Mutual Information Collapse Explains Disentanglement Failure in β-VAEs

Anonymous authors
Paper under double-blind review

Abstract

The β-VAE is a foundational framework for unsupervised disentanglement, utilizing a regularization parameter β to balance latent factorization against reconstruction fidelity. However, disentanglement performance often exhibits a non-monotonic dependence on β: standard metrics, such as MIG and SAP, typically peak at intermediate values and deteriorate under stronger regularization. We characterize this phenomenon as informational collapse—an information-theoretic failure in which excessive regularization drives the mutual information between latent variables and ground-truth generative factors toward zero. By analyzing the stationarity conditions in a linear-Gaussian setting, we prove that for $\beta > 1$, alternating optimization induces a spectral contraction of the encoder gain. This leads to an exponential decay of its spectral norm and the subsequent vanishing of latent–factor mutual information. To mitigate this failure mode, we investigate the $\lambda\beta$-VAE, which augments the objective with an auxiliary L_2 reconstruction penalty. Our analysis demonstrates that this term modifies the encoder stationarity conditions to counteract spectral decay, thereby stabilizing information flow within the latent representation. Extensive experiments on dSprites, Shapes3D, and MPI3D-real confirm that $\lambda > 0$ enhances the stability of disentanglement and preserves latent informativeness across a significantly broader range of β, providing a principled justification for dual-parameter regularization in variational inference.

1 Introduction

Disentangled representation learning aims to map high-dimensional observations into a latent space where individual dimensions correspond to distinct, often statistically independent, generative factors (Bengio et al., 2014; Tschannen et al., 2018; Wang et al., 2024). This structural alignment is critical for interpretability in scientific domains—such as genomics, robotics, and physics—where latent variables are expected to reflect underlying physical or biological mechanisms (Eguchi et al., 2022; Jiao et al., 2024; Cropsal & Mercado, 2025). Variational autoencoders (VAEs) provide a principled probabilistic framework for learning such representations (Kingma & Welling, 2022), with the β-VAE (Higgins et al., 2017) serving as the canonical baseline. By upweighting the Kullback–Leibler (KL) divergence term in the evidence lower bound (ELBO), the β-VAE imposes an information bottleneck that promotes latent factorization and enables unsupervised disentanglement under suitable conditions.

However, the β-VAE exhibits a well-documented failure mode under strong regularization. Specifically, it is prone to *posterior collapse* (He et al., 2019; Lucas et al., 2019; Razavi et al., 2019; Wang & Ziyin, 2022; Ichikawa & Hukushima, 2024; 2025), where the approximate posterior converges to the prior, rendering the latent variables uninformative. In this work, we adopt a broader perspective and refer to this phenomenon, as well as the more general degradation of latent information, as *informational collapse*. Empirically, disentanglement performance—measured by metrics such as the Mutual Information Gap (MIG) (Chen et al., 2018) and Separated Attribute Predictability (SAP) (Kumar et al., 2018)—typically exhibits a non-monotonic dependence on β: performance improves at moderate values but deteriorates as β increases (Locatello et al., 2019; Schott et al., 2022). While this trend is often attributed to a trade-off between reconstruction fidelity and latent factorization, a mechanistic explanation for the loss of factor-relevant information in high-regularization regimes remains limited.

We argue that the utility of disentanglement metrics fundamentally depends on *latent informativeness*—the extent to which latent variables retain information regarding the ground-truth generative factors. When the latent–factor mutual information becomes negligible, metrics like MIG and SAP lose discriminative power and fail to distinguish between structured and uninformative representations. Understanding this regime is essential for diagnosing instability in disentanglement and for addressing broader challenges in dimensionality reduction, where preventing the collapse of learned representations is a primary concern (Kalantidis et al., 2022; Baptista et al., 2025).

To address this gap, we provide a formal analysis of β-VAE optimization dynamics within a tractable linear-Gaussian setting. We prove that for $\beta > 1$, the stationarity conditions induce a *spectral contraction* of the encoder gain, causing its spectral norm to decay exponentially under alternating optimization. Consequently, the iterates converge to a unique trivial fixed point where the mutual information between the latent representation and generative factors vanishes. Motivated by this analysis, we investigate the $\lambda\beta$-VAE (Vu, 2025), which augments the objective with an auxiliary L_2 reconstruction penalty weighted by λ. We demonstrate that this term modifies the encoder stationarity conditions to counteract the decay of the encoder gain, thereby stabilizing information flow and mitigating informational collapse even under strong regularization.

Summary of Contributions

- **Failure Mode of Disentanglement Metrics:** We demonstrate that standard benchmarks, such as MIG and SAP, suffer from a critical failure mode where they lose discriminative power as latent–factor mutual information vanishes. This provides a principled explanation for the empirical instability and reproducibility challenges widely reported in high-regularization regimes (Abdi et al., 2019; Fil et al., 2021; Carbonneau et al., 2024).
- **Mechanism of Informational Collapse:** Utilizing a linear-Gaussian framework, we prove that for $\beta > 1$, optimization dynamics induce a spectral contraction of the encoder gain. This is characterized by the exponential decay of its spectral norm under alternating optimization, forcing the system toward a unique trivial fixed point. This establishes a mechanism for why β-VAEs are inherently prone to informational collapse.
- **Stability via the $\lambda\beta$-VAE Framework:** Building on our analysis, we investigate the $\lambda\beta$-VAE as a structural remedy to preserve latent informativeness. We prove that the inclusion of a λ-weighted L_2 reconstruction penalty fundamentally alters the encoder's stationarity conditions. This modification introduces a restorative force that counteracts spectral decay, effectively stabilizing information flow through the bottleneck even under stringent regularization constraints.
- **Empirical Validation:** We verify our theoretical predictions across diverse benchmarks, including dSprites (Matthey et al., 2017), Shapes3D (Burgess & Kim, 2018), and MPI3D-real (Gondal et al., 2019). Our results confirm that the $\lambda\beta$-VAE framework significantly extends the operative regime of informative latent representations, yielding superior stability in both reconstruction and disentanglement performance relative to standard VAE baselines.

2 Related Work

VAE Regularization and Trade-offs. The β-VAE (Higgins et al., 2017) leverages an upweighted KL divergence to foster latent factorization by constraining the capacity of the latent information bottleneck. Building on this formulation, β-TCVAE (Chen et al., 2018) and FactorVAE (Kim & Mnih, 2018) decompose the ELBO to isolate and penalize the total correlation of the aggregated posterior, whereas DIP-VAE (Kumar et al., 2018) enforces disentanglement via moment-matching. Further extensions explore alternative latent structures, such as discrete variables via stochastic quantization (Takida et al., 2022) or hyperbolic geometries for hierarchical representations (Cho et al., 2023). Collectively, these methods seek to enforce latent independence by imposing structural constraints on the latent distribution. However, empirical evidence suggests that strong regularization can drive models into a regime of posterior collapse, where latent informativeness is sacrificed to satisfy the prior constraint, leading to the degraded and unstable representations observed in high-regularization settings (Fil et al., 2021). This highlights a critical trade-off between achieving latent independence and preserving factor-relevant information.

