning of Controllable Representations for Text without Supervision, August 2020.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. DisDiff: Unsupervised Disentanglement of Diffusion Probabilistic Models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Longxuan Yu, Xing Shi, Xianghao Kong, Tong Jia, and Greg Ver Steeg. MMG: Mutual Information Estimation via the MMSE Gap in Diffusion, November 2025.

Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised Representation Learning from Pre-trained Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 35:22117–22130, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8aff4ffcf2a9d41692a805b3987e29ea-Abstract-Conference.html.

A Appendix

A.1 Collapse of Disentanglement from Prior Matching

Motivation for the ellipse dataset. We use the ellipse dataset to examine whether generative factors acting at distinct spatial scales are recovered as separate latent coordinates, and how this recovery is affected by matching the aggregate posterior to a prescribed prior. The distinction between compression and prior matching follows from the Hoffman decomposition in Eq. 3:

$$\mathbb{E}_{p(x)}\mathrm{KL}\left(p_{e,\phi}(z \mid x) \,\|\, m(z)\right) = I_e(X;Z) + \mathrm{KL}\left(p_{e,\phi(z)} \,\|\, m(z)\right), \tag{23}$$

where $p_{e,\phi}(z) = \int p(x)\, p_{e,\phi}(z \mid x)\, dx$. Thus, the ELBO penalizes both the information retained by the representation and departures of its aggregate distribution from the prior. The latter constraint need not favor the coordinates corresponding to the underlying factors.

Rotational ambiguity. For an isotropic Gaussian prior, $m(z) = \mathcal{N}(0, I)$, prior matching cannot distinguish an aligned representation from rotated mixtures of its coordinates (Mathieu et al., 2019). To make this explicit, write the negative β-ELBO as

$$\mathcal{L}_\beta(p_e, p_d) = -\mathbb{E}_{p(x)\ p_e(z|x)}\left[\log p_e(x \mid z)\right] + \beta\, \mathbb{E}_{p(x)}\mathrm{KL}\left(p_e(z \mid x) \| m(z)\right). \tag{24}$$

For any orthogonal matrix R, define $z' = Rz$ and

$$p_{e,R}(z' \mid x) = p_e(R^\top z' \mid x), \qquad p_{d,R}(x \mid z') = p_d(x \mid R^\top z'). \tag{25}$$

The change of variables preserves the reconstruction term. For a Gaussian encoder $p_e(z|x) = \mathcal{N}(\mu(x), \Sigma(x))$, the KL term is also unchanged:

$$\begin{aligned}
&\mathrm{KL}\left(\mathcal{N}(R\mu, R\Sigma R^\top) \| \mathcal{N}(0, I)\right) \\
&\quad = \frac{1}{2}\left[\mathrm{tr}(R\Sigma R^\top) + \|R\mu\|^2 - d - \log\det(R\Sigma R^\top)\right] \\
&\quad = \frac{1}{2}\left[\mathrm{tr}\,\Sigma + \|\mu\|^2 - d - \log\det\Sigma\right].
\end{aligned}$$

Consequently,

$$\mathcal{L}_\beta(q_{\phi,R}, p_{\theta,R}) = \mathcal{L}_\beta(p_e, p_d). \tag{26}$$

Whenever both parameterizations belong to the admissible model families, the objective therefore assigns equal value to factor-aligned and factor-mixing representations. This establishes non-identifiability of the factor axes (as opposed to it being impossible to recover the factors).

Distinct scales and prior matching. A linear Gaussian analogue illustrates why different factor scales do not resolve this ambiguity through prior matching alone. Let $Y \sim \mathcal{N}(0, \Lambda)$, where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ and $\lambda_1 \neq \lambda_2$. Although the covariance has distinct principal axes, every code

$$Z_R = R\Lambda^{-1/2}Y \tag{27}$$

satisfies

$$\mathrm{Cov}(Z_R) = R\Lambda^{-1/2}\Lambda\Lambda^{-1/2}R^\top = I, \qquad Z_R \sim \mathcal{N}(0, I). \tag{28}$$

Thus both an aligned whitening transformation and rotations that mix the factors will match the prior. Since all remain invertible with $Y = \Lambda^{1/2}R^\top Z_R$, loss of axis alignment does not entail loss of information. The ellipse dataset tests whether the additional prior-matching constraint impairs alignment in practice. Rotational symmetry alone establishes indifference to these equivalent representations.

Implications for InfoMax. The encoder family can break the rotational equivalence: if $\Sigma(x)$ is constrained to be diagonal, then $R\,\Sigma(x)\,R^\top$ is generally inadmissible. Any resulting preference for alignment therefore depends on restrictions on the encoder and the structure of posterior uncertainty rather than on the isotropic prior. Removing the aggregate prior-matching term eliminates a potentially competing constraint, but does not by itself identify the factors, since

$$I(X; T(Z)) = I(X; Z) \tag{29}$$

for any invertible transformation T. In our framework, the proposed source of alignment is instead the combination of a diagonal encoder and the factors' distinct uncertainty profiles across noise levels.

A.2 Barber-Agakov bound on Mutual Information

Here we provide a derivation of the Barber-Agakov bound on mutual information (Barber & Agakov, 2003). Given a fixed data distribution $p(x)$ and an encoder $p(z|x)$, the mutual information is defined as

$$\begin{aligned} I(X;Z) &= h(X) - h(X|Z) \\ &= h(X) + \mathbb{E}_{p(x,z)}[\log p(x|z)]. \end{aligned}$$

Let $q(x|z)$ be a variational approximation to $p(x|z)$. The mutual information can be written in terms of $q(x|z)$:

$$\begin{aligned} I(X;Z) &= h(X) + \mathbb{E}_{p(x,z)}\left[\log p(x|z) + \log \frac{q(x|z)}{q(x|z)}\right] \\ &= h(X) + \mathbb{E}_{p(x,z)}\big[\log q(x|z)\big] + \mathbb{E}_{p(x,z)}\left[\log \frac{p(x|z)}{q(x|z)}\right] \\ &= h(X) + \mathbb{E}_{p(x,z)}\big[\log q(x|z)\big] + \mathrm{KL}\big(p(x|z)\|q(x|z)\big). \end{aligned}$$

Defining the variational approximation to the mutual information as $\tilde{I}(X;Z) := h(X) + \mathbb{E}_{p(x,z)}\big[\log q(x|z)\big]$, we get

$$I(X;Z) = \tilde{I}(X;Z) + \mathrm{KL}\big(p(x|z)\|q(x|z)\big).$$

Since $\mathrm{KL}\big(p(x|z)\|q(x|z)\big) \geq 0$, the variational approximation fo the mutual information is a lower bound on the true mutual information,

$$\tilde{I}(X;Z) \leq I(X;Z), \tag{30}$$

where equality occurs when $q(x|z) = p(x|z)$, such that $\mathrm{KL}\big(p(x|z)\|q(x|z)\big) = 0$.

A.3 Entropy and Mutual Information through I-MMSE

For an absolutely continuous d-dimensional random variable X with finite second moment and finite differential entropy, the I-MMSE relation gives the entropy representation (Kong et al., 2023)

$$h(X) = \frac{d}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left[\frac{d}{1+\gamma} - \mathrm{mmse}(\gamma)\right] d\gamma. \tag{31}$$

The reference term $d/(1+\gamma)$ is the MMSE for a standard Gaussian input to the channel $X_\gamma = \sqrt{\gamma}X + \epsilon$.

To obtain this identity, let $G \sim \mathcal{N}(0, I)$ and apply I-MMSE to the difference between $I(G; \sqrt{\gamma}G + \epsilon) = \frac{d}{2}\log(1+\gamma)$ and $I(X; X_\gamma)$. Both mutual informations vanish at $\gamma = 0$. At high SNR, their difference tends to $h(G) - h(X)$, yielding Eq. 31.

Provided the corresponding conditional entropies and integrals are finite, the conditional identity is

$$h(X|Z) = \frac{d}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left[\frac{d}{1+\gamma} - \mathrm{mmse}(\gamma|Z)\right] d\gamma. \tag{32}$$

Subtracting it from Eq. 31 gives the first equality in Eq. 12. For the second, write

$$X - f^* = (X - f_Z^*) + (f_Z^* - f^*). \tag{33}$$

The cross term has zero expectation because $\mathbb{E}[X - f_Z^*|X_\gamma, Z] = 0$. Consequently,

$$\mathrm{mmse}(\gamma) - \mathrm{mmse}(\gamma|Z) = \mathbb{E}\left[\|f_Z^* - f^*\|^2\right], \tag{34}$$

where the denoisers' arguments have been suppressed for readability.

A.4 Mutual Information as the Squared Norm of the Guidance Score

Relation of marginal Mutual Information to Fisher Information Let X be any random variable with a finite variance with a density $f(x)$. Let ϵ be an independent normally distributed random variable with zero mean and unit variance, and suppose $X_\gamma = X + \sqrt{\gamma}\epsilon$. We want to show that

$$R_\gamma = \partial_\gamma I(Z; X_\gamma) = \frac{1}{2}\,\mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|\nabla_{x_\gamma}\log p(z \mid x_\gamma)\right\|^2\right] \geq 0.$$

de Bruijn's identity connects the entropy to Fisher information:

$$\frac{\partial}{\partial\gamma}h_e(X_\gamma) = \frac{1}{2}J(X_\gamma).$$

where h_e is the differential entropy to base e and J is the Fisher information. The proof can be found in Theorem 17.7.2 of Cover & Thomas (2006).

Meanwhile, given conditioning variable Z, conditional de Bruijn's identity (Lemma A.1) is

$$\frac{\partial}{\partial\gamma}h_e(X_\gamma|Z) = \frac{1}{2}J(X_\gamma|Z).$$

The marginal mutual information is

$$\partial_\gamma I(Z; X_\gamma) = \partial_\gamma h(X_\gamma) - \partial_\gamma h(X_\gamma|Z) = \frac{1}{2}\big(\, J(X_\gamma) - J(X_\gamma|Z)\,\big) = \frac{1}{2}J(Z \mid X_\gamma).$$

Using the definition of the Fisher information,

$$R_\gamma = J(Z \mid X_\gamma) = \mathbb{E}_{p(x,x_\gamma,z)}\Big[\left\|\nabla_{x_\gamma}\log p(z|x_\gamma)\right\|^2\Big].$$

Lemma A.1 (Conditional de Bruijn's identity)**.** *Let X be any random variable with a finite variance with a density $f(x)$. Let ϵ be an independent normally distributed random variable with zero mean and unit variance. Let W be a random variable that we wish to condition on. Suppose $Y_\gamma = X + \sqrt{\gamma}\epsilon$. The conditional de Bruijn's identity is given by*

$$\frac{\partial}{\partial\gamma}h_e(Y_\gamma|W) = \frac{1}{2}J(Y_\gamma|W). \tag{35}$$

Proof. The density of Y_γ is

$$g_\gamma(y) = \int_{-\infty}^{\infty} f(x)\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx.$$

Since ϵ is independent of everything, conditioning on $W = w$ only affects the distribution of X, but not the additive Gaussian noise. Hence the conditional density of Y_γ given W is

$$g_\gamma(y|w) = \int_{-\infty}^{\infty} f(x|w)\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx.$$

Likewise, the joint distribution is defined as

$$\begin{aligned} g_\gamma(y,w) &= f_W(w)g_\gamma(y|w) \\ &= f_W(w)\int_{-\infty}^{\infty} f(x|w)\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx \\ &= \int_{-\infty}^{\infty} f(x,w)\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx. \end{aligned}$$

Since the identities in de Bruijn's identity (derivatives with respect to γ and y) do not affect $f(x)$, we can simply replace $f(x)$ with $f(x|w)$ to get similar identities for $g_\gamma(y|w)$:

$$\frac{\partial}{\partial\gamma}g_\gamma(y|w) = \int_{-\infty}^{\infty} f(x|w)\left[-\frac{1}{2\gamma}\frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right) + \frac{(y-x)^2}{2\gamma^2}\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)\right]dx$$

$$\frac{\partial}{\partial y}g_\gamma(y|w) = \int_{-\infty}^{\infty} f(x|w)\frac{1}{\sqrt{2\pi t}}\left(-\frac{y-x}{\gamma}\right)\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx$$

$$\frac{\partial^2}{\partial y^2}g_\gamma(y|w) = \int_{-\infty}^{\infty} f(x|w)\frac{1}{\sqrt{2\pi\gamma}}\left[-\frac{1}{\gamma}\exp\left(-\frac{(y-x)^2}{2\gamma}\right) + \frac{(y-x)^2}{\gamma^2}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)\right]dx$$

Thus we can see that

$$\frac{\partial}{\partial\gamma}g_\gamma(y|w) = \frac{1}{2}\frac{\partial^2}{\partial y^2}g_\gamma(y|w).$$

Now let's use this to get de Bruijn's identity for $\frac{\partial}{\partial\gamma}h_e(Y_\gamma|W)$. We can define $h_e(Y_\gamma|W)$ as

$$\begin{aligned} h_e(Y_\gamma|W) &= -\int_{-\infty}^{\infty} g_\gamma(y,w)\ln g_\gamma(y|w)dy \\ &= -\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty} g_\gamma(y|w)\ln g_\gamma(y|w)\,dy\,dw \end{aligned}$$

Differentiating, we get

$$\begin{aligned} \frac{\partial}{\partial\gamma}h_e(Y_\gamma|W) &= -\frac{\partial}{\partial\gamma}\int_{-\infty}^{\infty} f(w)h(Y_\gamma|W=w)\,dy\,dw \\ &= -\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial g_\gamma(y|w)}{\partial\gamma}\ln g_\gamma(y|w) + g_\gamma(y|w)\frac{1}{g_\gamma(y|w)}\frac{\partial g_\gamma(y|w)}{\partial y}dy\,dw \\ &= -\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial g_\gamma(y|w)}{\partial t}\ln g_\gamma(y|w) + \frac{\partial g_\gamma(y|w)}{\partial t}dy\,dw \\ &= -\int_{-\infty}^{\infty} f(w)\frac{\partial}{\partial t}\int_{-\infty}^{\infty} g_\gamma(y|w)\,dy\,dw - \int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial g_\gamma(y|w)}{\partial t}\ln g_\gamma(y|w)\,dy\,dw \end{aligned}$$

The first term goes to zero since the inner integrand $\int g_\gamma(y|w)dy = 1$. We can use the identity above to get

$$\frac{\partial}{\partial t}h_e(Y_\gamma|W) = -\frac{1}{2}\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial^2}{\partial y^2}g_\gamma(y|w)\ln g_\gamma(y|w)\,dy\,dw$$

Let's do integration by parts, just like before.

$$\frac{\partial}{\partial t}h_e(Y_\gamma|W) = \int_{-\infty}^{\infty} f(w)\left(\frac{1}{2}\left[\frac{\partial}{\partial y}g_\gamma(y|w)\ln g_\gamma(y|w)\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty}\left[\frac{\partial}{\partial y}g_\gamma(y|w)\right]^2\frac{1}{g_\gamma(y|w)}dy\right)dw.$$

Since the definition of $J(X|W)$ is

$$\begin{aligned} J(X|W) &= \int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty} f(x|w)\left[\frac{\partial}{\partial x}\ln f(x|w)\right]^2 dx\,dw \\ &= \int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty} f(x|w)\left[\frac{1}{f(x|w)}\frac{\partial}{\partial x}f(x|w)\right]^2 dx\,dw, \end{aligned}$$

the second term is the conditional Fisher information $\frac{1}{2}J(Y_\gamma|W)$. All that remains is to show that the first term is zero. We can rewrite the first term as

$$\frac{\partial}{\partial y}g_\gamma(y|w)\ln g_\gamma(y|w) = \left[\frac{1}{\sqrt{g_\gamma(y|w)}}\frac{\partial g_\gamma(y|w)}{\partial y}\right]\left[2\sqrt{g_\gamma(y|w)}\ln\sqrt{g_\gamma(y|w)}\right]$$

This goes to zero, just like for $\frac{\partial}{\partial y}g_\gamma(y)\ln g_\gamma(y)$. Thus

$$\frac{\partial}{\partial\gamma}h_e(Y_\gamma|W) = \frac{1}{2}J(Y_\gamma|W).$$

□

A.5 Rate as a Dot Product

We want to show that the norm-based form of the marginal rate (Eq. 20) is equivalent to the dot-product form of the marginal rate:

$$R_\gamma \equiv \partial_\gamma I(Z;X_\gamma) = \frac{1}{2}\,\mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|\nabla_{x_\gamma}\log q(z\mid x_\gamma)\right\|^2\right] = \mathbb{E}_{p(x,x_\gamma,z)}\left[\nabla_{x_\gamma}\log q(z\mid x_\gamma)\cdot\partial_\gamma x_\gamma\right].$$