Challenges in Unsupervised Disentanglement. Unsupervised disentanglement is fundamentally non-identifiable without specific inductive biases or structural assumptions regarding the data-generating process (Esmaeili et al., 2019; Locatello et al., 2019; Schott et al., 2022). Within the β-VAE framework, this limitation manifests as a heightened sensitivity to the regularization parameter β: increasing the weight of the KL divergence constrains the informational capacity of the latent channel, driving the model toward factorized aggregated posteriors at the potential expense of reconstruction fidelity. Recent high-dimensional analyses (Ichikawa & Hukushima, 2024; 2025) identify phase-transition regimes in which strong regularization triggers posterior collapse, a state where the approximate posterior converges to the prior and becomes uninformative. Prior research has attributed this failure mode to the inherent properties of the ELBO objective (Lucas et al., 2019), the competition between likelihood and prior regularization (Wang & Ziyin, 2022), and the lagging dynamics of the inference network (He et al., 2019). However, a precise characterization of how optimization dynamics interact with encoder parameterization to induce a systematic loss of latent information remains limited.

Robustness of Disentanglement Evaluation. Disentangled representations are typically evaluated using metrics such as SAP (Kumar et al., 2018), MIG (Chen et al., 2018), and DCI (Eastwood & Williams, 2018), which quantify the alignment between latent variables and ground-truth generative factors via predictability, mutual information, or feature-importance measures. These metrics implicitly assume that the latent representation retains sufficient information regarding the underlying factors to serve as a meaningful signal for evaluation. While prior research has examined these properties from an information-theoretic perspective (Burgess et al., 2018), their behavior in low-informativeness regimes remains poorly understood. Specifically, when the latent–factor mutual information vanishes, these metrics lose discriminative power and fail to reliably distinguish between structured, disentangled representations and uninformative noise. In this work, we demonstrate that this failure mode is a direct consequence of the optimization dynamics that drive informational collapse.

Information Preservation in Generative Models. VAEs serve as foundational representation backbones in generative pipelines, such as latent diffusion models (Rombach et al., 2022; Pynadath et al., 2025), where latent quality is a primary determinant of downstream performance. Empirical studies across diverse domains—including protein modeling (Eguchi et al., 2022), robotics (Abdulsamad et al., 2024), 3D shape representation (Jiao et al., 2024), and drug discovery (Cropsal & Mercado, 2025)—consistently report performance degradation under strong regularization, a phenomenon frequently attributed to reduced latent informativeness. To address these instabilities, recent variants such as the $\lambda\beta$-VAE and $\gamma\beta$-VAE (Vu, 2025) introduce auxiliary structural objectives. The $\lambda\beta$-VAE incorporates an L_2 reconstruction penalty to stabilize the reconstruction signal in high-β regimes without compromising disentanglement. Conversely, the $\gamma\beta$-VAE utilizes a mutual information penalty within re-encoding cycles to reinforce factorization at lower β values while maintaining high reconstruction fidelity.

3 Linear-Gaussian Framework

We leverage the tractability of the linear-Gaussian regime to formalize the informational collapse hypothesized in Section 1. This framework allows for a closed-form characterization of the stationarity conditions and the resulting optimization dynamics.

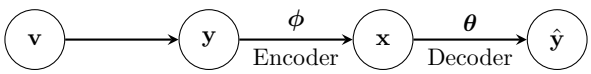


Figure 1: The linear-Gaussian generative and inference process. Generative factors $\mathbf{v}$ map to observations $\mathbf{y}$ via $\mathbf{\Gamma}$. The encoder $(\mathbf{B}, \mathbf{\Sigma_w})$ transforms $\mathbf{y} \to \mathbf{x}$, while the decoder $(\mathbf{A}, \mathbf{\Sigma_z})$ reconstructs $\hat{\mathbf{y}}$ from $\mathbf{x}$.

3.1 Generative Model

We model the informational flow within the VAE as a Markov chain $\mathbf{v} \to \mathbf{y} \to \mathbf{x} \to \hat{\mathbf{y}}$ (Fig. 1). Let $\mathbf{v} \in \mathbb{R}^s$ denote the ground-truth generative factors, distributed as $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_v})$, where $\mathbf{\Sigma_v} \in \mathbb{S}^s_{++}$ is diagonal and

positive definite. Observations $\mathbf{y} \in \mathbb{R}^n$ are generated via a linear mixing matrix $\mathbf{\Gamma} \in \mathbb{R}^{n \times s}$ with additive Gaussian noise:

$$\mathbf{y} = \mathbf{\Gamma}\mathbf{v} + \tilde{\mathbf{z}}, \qquad \tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n), \tag{1}$$

yielding the observation covariance

$$\mathbf{\Sigma_y} = \mathbf{\Gamma}\mathbf{\Sigma_v}\mathbf{\Gamma}^\top + \sigma^2 \mathbf{I}_n.$$

The latent representation $\mathbf{x} \in \mathbb{R}^m$ is subject to an isotropic Gaussian prior $p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. We define a Gaussian encoder $q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y})$ parameterized by the encoder gain $\mathbf{B} \in \mathbb{R}^{m \times n}$ and noise covariance $\mathbf{\Sigma_w} \in \mathbb{S}^m_{++}$:

$$\mathbf{x} = \mathbf{B}\mathbf{y} + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_w}). \tag{2}$$

The decoder $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ reconstructs the observations via the decoder gain $\mathbf{A} \in \mathbb{R}^{n \times m}$ and noise covariance $\mathbf{\Sigma_z} \in \mathbb{S}^n_{++}$:

$$\hat{\mathbf{y}} = \mathbf{A}\mathbf{x} + \mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_z}). \tag{3}$$

The statistical relationship between the latent variables and generative factors is captured by the joint covariance:

$$\mathbf{\Sigma} = \mathrm{Cov}\left(\begin{bmatrix}\mathbf{x} \\ \mathbf{v}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{\Sigma_x} & \mathbf{\Sigma_{xv}} \\ \mathbf{\Sigma_{vx}} & \mathbf{\Sigma_v}\end{bmatrix} = \begin{bmatrix}\mathbf{B}\mathbf{\Sigma_y}\mathbf{B}^\top + \mathbf{\Sigma_w} & \mathbf{B}\mathbf{\Gamma}\mathbf{\Sigma_v} \\ \mathbf{\Sigma_v}\mathbf{\Gamma}^\top\mathbf{B}^\top & \mathbf{\Sigma_v}\end{bmatrix}. \tag{4}$$

3.2 Variational Objectives

The standard β-VAE objective is formulated as a weighted ELBO, which we minimize with respect to the encoder and decoder parameters:

$$\mathcal{L}_\beta(\boldsymbol{\phi}, \boldsymbol{\theta}) = \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y})}\big[-\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})\big]}_{\text{Reconstruction}} + \beta \underbrace{\mathrm{D}_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y}) \,\|\, p(\mathbf{x})\big)}_{\text{Regularization}}. \tag{5}$$