Since $I(Z; X_\gamma) = H(Z) - H(Z|X_\gamma)$ and $H(Z)$ is independent of γ, we have

$$\begin{aligned}\partial_\gamma I(Z; X_\gamma) &= \partial_\gamma \mathbb{E}\left[\log p(z|x_\gamma)\right] \\ &= \mathbb{E}\left[\frac{d}{d\gamma} \log p(z \mid x_\gamma)\right] \\ &= \mathbb{E}\left[\partial_\gamma \log p(z|x_\gamma) + \nabla_{x_\gamma} \log p(z|x_\gamma) \cdot \partial_\gamma x_\gamma\right]\end{aligned}$$

The first term can be written as

$$\begin{aligned}\mathbb{E}_{p(z,x_\gamma)}\left[\partial_\gamma \log p(z|x_\gamma)\right] &= \mathbb{E}_{p(x_\gamma)}\left[\int p(z|x_\gamma) \partial_\gamma \log p(z|x_\gamma)\, dz\right] \\ &= \mathbb{E}_{p(x_\gamma)}\left[\int \partial_\gamma p(z|x_\gamma)\, dz\right] \\ &= \mathbb{E}_{p(x_\gamma)}\left[\partial_\gamma \int p(z|x_\gamma)\, dz\right].\end{aligned}$$

where in the second line we have used the identity $p_\gamma \partial_\gamma \log p_\gamma = \partial_\gamma p_\gamma$. Since $\int p(z|x_\gamma)\, dz = 1$, the entire expectation goes to 0. Thus we are left with

$$\partial_\gamma I(Z; X_\gamma) = \mathbb{E}_{p(x,z,x_\gamma)}\left[\nabla_{x_\gamma} \log p(z|x_\gamma) \cdot \partial_\gamma x_\gamma\right].$$

A.6 Proof of Multiscale Disentanglement from Noise-level Structure

Setup. Suppose an image $x \in \mathbb{R}^D$ is generated from K independent latent *features* $\mathbf{y} = (y_1, \ldots, y_K)$ through a smooth injective map $x = f(\mathbf{y})$ that may otherwise be arbitrarily nonlinear. Each feature is standardized to unit variance. We never observe $\mathbf{y}$ directly, but only through the noised observation

$$x_\gamma \;=\; \sqrt{\gamma}\, f(\mathbf{y}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_D),$$

where γ is the signal-to-noise ratio (SNR): $\gamma \to \infty$ is a clean image and $\gamma \to 0$ is one buried in noise.

We build the theorem from the following two quantities. Denoting $J(\mathbf{y}) = \partial f / \partial \mathbf{y} \in \mathbb{R}^{D \times K}$ for the Jacobian and $J_k(\mathbf{y})$ for its k-th column, define the *energy* of feature y_k and its *SNR threshold* as

$$a_k(\mathbf{y}) \;:=\; \|J_k(\mathbf{y})\|^2, \qquad \gamma_k \;\asymp\; 1/a_k.$$

The energy a_k measures how much the image moves when y_k is perturbed, and γ_k is the noise level at which that motion first rises above the unit noise floor. Below γ_k the observation is *essentially uninformative* about y_k, carrying only $o(\gamma\, a_k)$ nats (Lemma A.4); above it, information about y_k accumulates. We assume the SNR thresholds are *distinct and ordered*, $\gamma_1 > \cdots > \gamma_K$. Equivalently, since $\gamma_k \asymp 1/a_k$, the energies run the *opposite* way, $a_1 < \cdots < a_K$: a feature with a large threshold has small energy. As γ increases from 0 we cross γ_K first, so features become detectable one at a time in the order $y_K, \ldots, y_1$.

Our goal is to show that a conditional Gaussian encoder with diagonal covariance,

$$p_\phi(z \mid x_\gamma) = \mathcal{N}\big(\mu_\phi(x_\gamma),\, \mathrm{diag}(\sigma_\phi(x_\gamma))\big),$$

trained on the InfoMax objective, learns latents $\mathbf{z} = (z_1, \ldots, z_M)$ that are disentangled according to the SNR threshold: the encoder assigns each feature to a unique latent and maintains that bijective map across all noise levels.

Assumption 1.

- *(A1)* *There exist features* $\mathbf{y} = (y_1, \ldots, y_K)$ *that are mutually independent, with smooth density, standardized so that* $\mathbb{E}[y_k] = 0$, $\mathrm{Var}(y_k) = 1$.
- *(A2)* $f : \mathbb{R}^K \to \mathbb{R}^D$ *is smooth and injective, and its Jacobian* $J(\mathbf{y}) = \partial f / \partial \mathbf{y} \in \mathbb{R}^{D \times K}$ *has full column rank.*
- *(A3)* (Separability) $J(\mathbf{y})^\top J(\mathbf{y})$ *is diagonal for every* $\mathbf{y}$*; that is,* $J_k(\mathbf{y})^\top J_j(\mathbf{y}) = 0$ *for* $k \neq j$*, with diagonal entries* $a_k(\mathbf{y}) = \|J_k(\mathbf{y})\|^2$.

(A4) (Distinct ordered energies) *The energies occupy non-overlapping bands,* $a_k(\mathbf{y}) < a_{k+1}(\mathbf{y})$*, so the SNR thresholds* $\gamma_k \asymp 1/a_k$ *are distinct and* oppositely *ordered,* $\gamma_1 > \gamma_2 > \cdots > \gamma_K$.

(A5) (Bounded, weakly nonlinear energies) *Over the support of the prior* $p(\mathbf{y})$*, the energies* $a_k(\mathbf{y})$ *are (i)* uniformly bounded*, with fixed constants* $0 < a_{\min} \le a_k(\mathbf{y}) \le a_{\max} < \infty$ *for all* $\mathbf{y}$*; and (ii)* slowly varying*, the curvatures* $\partial^2 f/\partial y_k^2$ *being small relative to the unit spread of* y_k*, so each column* $J_k(\mathbf{y})$ *changes little as* $\mathbf{y}$ *ranges over the prior.*

(A6) (Regularity of generative posterior) *The Bayesian posterior* $p(\mathbf{y} \mid x_\gamma)$ *induced by the generative model concentrates at high SNR in a local Gaussian about the maximum-likelihood estimate. We assume* f *is proper.*

Theorem A.2 (Disentangled hierarchical representation)**.** *Under Assumption 1, the maximally informative code for* x_γ *is the assignment* $z_i = \hat{y}_{k(i)}$ *of one latent per feature, and this assignment is preserved across noise levels, the latents activating in the order* $y_K, \ldots, y_1$ *as* γ *increases. The assignment holds up to a permutation and a per-coordinate reparametrization.*

Proof sketch. The claim rests on two facts: because each feature has a distinct noise threshold, the optimal code assigns each feature to a single latent; and because information only grows with γ, that assignment is stable as the image gets cleaner. The five steps below outline this argument and point to the relevant lemma.

Step 1 (visibility is energy). To detect y_k we must notice its effect on the image. Perturbing y_k by a small δ shifts the noised mean by $\sqrt{\gamma}\,\delta\,J_k(\mathbf{y})$; against unit-variance noise this is detectable only when the *effective per-feature SNR*

$$s_k(\gamma) \;=\; \gamma\, a_k(\mathbf{y}) \tag{36}$$

exceeds one. This is made precise by Lemma A.3.

Step 2 (features appear one at a time). Feature k is invisible while $s_k \ll 1$ and visible once $s_k \gtrsim 1$, i.e. once $\gamma \gtrsim \gamma_k = 1/a_k$. "Loud" features (large a_k, small γ_k) survive heavy noise, while "faint" features (small a_k, large γ_k) are only visible in a clean image. Distinct energies correspond to distinct thresholds, so as we lower the noise the features appear in a fixed order. This monotonicity is given by Lemma A.4(ii)-(iii).

Step 3 (the order labels the features). Because the features switch on at *distinct* noise levels, the crossing level itself identifies which feature has appeared: γ acts as a label sorting the features. This step is laid out in Lemma A.5.

Step 4 (the best code assigns latents to features). Observe a single latent z_i as we sweep γ. If z_i carries one feature it has a *single onset*, defined as a latent that is silent below γ_k and informative above. If the latent mixes two features with different thresholds it has *two* onsets. Among informative codes we show the noise-level structure forces a bijection $z_i \leftrightarrow y_k$ up to relabeling. This step is given by Lemma A.5.

Step 5 (the arrangement is stable across noise levels). This assignment persist across noise levels for two reasons: features appear in a fixed order set by their energy, and since information is monotone in γ, once a feature is visible it stays visible. The *active set* of features therefore grows in a *nested* fashion, with a loud feature appearing low in the hierarchy and a faint feature appearing high. This result comes from combining Lemma A.4(i) with (A4).

The three lemmas below examine a single object, the per-feature threshold γ_k, from three angles. Lemma A.3 has to do with *precision*: we use the Cramér-Rao bound to determine γ_k. Lemma A.4 approaches this from an *information* perspective: the I-MMSE relation recovers the same γ_k and shows that information is monotone. Lemma A.5 is the *labeling* step: distinct thresholds turn onsets into a coordinate frame.

Lemma A.3 (Fisher information and the SNR threshold)**.** *The Fisher information that* x_γ *carries about* $\mathbf{y}$ *is*

$$\mathcal{I}_\gamma(\mathbf{y}) = \gamma\, J(\mathbf{y})^\top J(\mathbf{y}), \tag{37}$$

which under (A3) *is diagonal with entries* $\gamma\, a_k(\mathbf{y})$*. Any unbiased estimator of* y_k *from* x_γ *therefore has variance at least* $(\gamma\, a_k)^{-1}$*. Since* y_k *carries unit prior variance,* x_γ *is essentially uninformative about* y_k *while* $\gamma\, a_k \lesssim 1$ *and becomes informative once* $\gamma\, a_k \gtrsim 1$*, fixing the SNR threshold* $\gamma_k \asymp 1/a_k$*.*

Proof. Since $x_\gamma \mid \mathbf{y} \sim \mathcal{N}(\sqrt{\gamma}\, f(\mathbf{y}), I_D)$, the score is $\nabla_{\mathbf{y}} \log p(x_\gamma \mid \mathbf{y}) = \sqrt{\gamma}\, J(\mathbf{y})^\top \big(x_\gamma - \sqrt{\gamma}\, f(\mathbf{y})\big) = \sqrt{\gamma}\, J(\mathbf{y})^\top \epsilon$. The covariance of the score is the Fisher information, $\gamma\, J^\top J$, giving us (37). (A3) makes it diagonal.

By Cramér-Rao, any unbiased estimate of y_k from x_γ has standard deviation at least $(\gamma a_k)^{-1/2}$, the inverse square root of its Fisher information. The observation becomes informative about y_k only once this variance falls below the unit feature variance (Recall (A1)), which occurs once $\gamma a_k \gtrsim 1$. The crossover at $\gamma a_k \asymp 1$ gives us the SNR threshold $\gamma_k \asymp 1/a_k$. □

Lemma A.4 (Monotone per-feature information)**.** *Denote* $\mathbf{y}_{-k} = (y_j)_{j \neq k}$ *for the features in* $\mathbf{y}$ *other than* y_k*.*

(i) For $\gamma < \gamma'$ *the chain* $\mathbf{y} \to x_{\gamma'} \to x_\gamma$ *is Markov; hence* $\gamma \mapsto I(y_k; x_\gamma)$ *is non-decreasing.*

(ii) Let $\mathit{mmse}(\gamma) = \mathbb{E}\|f(\mathbf{y}) - \mathbb{E}[f(\mathbf{y}) \mid x_\gamma]\|^2$*, and let* $\mathit{mmse}(\gamma \mid y_k) = \mathbb{E}\|f(\mathbf{y}) - \mathbb{E}[f(\mathbf{y}) \mid x_\gamma, y_k]\|^2$ *be the residual error when* y_k *is known. Then, with no assumption on* f *beyond injectivity,*

$$\frac{d}{d\gamma} I(y_k; x_\gamma) \;=\; \tfrac{1}{2}\, \mathit{mmse}_k(\gamma), \qquad \mathit{mmse}_k(\gamma) := \mathit{mmse}(\gamma) - \mathit{mmse}(\gamma \mid y_k) \;\geq\; 0.$$

The per-feature rate of change in information is thus exactly half the error that knowing y_k *removes, and its non-negativity recovers the monotonicity of (i).*

(iii) The onset of y_k *is at* γ_k*. As* $\gamma \to 0$*,*

$$I(y_k; x_\gamma) = \tfrac{1}{2}\, c_k\, \gamma + o(\gamma), \qquad c_k := \mathit{mmse}_k(0) = \operatorname{tr} \operatorname{Cov}\big(\mathbb{E}[f(\mathbf{y}) \mid y_k]\big),$$

with $c_k \asymp a_k$ *under* (A5)*. Thus* $I(y_k; x_\gamma) = O(\gamma\, a_k)$ *until* $\gamma \gtrsim \gamma_k \asymp 1/a_k$*, recovering the threshold of Lemma A.3.*

Proof. (i) Degrading a cleaner observation yields a noisier one: $x_\gamma \overset{d}{=} \sqrt{\gamma/\gamma'}\, x_{\gamma'} + \tilde{\epsilon}$ with $\tilde{\epsilon} \sim \mathcal{N}\big(0, (1 - \gamma/\gamma')I_D\big)$ independent of $x_{\gamma'}$, so $\mathbf{y} \to x_{\gamma'} \to x_\gamma$ is Markov and the data-processing inequality (Cover & Thomas, 2006) gives $I(y_k; x_\gamma) \leq I(y_k; x_{\gamma'})$.

(ii) Since f is injective, $\mathbf{y} \leftrightarrow f(\mathbf{y})$ is a bijection, and given y_k so is $\mathbf{y}_{-k} \leftrightarrow f(\mathbf{y})$:

$$I(\mathbf{y}; x_\gamma) = I(f(\mathbf{y}); x_\gamma), \quad I(\mathbf{y}_{-k}; x_\gamma \mid y_k) = I(f(\mathbf{y}); x_\gamma \mid y_k).$$

Both are governed by the (conditional) I-MMSE relation (Guo et al., 2005),

$$\frac{d}{d\gamma} I(\mathbf{y}; x_\gamma) = \tfrac{1}{2}\, \mathrm{mmse}(\gamma), \qquad \frac{d}{d\gamma} I(\mathbf{y}_{-k}; x_\gamma \mid y_k) = \tfrac{1}{2}\, \mathrm{mmse}(\gamma \mid y_k).$$

They are related by the chain rule for mutual information,

$$I(\mathbf{y}; x_\gamma) = I(y_k; x_\gamma) + I(\mathbf{y}_{-k}; x_\gamma \mid y_k),$$

so differentiating and subtracting gives

$$\frac{d}{d\gamma} I(y_k; x_\gamma) \;=\; \frac{1}{2}\big(\mathrm{mmse}(\gamma) - \mathrm{mmse}(\gamma \mid y_k)\big) \;:=\; \frac{1}{2}\mathrm{mmse}_k(\gamma).$$

where we have defined $\mathrm{mmse}_k(\gamma)$ as the mmse error removed by knowing y_k. This can essentially be interpreted as the amount of residual uncertainty in f that y_k explains at SNR γ. We know that $\mathrm{mmse}(\gamma|y_k) \leq \mathrm{mmse}(\gamma)$ because $\mathbb{E}[f|x_\gamma]$ is a suboptimal estimator when we also know y_k. Hence $\mathrm{mmse}_k(\gamma) \geq 0$. This independently reproves (i) that $I(y_k; x_\gamma)$ is non-decreasing.