In the linear-Gaussian regime, this objective admits a closed-form analytical expansion:

$$\begin{aligned}\mathcal{L}_\beta(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_z}, \mathbf{\Sigma_w}) = -\frac{1}{2}\Bigg(&\mathrm{Tr}\Big[\mathbf{A}^\top\mathbf{\Sigma_z}^{-1}\mathbf{\Sigma_y}\mathbf{B}^\top + \mathbf{\Sigma_z}^{-1}\mathbf{A}\mathbf{B}\mathbf{\Sigma_y} - \mathbf{\Sigma_z}^{-1}\mathbf{\Sigma_y} \\ &- \mathbf{A}^\top\mathbf{\Sigma_z}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma_y}\mathbf{B}^\top + \mathbf{\Sigma_w})\Big] - n\log(2\pi) - \log|\mathbf{\Sigma_z}|\Bigg) \\ &+ \frac{\beta}{2}\Big[\mathrm{Tr}(\mathbf{B}\mathbf{\Sigma_y}\mathbf{B}^\top + \mathbf{\Sigma_w}) - \log|\mathbf{\Sigma_w}| - m\Big].\end{aligned} \tag{6}$$

As demonstrated in Section 3.3, for $\beta > 1$, alternating optimization dynamics induce a spectral contraction of the encoder gain, driving $\mathbf{B} \to \mathbf{0}$. This informational collapse renders the latent channel uninformative, causing evaluation metrics to lose their discriminative power as they can no longer distinguish structured representations from uninformative noise. To mitigate this failure mode, we investigate the $\lambda\beta$-VAE formulation, which augments the objective with an auxiliary L_2 reconstruction penalty weighted by λ. The primary goal of this term is to enhance reconstruction accuracy in high-β regimes while maintaining latent disentanglement. Simultaneously, we demonstrate that this intervention provides a theoretical basis for mitigating informational collapse by modifying the system's stationarity conditions. Specifically, we utilize the reconstruction $\hat{\mathbf{y}} = \mathbf{A}\mathbf{x}$ by omitting the additive decoder noise $\mathbf{z}$. This approach follows established research suggesting that a deterministic decoder can effectively reduce blurriness and improve reconstruction fidelity (Ghosh et al., 2020; Bredell et al., 2023; Kim & Lee, 2025). This auxiliary term penalizes the squared error between the reconstructed output and the observation $\mathbf{y}$:

$$\mathcal{L}_{\lambda\beta}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathcal{L}_\beta(\boldsymbol{\phi}, \boldsymbol{\theta}) + \lambda\, \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\big[\|\mathbf{y} - \hat{\mathbf{y}}\|^2\big], \qquad \lambda \geq 0, \tag{7}$$

where the expectation is taken over the joint distribution $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{y}) = q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y})p(\mathbf{y})$. Substituting the encoder relation $\mathbf{x} = \mathbf{B}\mathbf{y} + \mathbf{w}$ into the auxiliary term yields the augmented linear-Gaussian objective:

$$\mathcal{L}_{\lambda\beta}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_z}, \mathbf{\Sigma_w}) = \mathcal{L}_\beta(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_z}, \mathbf{\Sigma_w}) + \lambda\, \mathrm{Tr}\Big[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{\Sigma_y}(\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top + \mathbf{A}\mathbf{\Sigma_w}\mathbf{A}^\top\Big]. \tag{8}$$

The full derivations for these closed-form expressions are provided in Appendix A.

3.3 Informational Collapse

We characterize informational collapse through its manifestation in the linear-Gaussian regime. Specifically, this phenomenon corresponds to the asymptotic vanishing of the latent signal at the objective's stationary points. These points are governed by a system of coupled stationarity conditions, as formalized in the following lemma:

Lemma 3.1 (β-VAE Stationarity Conditions)**.** *Any stationary point of the β-VAE objective satisfies the following system of coupled equations for the decoder $(\mathbf{A}, \mathbf{\Sigma_z})$ and encoder $(\mathbf{B}, \mathbf{\Sigma_w})$:*

$$\mathbf{A} = (\mathbf{\Sigma_y}^{-1} + \mathbf{B}^\top \mathbf{\Sigma_w}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top \mathbf{\Sigma_w}^{-1} \qquad \mathbf{B} = \left[\mathbf{I}_m + \mathbf{A}^\top(\mathbf{\Sigma_z}^{-1}/\beta)\mathbf{A}\right]^{-1}\mathbf{A}^\top(\mathbf{\Sigma_z}^{-1}/\beta)$$
$$\mathbf{\Sigma_z} = (\mathbf{\Sigma_y}^{-1} + \mathbf{B}^\top \mathbf{\Sigma_w}^{-1}\mathbf{B})^{-1} \qquad \mathbf{\Sigma_w} = \left[\mathbf{I}_m + \mathbf{A}^\top(\mathbf{\Sigma_z}^{-1}/\beta)\mathbf{A}\right]^{-1}$$

Proof. See Appendix B.

The β^{-1} scaling in the encoder gain update serves as the primary driver of informational suppression. Since the optimal encoder and decoder parameters are mutually dependent, we resolve these conditions via the alternating optimization procedure detailed in Algorithm 1. This iterative process can be viewed as a continuous-variable analog of the Blahut–Arimoto algorithm (Arimoto, 1972; Blahut, 1972; Yang & Mandt, 2022), solving the closed-form stationarity condition for one parameter block while holding others fixed. For $\beta > 1$, this sequence of updates induces a spectral contraction of the encoder gain, leading to the following result:

Theorem 3.2 (Informational Collapse for $\beta > 1$)**.** *For any regularization strength $\beta > 1$, the alternating optimization dynamics described in Lemma 3.1 and Algorithm 1 converge to the trivial fixed point:*

$$(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_z}, \mathbf{\Sigma_w}) = (\mathbf{0}, \mathbf{0}, \mathbf{\Sigma_y}, \mathbf{I}_m).$$

Consequently, the encoder gain $\mathbf{B}$ vanishes, the latent representation $\mathbf{x}$ becomes statistically independent of the generative factors $\mathbf{v}$, and the mutual information decays to zero:

$$\lim_{t\to\infty} I(\mathbf{x}; \mathbf{v})^{(t)} = 0 \quad \textit{for } \beta > 1.$$

Proof. See Appendix D.

Theorem 3.2 provides a mechanistic explanation for the sharp decline in disentanglement performance observed in high-regularization regimes. In this state, the latent channel fails to preserve factor-relevant information, rendering the representation uninformative regardless of its factorization.

4 Evaluation Robustness and Informational Restoration

4.1 Analytical Metric Formulations

We consider three representative metrics to evaluate the structural integrity of the latent space: Separated Attribute Predictability (SAP) (Kumar et al., 2018), Mutual Information Gap (MIG) (Chen et al., 2018), and the Latent Informativeness Metric (LIM) (Vu, 2025). While MIG is widely utilized in nonlinear settings, LIM offers a principled alternative for continuous variables within the linear-Gaussian regime. All three metrics explicitly depend on the magnitude of the latent signal; consequently, they lose their discriminative power when the encoder gain undergoes spectral contraction.

Separated Attribute Predictability (SAP). The SAP score utilizes the squared correlation matrix $\mathbf{S} \in \mathbb{R}^{m\times s}$, where each element $S_{i,j}$ quantifies the linear predictability of generative factor v_j from latent dimension x_i:

$$S_{i,j} = \frac{[\mathbf{B\Gamma\Sigma_v}]^2_{i,j}}{[\mathbf{\Sigma_x}]_{i,i}[\mathbf{\Sigma_v}]_{j,j}}. \tag{9}$$

SAP is defined as the average difference between the two largest squared correlations for each factor:

$$\mathrm{SAP}(\mathbf{x}, \mathbf{v}) = \frac{1}{s} \sum_{j=1}^{s} \left(S_{i^{(j)},j} - S_{i'^{(j)},j} \right), \tag{10}$$

where $i^{(j)}$ and $i'^{(j)}$ denote the indices of the most and second-most predictive latent variables for factor v_j, respectively. As alternating optimization drives $\mathbf{B} \to \mathbf{0}$, the numerator in the correlation expression vanishes while the latent variances $[\mathbf{\Sigma_x}]_{i,i}$ remain at the prior noise level. Consequently, $S_{i,j} \to 0$ for all (i, j), causing the SAP score to converge to zero and lose its ability to distinguish between distinct latent structures.

Mutual Information Gap (MIG). MIG measures the normalized difference in mutual information between the top two latent candidates for each generative factor:

$$\mathrm{MIG}(\mathbf{x}, \mathbf{v}) = \frac{1}{s} \sum_{j=1}^{s} \frac{I(x_{i^{(j)}}; v_j) - I(x_{i'^{(j)}}; v_j)}{H(v_j)}, \tag{11}$$

where $H(v_j)$ represents the entropy of factor v_j. In the linear-Gaussian regime, mutual information is a monotonic function of the squared correlation:

$$I(x_i; v_j) = -\frac{1}{2} \log(1 - S_{i,j}).$$

As $\|\mathbf{B}\|_2 \to 0$ during informational collapse, $S_{i,j} \to 0$, which in turn drives $I(x_i; v_j) \to 0$, rendering the MIG score uninformative.

Latent Informativeness Metric (LIM). LIM quantifies the total factor-relevant information preserved within the latent space, accounting for potential statistical correlations between generative factors $\mathbf{v}$. It identifies an optimal partition $\mathcal{P} = \{v_{s_k}\}_{k=1}^{m}$ of the s factors into m disjoint subsets, each assigned to a latent dimension x_k:

$$\mathrm{LIM}(\mathbf{x}, \mathbf{v}) = \max_{\mathcal{P}} \left[\sum_{k=1}^{m} I(x_k; v_{s_k}) - \sum_{i<j} I(v_{s_i}; v_{s_j}) \right]. \tag{12}$$

The second term serves to remove informational redundancy among factor groups. In our controlled linear-Gaussian setting with independent factors ($\mathbf{\Sigma_v}$ diagonal), this redundancy term vanishes, reducing LIM to a maximum-weight assignment problem. Furthermore, the linear-Gaussian assumption enables the exact computation of $I(x_k; v_{s_k})$ via log-determinants of covariance sub-blocks, bypassing the numerical estimation issues that MIG can exhibit for continuous variables.

4.2 λ-Restoration of Latent Informativeness

In the standard β-VAE framework, alternating optimization dynamics for $\beta > 1$ induce a spectral contraction of the encoder gain, driving $\|\mathbf{B}\|_2 \to 0$ and causing informativeness-based metrics to lose their discriminative power. To counteract this collapse, the $\lambda\beta$-VAE objective introduces a structural intervention to the underlying stationarity conditions via an auxiliary L_2 reconstruction penalty.

Lemma 4.1 ($\lambda\beta$-VAE Stationarity Conditions)**.** *Let $\mathbf{M} \coloneqq \beta^{-1}(\mathbf{\Sigma_z}^{-1} + 2\lambda\mathbf{I}_n)$. At any stationary point of the $\lambda\beta$-VAE objective, the decoder parameters $(\mathbf{A}, \mathbf{\Sigma_z})$ satisfy the conditions established in Lemma 3.1, while the encoder parameters satisfy the following augmented system:*

$$\mathbf{B} = (\mathbf{I}_m + \mathbf{A}^\top \mathbf{M} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{M}$$
$$\mathbf{\Sigma_w} = (\mathbf{I}_m + \mathbf{A}^\top \mathbf{M} \mathbf{A})^{-1}.$$

Proof. See Appendix C.

The λ parameter serves as a restorative mechanism that modulates the spectral decay induced by the KL penalty. By embedding $2\lambda\mathbf{I}_n$ within the operator $\mathbf{M}$, the objective shifts the threshold of the informational phase transition, preserving latent informativeness over a significantly broader range of regularization

strengths. Crucially, this intervention ensures that the encoder gain remains non-trivial even when $\beta > 1$. However, as this effect is fundamentally constrained by the data covariance $\|\mathbf{\Sigma}_\mathbf{y}\|_2$, the global contraction factor β^{-1} eventually dominates as $\beta \to \infty$. In this asymptotic limit, the encoder gain still vanishes, indicating that while $\lambda > 0$ significantly extends the regime of informational stability, it does not preclude eventual collapse under infinite regularization. A comprehensive analysis of these limiting dynamics is provided in Appendix E.

5 Empirical Evaluation

We evaluate the $\lambda\beta$-VAE framework through two complementary studies: (i) controlled linear-Gaussian simulations to verify the stationarity-induced phase transitions derived in Section 3.3, and (ii) benchmarks on deep convolutional architectures to assess the robustness of informational restoration in nonlinear settings.[1]

5.1 Linear-Gaussian Regime

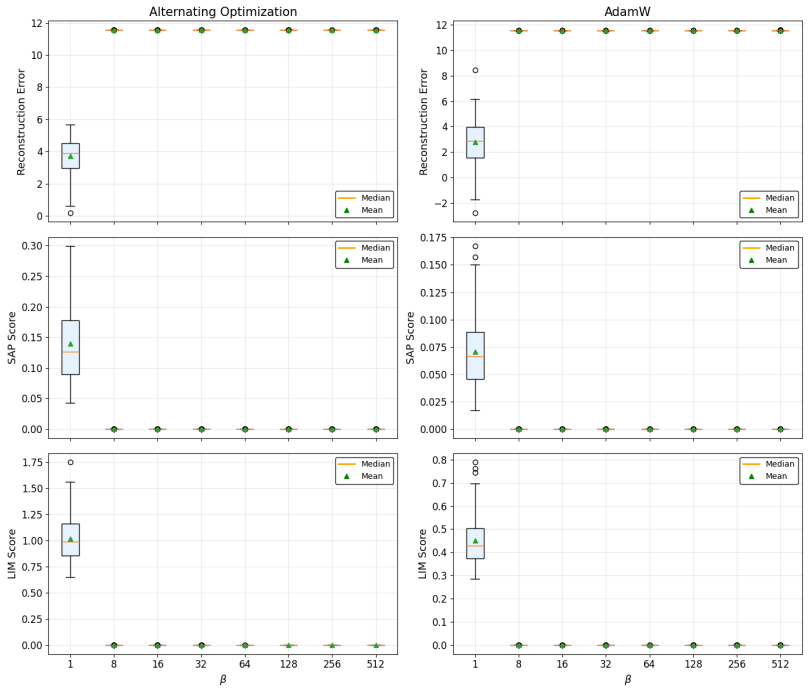


Figure 2: Reconstruction error, SAP, and LIM scores for the standard β-VAE ($n = 50, m = 10, s = 5$). Both alternating optimization and AdamW exhibit a sharp informational collapse for $\beta > 1$, where the vanishing encoder gain $\mathbf{B}$ drives latent–factor mutual information to zero.