(iii) At $\gamma = 0$ (high noise) the observation $x_0 = \epsilon$ is independent of $\mathbf{y}$, so the estimators reduce to prior means and the errors to prior variances:

$$\mathrm{mmse}(0) = \operatorname{tr} \operatorname{Cov}\big(f(\mathbf{y})\big), \qquad \mathrm{mmse}(0 \mid y_k) = \mathbb{E}\big[\operatorname{tr} \operatorname{Cov}\big(f(\mathbf{y}) \mid y_k\big)\big].$$

The law of total variance, $\mathrm{Cov}(f) = \mathbb{E}[\mathrm{Cov}(f \mid y_k)] + \mathrm{Cov}(\mathbb{E}[f \mid y_k])$, gives

$$c_k := \mathrm{mmse}_k(0) = \mathrm{mmse}(0) - \mathrm{mmse}(0 \mid y_k) = \mathrm{tr}\,\mathrm{Cov}\big(\mathbb{E}[f(\mathbf{y}) \mid y_k]\big),$$

the variance of the k-th component of f explained by y_k. To first order, a unit change in y_k moves the image along the Jacobian J_k, so $\mathbb{E}[f \mid y_k] - \mathbb{E}[f] \approx J_k\, y_k$. Thus

$$\mathrm{Cov}\left(\mathbb{E}[f \mid y_k]\right) \approx J_k J_k^\top \mathrm{Var}(y_k) = J_k J_k^\top$$

where in the equality we have used the unit variance assumption (A1). This means that

$$c_k \approx \mathrm{tr}\left(J_k J_k^\top\right) = \|J_k\|^2 = a_k.$$

where we have used assumption (A5)(ii) that J_k varies slowly. Thus $c_k \asymp a_k$.

Integrating the relation from (ii) from 0, and using $I(y_k; x_0) = 0$, we get

$$I(y_k; x_\gamma) = \tfrac{1}{2}\int_0^\gamma \mathrm{mmse}_k(s)\, ds = \tfrac{1}{2}\, c_k\, \gamma + \tfrac{1}{2}\int_0^\gamma \big(\mathrm{mmse}_k(s) - c_k\big)\, ds = \tfrac{1}{2}\, c_k\, \gamma + o(\gamma).$$

The first term is linear and the remainder is $o(\gamma)$ since $\mathrm{mmse}_k(s) \to c_k$ as $s \to 0$ (A5).

The information grows linearly at rate $\frac{1}{2}c_k \asymp \frac{1}{2}a_k$. This means it is negligible while $\gamma a_k \ll 1$ and order one once $\gamma a_k \asymp 1$, i.e. once $\gamma \asymp 1/a_k = \gamma_k$. (At $\gamma = \gamma_k$ the leading term is $\frac{1}{2}a_k \cdot (1/a_k) = \frac{1}{2}$ nat.) This low-SNR slope, extrapolated to where I reaches $O(1)$, places the onset at the same $\gamma_k \asymp 1/a_k$ as the high-SNR Cramér-Rao bound of Lemma A.3; the two regimes meet at the SNR threshold. □

Proof. (i) We can show the Markov relation by writing x_γ in terms of $x_{\gamma'}$: $x_\gamma \overset{d}{=} \sqrt{\gamma/\gamma'}\, x_{\gamma'} + \tilde{\epsilon}$ with $\tilde{\epsilon} \sim \mathcal{N}\big(0, (1 - \gamma/\gamma')I_D\big)$ independent of $x_{\gamma'}$. By the data-processing inequality (Cover & Thomas, 2006) we get $I(y_k; x_\gamma) \le I(y_k; x_{\gamma'})$.

(ii) The scalar identity is given by Guo et al. (2005); under (A3) the orthogonal columns decouple the per-feature estimation errors to leading order, so $\mathrm{mmse}(\gamma) = \sum_k \mathrm{mmse}_k(\gamma)$ and the derivative can be split into per-feature terms.

(iii) By part (ii), the I-MMSE relation gives $\frac{d}{d\gamma} I(y_k; x_\gamma) = \frac{1}{2}\,\mathrm{mmse}_k(\gamma)$, and we will use this relation to analyze the behavior of the information near $\gamma = 0$. At $\gamma = 0$ the observation $x_0 = \epsilon$ is pure noise, independent of $\mathbf{y}$, so the MMSE estimator of feature k's contribution to the image is its prior mean, and the residual error is the full prior variance,

$$c_k \;:=\; \mathrm{mmse}_k(0) \;=\; \mathrm{Var}\big(\text{contribution of } y_k \text{ to } f(\mathbf{y})\big).$$

To first order, a unit perturbation of y_k moves the image along $J_k(\mathbf{y})$, so this variance scales as $\|J_k\|^2 = a_k$, giving $c_k \asymp a_k$ under (A5). Integrating the I-MMSE relation from 0 to γ, and using $I(y_k; x_0) = 0$,

$$I(y_k; x_\gamma) \;=\; \tfrac{1}{2}\int_0^\gamma \mathrm{mmse}_k(s)\, ds \;=\; \tfrac{1}{2}\, c_k\, \gamma + \tfrac{1}{2}\int_0^\gamma \big(\mathrm{mmse}_k(s) - c_k\big)\, ds \;=\; \tfrac{1}{2}\, c_k\, \gamma + o(\gamma),$$

where the remainder is $o(\gamma)$ because the integrand vanishes as $s \to 0$, by the regularity of mmse_k near the origin (A5).

This indicates that as we move away from $\gamma = 0$, the growth in information is linear of order $O(\gamma\, a_k)$. This is negligible while $\gamma\, a_k \ll 1$ and of order one once $\gamma\, a_k \asymp 1$, i.e. once $\gamma \asymp 1/a_k = \gamma_k$. (At $\gamma = \gamma_k$ the leading term is $\frac{1}{2}a_k \cdot (1/a_k) = \frac{1}{2}$ nat, so roughly half a nat has accumulated by the SNR threshold.) Strictly, this is an expansion around $\gamma = 0$, so our statement above extrapolates the low-SNR slope out to where the information reaches $O(1)$. This extrapolation agrees with the high-information estimate of Lemma A.3, which places the onset at the same SNR $\gamma_k \asymp 1/a_k$; both the Cramér-Rao bound and the I-MMSE slope perspective approach the threshold from opposite regimes. □

Definition A.1 (Onset). *Let $z = \psi(x_\gamma)$ be any scalar statistic of the observation. Its information $\gamma \mapsto I(z; x_\gamma)$ is non-decreasing (data processing, Lemma A.4(i)). Say z has an* onset *at γ_k if $I(z; x_\gamma)$ crosses from $o(1)$ to $\Omega(1)$ as γ passes γ_k, where $o(1)$ and $\Omega(1)$ are taken in the effective-SNR sense ($s_k = \gamma a_k \to 0$ and $\to \infty$ respectively). Onsets are counted only at the threshold set $\{\gamma_1, \ldots, \gamma_K\}$.*

Lemma A.5 (Distinct thresholds determine the basis)**.** *Suppose a code is aligned to the feature basis: each coordinate z_i is an invertible reparametrization of a single feature $y_{k(i)}$, for some assignment $k(\cdot)$. Then*

(i) each coordinate becomes informative at the threshold of the feature it carries, and

(ii) the distinctness of these thresholds (A4) *recovers the assignment $k(\cdot)$ uniquely.*

Proof. (i) *Each feature has a single onset at its own SNR threshold.* By separability (A3) and Lemma A.4(ii), $I(y_k; x_\gamma)$ depends on the noise level only through the feature's effective SNR $s_k = \gamma a_k$. It is a fixed increasing function of s_k that vanishes at $s_k = 0$, is $o(1)$ for $s_k \ll 1$ and is $\Omega(1)$ for $s_k \gtrsim 1$. Such a curve crosses from $o(1)$ to $\Omega(1)$ exactly once, at $s_k \asymp 1$, i.e. $\gamma \asymp \gamma_k$. Because the other thresholds γ_j ($j \neq k$) never enter this curve, it has no second crossing. Hence y_k has a single onset at γ_k.

Each aligned coordinate inherits its feature's onset. By hypothesis, latent z_i is an invertible reparametrization of the single feature $y_{k(i)}$ and mutual information is invariant under such reparamaterization, so

$$I(z_i; x_\gamma) = I(y_{k(i)}; x_\gamma).$$

Latent z_i thus has the same information rate as $y_{k(i)}$ and, by Step 1, a single onset at $\gamma_{k(i)}$.