We first validate the theoretical existence of the uninformative trivial solution for $\beta > 1$ using a high-dimensional configuration ($n = 50, m = 10, s = 5$). Generative factors $\mathbf{v}$ are sampled i.i.d. with observation noise variance $\sigma^2 = 0.05$. We compare two distinct optimization dynamics: **Alternating Optimization** (Algorithm 1), utilizing the closed-form stationarity conditions established in Lemma 3.1, and **AdamW** using Cholesky-parameterized covariance matrices. For each (β, λ) pair, we conduct 100 independent trials with randomized initializations $(\mathbf{B}^{(0)}, \mathbf{\Sigma}_\mathbf{w}^{(0)})$ to ensure statistical robustness.

Empirical Findings. As predicted by Theorem 3.2, the baseline β-VAE ($\lambda = 0$) undergoes an abrupt phase transition (Fig. 2). For $\beta > 1$, both alternating and stochastic trajectories converge to the unique trivial fixed point, where the encoder gain $\mathbf{B}$ undergoes spectral contraction to zero, causing SAP and LIM

[1]Code repository: https://github.com/mh-vu/lambda-beta-vae

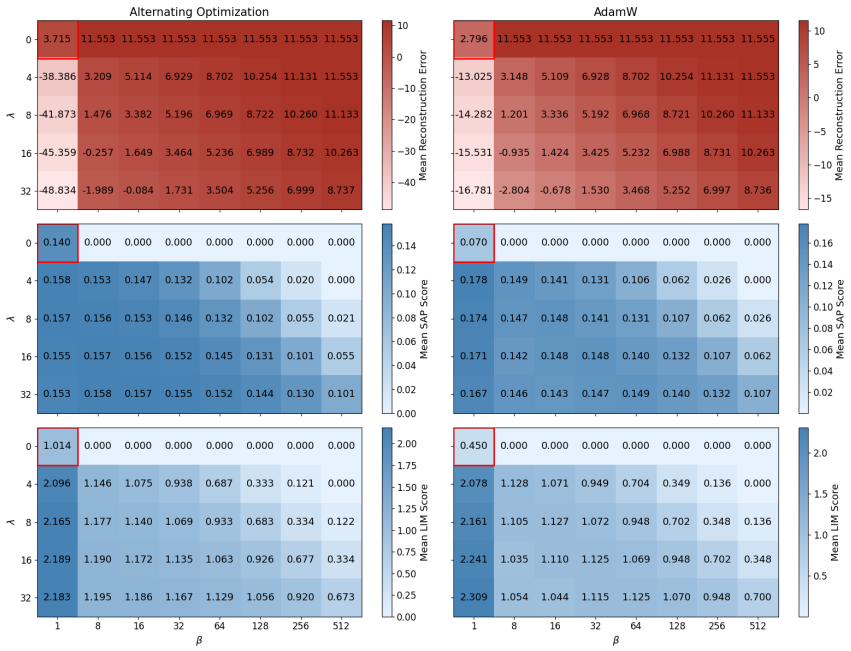


Figure 3: Reconstruction error, SAP, and LIM scores for the $\lambda\beta$-VAE ($n = 50, m = 10, s = 5$). Introducing a positive λ counteracts spectral contraction, substantially extending the informational stability range. At extreme β, the global contraction factor β^{-1} eventually dominates, leading to asymptotic informational collapse.

scores to vanish. Introducing $\lambda > 0$ fundamentally alters the stationarity conditions (Lemma 4.1), acting as a restorative force that counteracts this contraction (Fig. 3). Under the $\lambda\beta$-VAE objective, latent–factor coupling $\mathbf{\Sigma_{xv}}$ is preserved even under high regularization ($\beta = 512$), significantly extending the regime of informational stability. However, this restorative effect is ultimately bounded by the data covariance $\|\mathbf{\Sigma_y}\|_2$. In the limit $\beta \to \infty$, the global contraction factor β^{-1} dominates, resulting in asymptotic informational collapse. These results confirm that while λ shifts the informational threshold, it does not preclude eventual signal decay under extreme regularization.

5.2 Deep Nonlinear Architectures

To evaluate the generalizability of our theoretical analysis, we extend our study to high-dimensional visual domains: dSprites (Matthey et al., 2017), Shapes3D (Burgess & Kim, 2018), and MPI3D-real (Gondal et al., 2019). The MPI3D-real benchmark, in particular, introduces realistic visual correlations and complex backgrounds, which have been shown to accelerate informational decay under over-regularization (Schott et al., 2022).

Table 1: Convolutional VAE architecture for nonlinear experiments.

Encoder	**Decoder**
Input: $64 \times 64 \times C$	Input: $\mathbf{x} \in \mathbb{R}^m$
4×4 Conv, 32, stride 2, ReLU	FC 256, ReLU
4×4 Conv, 32, stride 2, ReLU	FC $4 \times 4 \times 64$, ReLU
4×4 Conv, 64, stride 2, ReLU	4×4 ConvTranspose, 64, stride 2, ReLU
4×4 Conv, 64, stride 2, ReLU	4×4 ConvTranspose, 32, stride 2, ReLU
FC 256, ReLU	4×4 ConvTranspose, 32, stride 2, ReLU
FC $2m$ $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$	4×4 ConvTranspose, C, stride 2

Table 2: Hyperparameter configurations for nonlinear experiments.

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	10^{-4}
Batch size	64
Latent dimension (m)	15
β grid	$\{1, 4, 8, 16, 32, 64, 128, 256\}$
λ grid	$\{0, 4, 8, 16, 32\}$
β-VAE training steps	150,000
$\lambda\beta$-VAE training steps	50,000
Seeds	10

Architecture and Training Protocol. We employ a deep convolutional VAE with a 15-dimensional latent space, utilizing a Gaussian encoder and a Bernoulli decoder following the architectural configurations of Locatello et al. (2019). Detailed specifications and hyperparameters are provided in Tables 1 and 2. To isolate the impact of the structural intervention, we adopt a two-phase training protocol:

1. **Stabilization Phase:** Models are initially trained as standard β-VAEs ($\lambda = 0$) for 150,000 steps across the β-grid to establish a baseline for latent factorization under standard regularization pressure.
2. **Restoration Phase:** The λ parameter is introduced, and training continues for an additional 50,000 steps under the $\lambda\beta$-VAE objective.

This sequential protocol evaluates whether the λ intervention can restore latent informativeness in representations already degraded by informational collapse, while ensuring the auxiliary penalty does not disrupt the learned factorization dynamics.

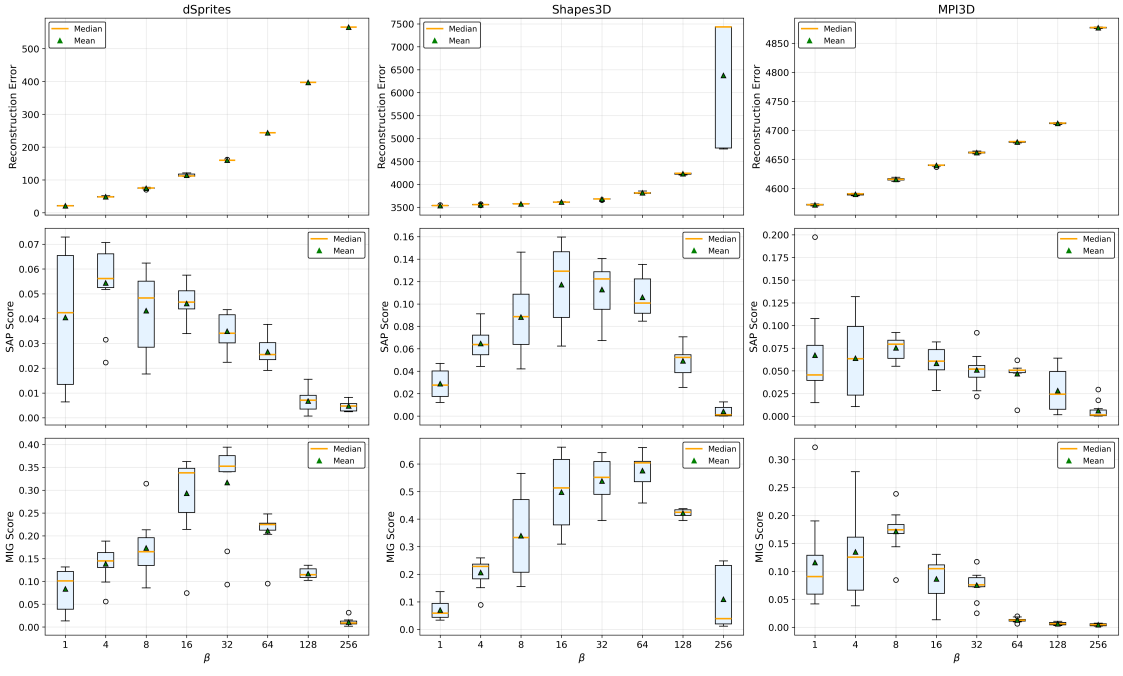


Figure 4: Reconstruction error, SAP, and MIG scores for the standard β-VAE across benchmarks. Disentanglement performance typically peaks at intermediate β values and deteriorates under high regularization.

Empirical Findings. As illustrated in Fig. 4, all datasets exhibit the predicted high-β performance degradation. While the transition is more gradual than in the linear-Gaussian case, SAP and MIG scores confirm

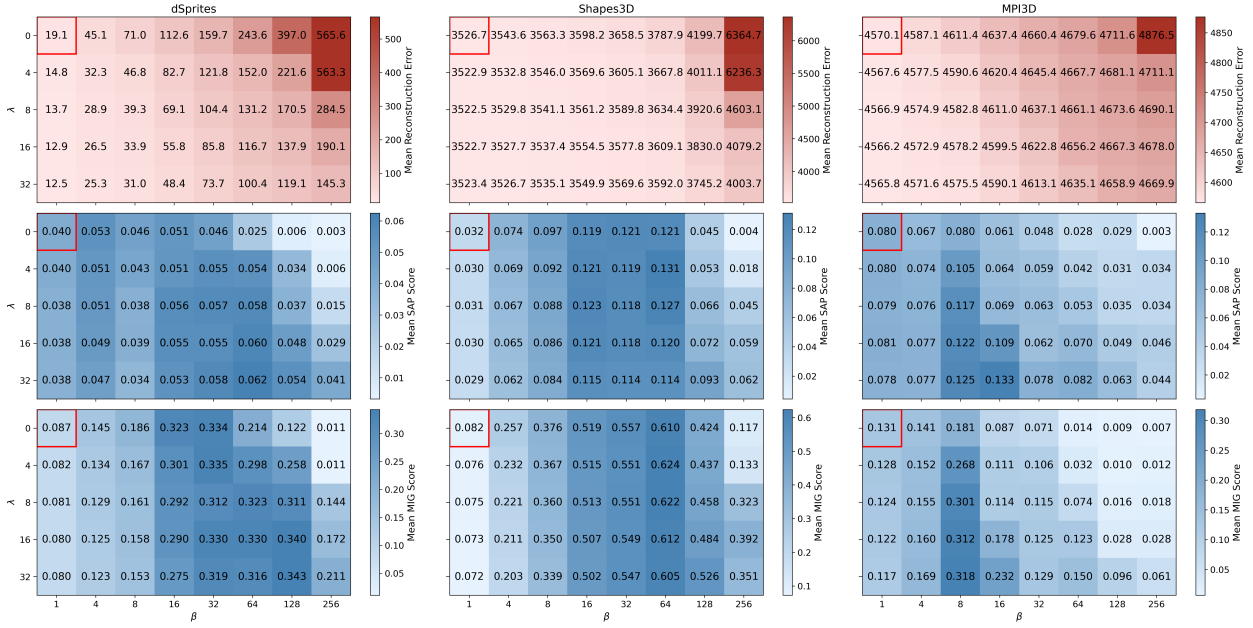


Figure 5: Reconstruction error, SAP, and MIG scores for the $\lambda\beta$-VAE. Introducing $\lambda > 0$ preserves latent informativeness under high β, substantially expanding the operative regime of the representation. While performance eventually declines as $\beta \to \infty$, this restorative mechanism significantly delays the onset of informational collapse.

that excessive KL pressure systematically erases factor-relevant information. This effect is most pronounced in MPI3D-real, consistent with findings that increased dataset complexity lowers the threshold for informational collapse (Ichikawa & Hukushima, 2025). Heatmap analysis (Fig. 5) demonstrates that λ provides a robust restorative effect across all benchmarks. For models nearing the collapse regime at high β, introducing $\lambda > 0$ stabilizes disentanglement metrics while simultaneously improving reconstruction fidelity. These results indicate that the $\lambda\beta$-VAE effectively preserves the latent signal, substantially expanding the hyperparameter regime in which the learned representation remains informative.

6 Hyperparameter Selection Strategy

Practical deployment of the $\lambda\beta$-VAE requires a systematic method for navigating the (β, λ) hyperparameter space to identify optimal operating points. We formalize this selection as a multi-objective optimization problem, seeking configurations that balance the competing demands of reconstruction fidelity and latent factorization. We define the objective vector $\mathbf{f} = [f_1, f_2]^\top$, where f_1 denotes the reconstruction error (negative log-likelihood) and f_2 represents the *entanglement cost*, $1 - \text{MIG}$.

To navigate the trade-offs between f_1 and f_2, we employ **augmented Tchebycheff scalarization** (Steuer & Choo, 1983; Miettinen, 1999; Dächert et al., 2012; Chugh, 2020). Each objective f_i is first normalized over the evaluation grid $\mathcal{G}$ to ensure scale invariance:

$$\bar{f}_i(\beta, \lambda) = \frac{f_i(\beta, \lambda) - \min_{\mathcal{G}} f_i}{\max_{\mathcal{G}} f_i - \min_{\mathcal{G}} f_i}, \quad i \in \{1, 2\}. \tag{13}$$

Given a preference vector $\boldsymbol{\omega}$ (where $\omega_1 + \omega_2 = 1$), we seek to minimize the scalarized selection score $\mathcal{S}$:

$$\mathcal{S}(\beta, \lambda; \boldsymbol{\omega}) = \max_{i \in \{1,2\}} \{\omega_i \bar{f}_i(\beta, \lambda)\} + \rho \sum_{i=1}^{2} \omega_i \bar{f}_i(\beta, \lambda), \tag{14}$$

where $\rho = 10^{-3}$ is a small augmentation parameter used to guarantee strict Pareto optimality.