(ii) The thresholds $\gamma_1 > \cdots > \gamma_K$ are distinct (A4), so the map $i \mapsto \gamma_{k(i)}$ is injective. Reading each coordinate's onset level off against $\{\gamma_k\}$ therefore uniquely recovers the assignment $k(\cdot)$. □

Assembling the theorem. The proof has three steps. *Concentration*: at high information the posterior over the active features is Gaussian and, under separability, diagonal in the feature frame — but only up to a rotation within the active block, which sufficiency alone cannot remove. *De-rotation*: the diagonal encoder, under InfoMax, is exact only in the frame that diagonalizes the posterior, and distinct energies make that frame unique — so it lands on the feature frame, and Lemma A.5 labels the axes. *Nesting*: monotone information grows the active set in threshold order.

Proof. Concentration. By Lemma A.3 the posterior precision is $\mathcal{I}_\gamma(\mathbf{y}) = \gamma\ \mathrm{diag}(a_k)$. Restrict to the active features $A(\gamma) = \{k : \gamma > \gamma_k\}$; inactive features carry $o(1)$ information (Lemma A.4(iii)) and do not concentrate. Let $\hat{\mathbf{y}}$ denote the maximum-likelihood estimate of $\mathbf{y}$ from x_γ, i.e. the minimizer of $\|x_\gamma - \sqrt{\gamma} f(\mathbf{y})\|^2$, and $\hat{y}_k$ its k-th coordinate.

As γ increases the posterior sharpens, and by Bernstein-von Mises theorem (Vaart, 1998) (otherwise known as *local asymptotic normality* in the frequentist literature) the posterior over the active features $\mathbf{y}_{A(\gamma)}$ converges, to leading order $1/\gamma$, to a normal distribution about $\hat{\mathbf{y}}$ with covariance $\mathcal{I}_\gamma^{-1}$. Diagonality of $\mathcal{I}_\gamma$ (A3) makes this covariance diagonal, so the posterior factorizes over features,

$$p\big(\mathbf{y}_{A(\gamma)} \mid x_\gamma\big) \;\approx\; \prod_{k \in A(\gamma)} \mathcal{N}\big(\hat{y}_k,\, (\gamma a_k)^{-1}\big), \qquad A(\gamma) = \{k : \gamma > \gamma_k\}.$$

Because this depends on x_γ only through the centers $\hat{y}_k$ (the widths being fixed), the vector $(\hat{y}_k)_{k \in A(\gamma)}$ is a minimal sufficient statistic for $\mathbf{y}_{A(\gamma)}$. Moreover, since the posterior factorizes, this statistic is coordinate-wise: each $\hat{y}_k$ is sufficient for its own feature y_k alone. Hence the maximally informative code's coordinates are re-parametrizations of the active features $\{y_k\}_{k \in A(\gamma)}$. This is determined, however, only up to a rotation *within* the active block, since $I(z; x_\gamma)$ is invariant under any invertible reparametrization of z. This residual rotation is removed in the de-rotation step.

De-rotation. Concentration determined the code only up to a rotation within the active block; we show the diagonal encoder, under InfoMax, selects the feature basis out of this rotation family. By the *feature basis* we mean the coordinate axes aligned with $y_1, \ldots, y_K$; under (A3) this is exactly the eigenbasis of the posterior covariance $\mathcal{I}_\gamma^{-1} = \Lambda := \mathrm{diag}\left((\gamma a_k)^{-1}\right)$.

Information alone cannot make this selection. For any rotation R of the active block, $R\hat{\mathbf{y}}$ is an equally valid sufficient statistic and $I(Rz; x_\gamma) = I(z; x_\gamma)$, so InfoMax evaluated on a fixed code is blind to R. The tie is broken instead by the factorial *constraint* on the encoder.

The encoder is restricted to diagonal covariance, $p_\phi(z \mid x_\gamma) = \mathcal{N}\big(\mu_\phi, \mathrm{diag}(\sigma_\phi)\big)$, so its coordinates are conditionally independent given x_γ. A mean-field code matches the posterior exactly only when

the posterior is itself diagonal in the code coordinates; otherwise it approximates a correlated Gaussian by a factorized one and pays a gap equal to the posterior's total correlation. In a basis where the active block has covariance Σ, the information among the coordinates that a factorized code must discard is given by

$$\Delta(\Sigma) \;=\; \min_{q \text{ factorized}} \mathrm{KL}\big(p \,\|\, q\big) \;=\; \tfrac{1}{2}\left(\sum_k \log \Sigma_{kk} - \log\det \Sigma\right) \;\geq\; 0.$$

By Hadamard's inequality, $\Delta(\Sigma) = 0$ iff Σ is diagonal.

Applying this to the rotation family R gives covariance $R\Lambda R^\top$ with fixed $\det(R\Lambda R^\top) = \det\Lambda$, so

$$\Delta\big(R\Lambda R^\top\big) \;=\; \tfrac{1}{2}\log \frac{\prod_k \big(R\Lambda R^\top\big)_{kk}}{\det\Lambda}.$$

By contrast, in the feature basis $R = I$, the covariance Λ is already diagonal so $\Delta = 0$, meaning that a diagonal encoder is exact and InfoMax-optimal. For any $R \neq I$, $R\Lambda R^\top$ is non-diagonal because the eigenvalues $(\gamma a_k)^{-1}$ are distinct (A4), so $\Delta(R\Lambda R^\top) > 0$ and the rotated basis is suboptimal. The only rotations preserving diagonality are permutations and sign flips.

The InfoMax-optimal diagonal encoder thus selects the unique basis that diagonalizes the posterior, i.e. the feature basis, up to a permutation and a sign. Lemma A.5 fixes the permutation by matching each coordinate's onset to a threshold γ_k.

Remark: the mechanism uses the finite posterior width $(\gamma a_k)^{-1}$ and the distinctness of those widths (A4): in the deterministic $\sigma \to 0$ limit the gap vanishes and rotations are free again. The alignment is thus *derived* from the diagonal-encoder InfoMax objective, not imposed, and holds to leading order in $1/\gamma$, matching concentration. Without the diagonal constraint the rotation survives.

Nesting: As γ increases, the active set $A(\gamma) = \{k : \gamma > \gamma_k\}$ only gains elements, and by the monotonicity of $I(y_k; x_\gamma)$ (Lemma A.4(i)) a feature stays informative once past its threshold. The active set is therefore nested, $A(\gamma) \subseteq A(\gamma')$ for $\gamma < \gamma'$, and grows one feature at a time in the order set by (A4), $\gamma_1 > \cdots > \gamma_K$: features are represented as $y_K, \ldots, y_1$ as the image possesses less noise. This nested, ordered set of activations is the hierarchy. □

A.7 Comparison of disentanglement scores on dSprites

Table 2 compares IMAE with published disentanglement scores on dSprites. IMAE exceeds the VAE baselines reported by Kim et al. (2022) on both metrics and achieves a MIG numerically close to that of their diffusion-based LCG-DM. Other evaluations report higher scores for FactorVAE, InfoGAN-CR, and DOT-VAE, placing IMAE within the range of existing methods. Since architectures, hyperparameter selection, and evaluation protocols differ across studies, these results provide context rather than a controlled ranking.

A.8 dSprites Architecture and Training

Data and architecture. We used dSprites at 64×64 resolution, with $737{,}280$ single-channel images in the configured dataset. The model expects image intensities in $[-1, 1]$. The ELBO-based and InfoMax variants share the same encoder and diffusion denoiser architecture.

The encoder parameterizes an eight-dimensional diagonal Gaussian,

$$p_\phi(z \mid x, t) = \mathcal{N}(z; \mu_\phi(x,t), \mathrm{diag}[\exp(\ell_\phi(x,t))]), \tag{38}$$

where ℓ_ϕ denotes the predicted log variance. Three 3×3 convolutions with stride two and padding one produce 32, 64, and 128 channels at resolutions 32×32, 16×16, and 8×8. Each convolution uses an additive bias, eight-group normalization with learned affine parameters, and ReLU. Adaptive average pooling to 3×3 produces $1{,}152$ features, passed to separate eight-output linear heads for the mean and log variance. The log-variance head has zero-initialized weights and bias -2, giving initial variance e^{-2}.