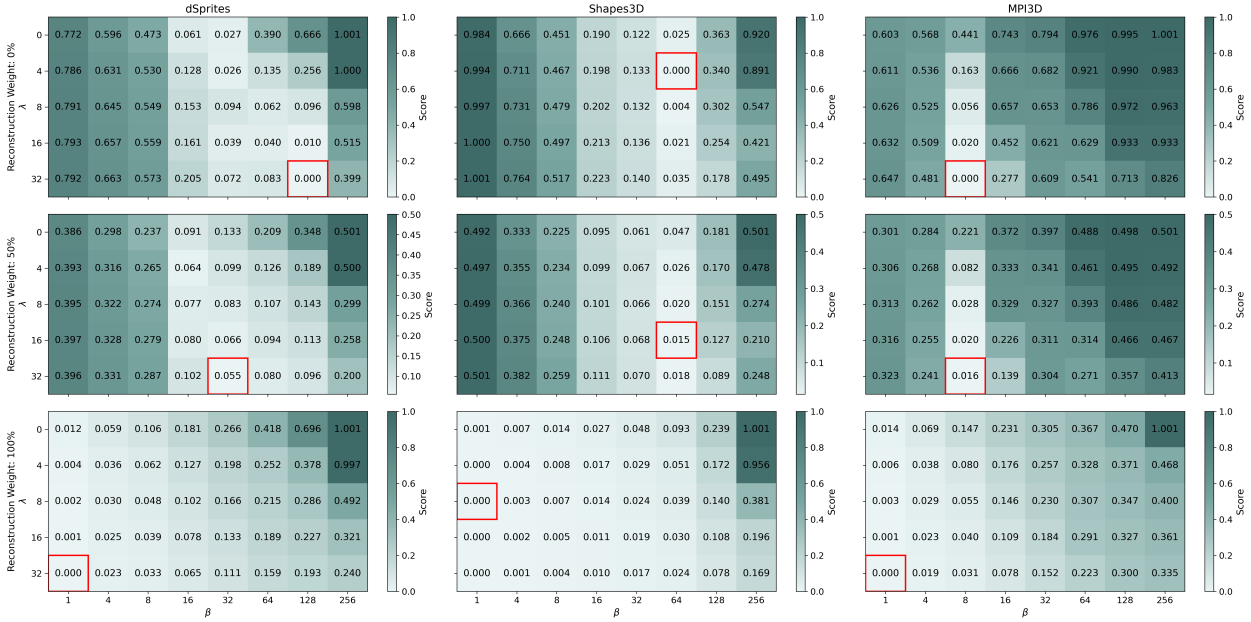


Figure 6: Augmented Tchebycheff scalarization heatmaps for three distinct preference profiles. Red markers indicate the optimal (β, λ) configurations that minimize the scalarized score over the evaluation grid $\mathcal{G}$.

Empirical Findings. Fig. 6 visualizes these selection scores under three representative preference profiles. As the priority shifts from reconstruction to disentanglement, the optimal configuration transitions from low-β to high-β regimes. Crucially, within the over-regularized regime, choosing $\lambda > 0$ recovers substantial reconstruction fidelity while maintaining latent disentanglement. Adjusting the preference vector $\boldsymbol{\omega}$ thus allows the scalarization procedure to identify the specific (β, λ) pair that satisfies the desired trade-off between reconstruction quality and latent factorization.

7 Conclusion

This work identifies a fundamental pathology in variational representation learning: the informational collapse of the latent channel under strong regularization. We demonstrate that the widely observed non-monotonic behavior of β-VAE performance is not merely an empirical artifact but arises directly from the stationarity conditions of the objective. Within a linear-Gaussian framework, we prove that for $\beta > 1$, alternating optimization dynamics induce a contraction mapping driven by the spectral contraction of the encoder gain, ultimately forcing latent–factor mutual information to vanish. This characterization provides a principled explanation for the breakdown of standard disentanglement metrics, such as MIG and SAP, whose validity critically depends on the preservation of a non-collapsed latent signal.

Building on this diagnosis, we investigate the $\lambda\beta$-VAE as a structural mechanism to mitigate collapse. By decoupling the pressure for latent independence from the requirement for informational integrity, the auxiliary reconstruction weight λ modifies the encoder stationarity conditions to act as a restorative term against spectral decay. Our theoretical analysis and empirical evaluations across dSprites, Shapes3D, and MPI3D-real demonstrate that this intervention effectively counteracts contraction, substantially widening the operative regime in which the latent representation remains informative. We emphasize, however, that preserving latent informativeness does not resolve the inherent non-identifiability of unsupervised disentanglement (Locatello et al., 2019); rather, it ensures the representation retains the statistical signal necessary for downstream tasks or causal discovery to operate meaningfully.

These findings suggest several promising directions for future research. First, the fixed hyperparameter setting could be extended to adaptive scheduling strategies, where λ is dynamically adjusted to maintain a target level of mutual information via control-theoretic approaches (Shao et al., 2020). Second, the proposed

framework should be explored within modern generative pipelines, such as latent diffusion models (Rombach et al., 2022; Pynadath et al., 2025), where latent informativeness directly governs downstream controllability. Finally, extending the spectral contraction analysis beyond the linear-Gaussian setting to high-dimensional asymptotics (Ichikawa & Hukushima, 2024; 2025; Rahimi et al., 2025) and alternative latent geometries (Cho et al., 2023; Pynadath et al., 2025) remains a vital step toward a unified theory of latent informativeness. Overall, our results establish a critical design principle for stable representation learning: while β modulates the strength of latent factorization, a decoupled mechanism is required to preserve the informational integrity of the latent channel.

References

Amir H. Abdi, Purang Abolmaesumi, and Sidney S. Fels. A preliminary study of disentanglement with insights on the inadequacy of metrics. *CoRR*, abs/1911.11791, 2019. URL http://arxiv.org/abs/1911.11791.

Hany Abdulsamad, Peter Nickl, Pascal Klink, and Jan Peters. Variational hierarchical mixtures for probabilistic learning of inverse dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):1950–1963, 2024. doi: 10.1109/TPAMI.2023.3314670.

S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972. doi: 10.1109/TIT.1972.1054753.

Ricardo Baptista, Michael Brennan, and Youssef Marzouk. Dimension reduction via score ratio matching. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=mvbZBaqSXo.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014. URL https://arxiv.org/abs/1206.5538.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.

Gustav Bredell, Kyriakos Flouris, Krishna Chaitanya, Ertunc Erdil, and Ender Konukoglu. Explicitly minimizing the blur error of variational autoencoders. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9krnQ-ue9M.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018. URL https://arxiv.org/abs/1804.03599.

Marc-André Carbonneau, Julian Zaïdi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):8747–8761, 2024. doi: 10.1109/TNNLS.2022.3218982.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Seunghyuk Cho, Juyong Lee, and Dongwoo Kim. Hyperbolic VAE via latent gaussian distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=FNn4zibGvw.