Timestep conditioning uses a 64-dimensional embedding, a learned $64 \to 64$ affine map with SiLU activation, and separate FiLM heads for each convolution. After normalization and before ReLU,

Model	Source	FactorVAE ↑	MIG ↑
VAE	K	0.58	0.05
β-VAE	K	0.60	0.23
β-TCVAE	K	0.67	0.23
FactorVAE	K	0.71	0.29
LCG-DM	K	0.84	0.34
FactorVAE ($\gamma = 10$)	L	0.83 ± 0.01	0.40 ± 0.01
FactorVAE ($\gamma = 40$)	L	0.82 ± 0.01	0.43 ± 0.01
InfoGAN-CR	L	0.88 ± 0.01	0.37 ± 0.01
DOT-VAE$_1$	P	0.77 ± 0.12	0.38
DOT-VAE$_2$	P	0.72 ± 0.15	0.34
IMAE (ours)	–	0.73 ± 0.006	0.33 ± 0.0007

Table 2: Disentanglement on dSprites; higher is better for both metrics. Sources denote the papers reporting the numerical results: K: Kim et al. (2022), Table 1; L: Lin et al. (2020), Table 1; P: Patil et al. (2022), Table 1. Literature uncertainties are reproduced where reported and need not share the same statistical interpretation. Duplicate FactorVAE entries retain results from distinct evaluations.

features are modulated as $h \mapsto (1 + s(t)) \odot h + b(t)$. The FiLM heads are initialized to zero, so modulation initially acts as the identity. Clean images use $t = 0$; noisy images use their sampled timesteps.

The denoiser is a U-Net with contracting widths 32, 64, and 128 at resolutions 64×64, 32×32, and 16×16. Each stage has two 3×3 convolutions followed by stride-two 2×2 average pooling. Two bottleneck convolutions produce 256 channels at 8×8. The expanding path uses stride-two 2×2 transposed convolutions with widths 128, 64, and 32, concatenating the corresponding contracting features. These stages contain two, two, and one feature convolutions, respectively, followed by a single-channel linear output convolution. All 3×3 convolutions have unit stride, padding one, and no biases. Feature convolutions use ReLU and, except for the first, 32-group normalization without learned affine parameters. The output convolution has neither normalization nor activation. A learned affine projection maps a 64-dimensional sinusoidal timestep embedding to a 64×64 map, concatenated with the image as a second input channel.

Diffusion and conditioning. Training uses 200 timesteps with a cosine schedule in $\bar{\alpha}_t$:

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{4096}). \tag{39}$$

Timesteps are sampled from a custom timestep distribution that is uniform except the first $1/7$th of the timestep range, which follows a linear schedule. For a reparameterized clean latent $z \sim p_\phi(z \mid x, 0)$, the encoder provides the guidance score and conditional noise prediction

$$\begin{aligned} g_\phi(x_t, t, z) &= \nabla_{x_t} \log p_\phi(z \mid x_t, t), \\ \widehat{\epsilon}_{\theta,\phi}(x_t, t, z) &= \epsilon_\theta(x_t, t) - \sqrt{1 - \bar{\alpha}_t}\, g_\phi(x_t, t, z). \end{aligned} \tag{40}$$

The same conditioning mechanism is used for both objectives.

Training objectives. We trained the model on either InfoMax or ELBO. Both variants minimize the objective

$$\widehat{\mathcal{L}} = \frac{1}{BD} \sum_{i=1}^{B} \left[\|\epsilon_i - \widehat{\epsilon}_{\theta,\phi}(x_{t_i}, t_i, z_i)\|_2^2 + \beta r_i \right], \qquad D = 4096, \tag{41}$$

with the regularizer distinguishing the two variants:

$$r_i = \begin{cases} \dfrac{1}{2} \displaystyle\sum_{j=1}^{8} \left(\exp(\ell_{ij}) + \mu_{ij}^2 - 1 - \ell_{ij} \right), & \text{ELBO-based,} \\ \|g_\phi(x_{t_i}, t_i, z_i)\|_2^2, & \text{InfoMax.} \end{cases} \tag{42}$$

Table 3: Shared dSprites training settings.

Setting	Value
Optimizer / learning rate (both networks)	Adam / 10^{-3}
Training epochs	1,200
Distributed training	DDP, one node, three GPUs
Batch size per GPU / total	500 / 1,500
Rate coefficient / schedule	10^{-8} / constant
Precision / random seed	32-bit / 43
`sigma_minmax`	$(10^{-4}, 0.9999)$

Here μ_{ij} and ℓ_{ij} are clean-image encoder outputs. The first regularizer is equivalent to the KL divergence to $\mathcal{N}(0, I_8)$, while the second is the unscaled squared guidance norm. Neither term receives additional timestep weighting. The implementation sums over image coordinates (or latent coordinates for the KL), averages over examples, and divides both terms by D. Thus, pixel normalization preserves the relative weight $\beta = 10^{-8}$, which remains constant because its initial and final values coincide. Gradients propagate through both the guidance computation and the sampled clean latent.

Initialization and optimization. Only the denoiser is initialized from a pretrained dSprites DDPM. Both networks are then jointly optimized using Adam at a constant learning rate. Training settings shared by the two variants are summarized in Table 3.

A.9 CelebA Architecture and Training

Data. We used 64×64 greyscale CelebA images, normalized to $[-1, 1]$. The first 60,000 images in the stored dataset formed the training set and the remainder the test set, rather than using the official CelebA partitions.

Probabilistic encoder. The encoder parameterizes a 256-dimensional diagonal Gaussian $p_\phi(z \mid x) = \mathcal{N}(z; \mu_\phi(x), \mathrm{diag}[v_\phi(x)])$ (Eq. 22). Three stages with 64, 128, and 256 channels each apply two 3×3 convolutions followed by stride-two 2×2 average pooling, reducing the spatial resolution from 64×64 to 8×8. Two further convolutions map the channel width from 256 to 512 to 64. The flattened features feed separate, bias-free linear heads predicting the mean and an unconstrained scale $a_\phi(x)$, with

$$v_{\phi,i}(x) = \max\left\{\mathrm{softplus}(a_{\phi,i}(x))^2, 10^{-6}\right\}. \tag{43}$$

Latents are sampled as $z = \mu_\phi(x) + \sqrt{v_\phi(x)} \odot \eta$, where $\eta \sim \mathcal{N}(0, I_{256})$. The encoder processes both clean and noisy images without explicit timestep conditioning.

Diffusion denoiser. The noise-predicting U-Net has three contracting stages with 128, 256, and 512 channels, each comprising two 3×3 convolutions and stride-two 2×2 average pooling. Two 1,024-channel bottleneck convolutions operate at 8×8 resolution. The expanding stages use stride-two 2×2 transposed convolutions with 512, 256, and 128 output channels, concatenated with the corresponding residual contracting features. These stages contain 3×3 feature convolutions, followed by a single-channel linear 3×3 output convolution. A learned affine projection maps a 64-dimensional sinusoidal timestep embedding to a 64×64 tensor that is concatenated with the noisy image and enters into the U-Net as a single input.

Activations and normalization. Both networks use unit-stride, zero-padded 3×3 convolutions without additive biases and ReLU activations after all feature convolutions. Except for each network's first convolution, these convolutions are followed by a bias-free batch normalization layer before the activation. During training, this layer divides each channel by its standard deviation across the batch and spatial dimensions and applies a learned multiplicative gain; it does not subtract the mean or learn an additive offset. The variance stabilizer is 10^{-5}, and running standard deviations are updated with momentum 0.1 for evaluation. The encoder's linear heads also omit biases. The denoiser's final convolution has neither normalization nor an activation.

Table 4: CelebA architecture and optimization settings.

Setting	Value
Optimizer	Adam
Learning rate (both networks)	10^{-3}
Epochs	1,000
Batch size per GPU	512
Precision / random seed	32-bit / 43
`sigma_minmax`	$(10^{-4}, 0.9999)$
Denoiser parameterization	Noise prediction
Image shape	$1 \times 64 \times 64$
Training images	60,000
Latent dimension / covariance	256 / diagonal
Encoder stage widths	$64, 128, 256$
Denoiser stage widths	$128, 256, 512$

Noise and conditioning. Training uses a variance-preserving Gaussian corruption process,

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{4096}), \tag{44}$$

with 1,000 noise levels and uniform timestep sampling. The configuration selects a cosine schedule in $\bar{\alpha}_t$; the model adapter sets its offset to 0.01 to match the pretrained model. Conditioning is supplied through the encoder likelihood gradient,

$$g_\phi(x_t, z) = \nabla_{x_t} \log p_\phi(z \mid x_t), \qquad z \sim p_\phi(z \mid x), \tag{45}$$

which gives the conditional noise prediction

$$\widehat{\epsilon}_{\theta,\phi}(x_t, t, z) = \epsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t}\, g_\phi(x_t, z). \tag{46}$$

The latent therefore conditions the denoising prediction through the likelihood gradient, without a separate latent embedding in the U-Net.

Training objective. For a minibatch of B examples, we minimize

$$\widehat{\mathcal{L}}(\theta, \phi) = \frac{1}{B} \sum_{i=1}^{B} \left[\|\epsilon_i - \widehat{\epsilon}_{\theta,\phi}(x_{t_i}, t_i, z_i)\|_2^2 + \beta \, \|g_\phi(x_{t_i}, z_i)\|_2^2 \right], \qquad \beta = 10^{-3}. \tag{47}$$

Each example uses independently sampled corruption noise, timestep, and reparameterized clean latent. Both norms sum over all 4,096 pixels; the loss averages over examples, with constant β and no additional timestep weighting. The regularizer uses the unscaled guidance score. Automatic differentiation retains the score's derivative graph, allowing gradients through both the guidance and the sampled latent to the encoder.

Initialization and optimization. The denoiser is initialized from a pretrained CelebA checkpoint, and both networks are jointly optimized using Eq. 47. Training uses distributed data parallelism on four NVIDIA A100 GPUs; a summary of architecture and optimization settings are listed in Table 4.

A.10 Algorithm for conditional generation

To generate images conditioned on a test image x_{ref}, we draw a latent $z \sim p_\phi(z \mid x_{\text{ref}})$ and hold it fixed throughout reverse diffusion. At each timestep, the encoder evaluates the likelihood of this latent given the current noisy image. Its input gradient provides the guidance field, which modifies the unconditional noise prediction before a DDPM sampling step (Ho et al., 2020; Dhariwal & Nichol, 2021). Algorithm 1 uses the variance-preserving parameterization of the preceding subsection. Writing $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, with $\bar{\alpha}_0 = 1$, the forward marginal and reverse posterior variance are

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\, \beta_t, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{48}$$

Algorithm 1 Conditional generation with the InfoMax Autoencoder

Require: Reference image x_{ref}, encoder p_ϕ, noise predictor ϵ_θ, schedule $\{\beta_t\}_{t=1}^T$

$z \sim p_\phi(z \mid x_{\text{ref}})$ ▷ Encode once

$x_T \sim \mathcal{N}(0, I)$

for $t = T, T-1, \ldots, 1$ **do**

  $g_t \leftarrow \nabla_{x_t} \log p_\phi(z \mid x_t)$ ▷ Differentiate with z fixed

  $\hat{\epsilon}_t \leftarrow \epsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t}\, g_t$

  $\mu_t \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\epsilon}_t \right)$

  if $t > 1$ **then**

    $\xi_t \sim \mathcal{N}(0, I)$

  else

    $\xi_t \leftarrow 0$

  end if

  $x_{t-1} \leftarrow \mu_t + \sqrt{\tilde{\beta}_t}\, \xi_t$

end for

return x_0

Here β_t denotes the diffusion variance, distinct from the rate coefficient β in the training objective.

Both networks remain fixed during sampling; gradients are required only with respect to the noisy image. Gaussian initialization assumes that $\bar{\alpha}_T$ is sufficiently small for the terminal signal to be negligible. The procedure produces approximate samples from the image distribution conditioned on z, with variability arising from the initial noise and stochastic reverse steps.

A.11 Latent axes with large effects encode coarse image structure

Measuring the effect of each axis. To measure how much each latent axis changes generated images, we took the latent of a test image and traversed along a single axis until we reached the 2nd or 98th percentiles of its posterior mean across 2,000 test images. All other axes were held at the original image's posterior mean. We generated both images from these two latent endpoints with the same sampling noise, so that any difference between them can be attributed to that one axis. We define the effect of an axis as the mean squared error between the two decoded images, averaged over four test images.

We found that the largest effects are concentrated in a small number of axes (Fig. 4a). As measured by the MSE, the axes with the largest effect seem to change global properties of the image (Fig. 4b): the brightness of the background and hair (axis 210), left–right illumination (axis 138) and the brightness of the face (axis 187).

To see if the effect size of the axis is related to the scale of what it encodes, we computed its information-weighted mean log SNR which is the noise level around which its information density $1/2\sigma^2\mathbb{E}\left\|a_\phi^{(d)}\right\|^2$ is concentrated. Comparing these two quantities, we find that axes whose information persists to higher noise levels have larger effects on the image (Fig. 4c, Spearman $\rho = -0.94$, $n = 253$). This relationship is not explained by differences in the total information. Even when controlling for the information each axis carries, the partial correlation is still $\rho = -0.90$.

Three axes had guidance fields that peaked at very large amounts of noise, suggesting they respond to the noise rather than to image content and thus change the decoded image very little. As a result, they were excluded from the correlation set. As expected, excluding a broader set of 18 axes with substantial guidance at pure noise gave nearly identical results ($\rho = -0.93$, partial $\rho = -0.90$, $n = 238$).

Guidance fields are independent at fine scales and aligned at coarse ones. The per-axis rates $\left\|a_\phi^{(d)}\right\|^2$ add up to the rate for the full latent $\left\|\sum_d a_\phi^{(d)}\right\|^2$ only if the per-axis guidance fields are orthogonal on average. We compared the squared norm of the full guidance field, $\left\|\sum_d a_\phi^{(d)}\right\|^2$,

against the sum of the per-axis squared norms, $\sum_d \left\| a_\phi^{(d)} \right\|^2$ (Fig. 4d), and found that the two coincide at high SNR but the full norm is about twice the sum at low SNR. This indicates that the per-axis guidance fields are approximately orthogonal at high SNR, such that each axis moves the image in its own directions, while at low SNR, the axes' guidance fields become positively aligned. This means that information about coarse image structure is *shared* across axes rather than divided among them.



Figure 4: Latent axes with large effects on generated images encode coarse image structure. **a.** Mean squared error between images decoded with each latent axis at its 2nd and 98th percentile (mean ± st.dev. over four test images), for the 30 axes with the largest effect. **b.** The six largest-effect axes (rows, top to bottom) for four test images: images at the 2nd and 98th percentile and their difference (red, brighter; blue, darker). **c.** Effect size against each axis's information-weighted mean log SNR; red crosses mark noise-level axes, excluded from the correlation. **d.** Squared norm of the full guidance field (solid) and the sum of per-axis squared norms (dashed); where they coincide, per-axis information rates add up to the total. **e.** Signal and posterior variance, averaged over latent axes.

Encoder uncertainty overtakes image-driven variation below log SNR $\approx$ 0. Averaged over axes, the posterior variance grows from about 0.2 on clean images to several hundred at pure noise. Meanwhile the signal variance, the variance of the posterior mean across noisy test images, stays roughly constant. The two cross near $\log \text{SNR} \approx 0$ (Fig. 4e). The signal variance levels off above zero at high noise because each image is paired with its own noise sample, so it also reflects how the encoder responds to the noise itself.

An empirical estimate of total information If we assume the latents are roughly Gaussian distributed, we can calculate the approximate total information via the ratio of the signal variance and the posterior variance (inset in Fig. 3f).

The encoder maps an image x onto a Gaussian posterior for each axis as

$$z_d|x \sim \mathcal{N}(\mu_d(x), \sigma_d^2(x)).$$

By the law of total variance, the variance of z_d across images is

$$\mathrm{Var}(z_d) = \mathrm{Var}_x[\mu_d(x)] + \mathbb{E}_x[\sigma_d^2(x)] = S_d + N_d,$$

where S_d is the *signal variance* (i.e. the variance of the clean image latent means) and N_d is the posterior variance. To calculate the channel capacity, we start with the definition of mutual information:

$$I(z_d; x) = h(z_d) - h(z_d|x)$$

The Gaussian assumption on both distributions give us:

$$h(z_d) \approx \frac{1}{2}\log 2\pi e(S_d + N_d), \quad h(z_d|x) = \frac{1}{2}\log 2\pi e N_d$$

Subtracting gives

$$I(z_d; x) \approx \frac{1}{2}\log\left(1 + \frac{S_d}{N_d}\right).$$