Tinkle Chugh. Scalarizing functions in bayesian multiobjective optimization. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2020. doi: 10.1109/CEC48606.2020.9185706.

Télio Cropsal and Rocío Mercado. Compressing biology: Evaluating the stable diffusion VAE for phenotypic drug discovery. In *NeurIPS 2025 Workshop for Imageomics: Discovering Biological Knowledge from Images Using AI*, 2025. URL https://openreview.net/forum?id=ZgVGpQzmWy.

Kerstin Dächert, Jochen Gorski, and Kathrin Klamroth. An augmented weighted tchebycheff method with adaptively chosen parameters for discrete bicriteria optimization problems. *Computers Operations Research*, 39(12):2929–2943, 2012. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2012.02.021. URL https://www.sciencedirect.com/science/article/pii/S0305054812000470.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Raphael R. Eguchi, Christian A. Choe, and Po-Ssu Huang. Ig-vae: Generative modeling of protein structure by direct 3d coordinate generation. *PLOS Computational Biology*, 18(6):1–18, 06 2022. doi: 10.1371/journal.pcbi.1010271. URL https://doi.org/10.1371/journal.pcbi.1010271.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2525–2534. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/esmaeili19a.html.

Miroslav Fil, Munib Mesinovic, Matthew Morris, and Jonas Wildberger. Beta-vae reproducibility: Challenges and extensions, 2021. URL https://arxiv.org/abs/2112.14278.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Scholkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g7tpEYDS.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rylDfnCqF7.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981. ISSN 00361445. URL http://www.jstor.org/stable/2029838.

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*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear VAE: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with KL annealing. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1936–1944. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/ichikawa24a.html.

Yuma Ichikawa and Koji Hukushima. High-dimensional asymptotics of vaes: threshold of posterior collapse and dataset-size dependence of rate-distortion curve. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073402, jul 2025. doi: 10.1088/1742-5468/adde3e. URL https://doi.org/10.1088/1742-5468/adde3e.

Yining Jiao, Carlton Jude ZDANSKI, Julia S Kimbell, Andrew Prince, Cameron P Worden, Samuel Kirse, Christopher Rutter, Benjamin Shields, William Alexander Dunn, Jisan Mahmud, and Marc Niethammer. Naisr: A 3d neural additive model for interpretable shape representation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wg8NPfeMF9.

Yannis Kalantidis, Carlos Eduardo Rosar Kos Lassance, Jon Almazán, and Diane Larlus. TLDR: Twin learning for dimensionality reduction. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=86fhqdBUbx. Expert Certification.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Seunghwan Kim and Seungkyu Lee. Beta-sigma vae: Separating beta and decoder variance in gaussian variational autoencoder. In Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa, Cheng-Lin Liu, Saumik Bhattacharya, and Umapada Pal (eds.), *Pattern Recognition*, pp. 355–369, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-78389-0.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

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 Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7e3315fe390974fcf25e44a9445bd821-Paper.pdf.

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

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer, 1999.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. 2006. URL https://api.semanticscholar.org/CorpusID:1221763.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Candi: Hybrid discrete-continuous diffusion models, 2025. URL https://arxiv.org/abs/2510.22510.

Kobi Rahimi, Yehonathan Refael, Tom Tirer, and Ofir Lindenbaum. Unveiling multiple descents in unsupervised autoencoders. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=FqfHDs6unx.

Ali Razavi, Aaron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-VAEs. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJe0Gn0cY7.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9RUHPlladgh.

Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. ControlVAE: Controllable variational autoencoder. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8655–8664. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/shao20b.html.

Ralph E. Steuer and Eng Ung Choo. An interactive weighted tchebycheff procedure for multiple objective programming. *Mathematical Programming*, 26:326–344, 1983. URL https://api.semanticscholar.org/CorpusID:27527773.

Yuhta Takida, Takashi Shibuya, Weihsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. SQ-VAE: Variational Bayes on discrete representation with self-annealed stochastic quantization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20987–21012. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/takida22a.html.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning, 2018. URL https://arxiv.org/abs/1812.05069.

Minh Hong Vu. *Disentanglement in Representation Learning: Interpretability in Dimension Reduction with VAE*. PhD thesis, Louisiana State University, 2025. URL https://repository.lsu.edu/gradschool_dissertations/6744/.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9677–9696, 2024. doi: 10.1109/TPAMI.2024.3420937.

Zihao Wang and Liu Ziyin. Posterior collapse of a linear latent variable model. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=zAc2a6_0aHb.

Yibo Yang and Stephan Mandt. Towards empirical sandwich bounds on the rate-distortion function. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=H4PmOqSZDY.

A Analytical Expansion of the β-VAE and $\lambda\beta$-VAE

A.1 Regularization Term

Proof. The regularization term in the β-VAE objective is defined as the expected KL divergence between the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y})$ and the prior $p(\mathbf{x})$. In the linear-Gaussian regime, this expectation over the data distribution $p(\mathbf{y})$ is equivalent to the KL divergence between the joint distribution and the product of marginals:

$$\mathbb{E}_{p(\mathbf{y})}\left[\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y})\|p(\mathbf{x}))\right] = \mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})),$$

where $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\phi}}(\mathbf{x}|\mathbf{y})p(\mathbf{y})$. Under the encoder mapping $\mathbf{x} = \mathbf{B}\mathbf{y} + \mathbf{w}$, the joint distribution $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})$ and the product of marginals $p(\mathbf{x})p(\mathbf{y})$ are zero-mean multivariate Gaussians:

$$q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_q), \quad p(\mathbf{x})p(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_p),$$

with the respective joint covariance matrices defined as:

$$\boldsymbol{\Sigma}_q = \begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w} & \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} \\ \boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top & \boldsymbol{\Sigma}_\mathbf{y} \end{bmatrix}, \quad \boldsymbol{\Sigma}_p = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\mathbf{y} \end{bmatrix}.$$

Thus, the KL divergence between these two distributions is given by:

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}} &= \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - (m+n) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\mathbf{I}_m||\boldsymbol{\Sigma}_\mathbf{y}|}{|\boldsymbol{\Sigma}_\mathbf{w}||\boldsymbol{\Sigma}_\mathbf{y}|} - (m+n) + \mathrm{Tr}\left(\begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\mathbf{y}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w} & \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} \\ \boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top & \boldsymbol{\Sigma}_\mathbf{y} \end{bmatrix}\right)\right] \\
&= \frac{1}{2}\left[-\log|\boldsymbol{\Sigma}_\mathbf{w}| - m - n + \mathrm{Tr}\begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w} & \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} \\ \mathbf{B}^\top & \mathbf{I}_n \end{bmatrix}\right] \\
&= \frac{1}{2}\left[-\log|\boldsymbol{\Sigma}_\mathbf{w}| - m - n + \mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w}) + n\right] \\
&= \frac{1}{2}\left[\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w}) - \log|\boldsymbol{\Sigma}_\mathbf{w}| - m\right].
\end{aligned}$$

□

A.2 Reconstruction Term

Proof. The reconstruction term $\mathcal{L}_{\mathrm{recon}}$ is defined as the expected negative log-likelihood of the decoder. We can rewrite this expectation over the joint distribution $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{y})$:

$$\mathcal{L}_{\mathrm{recon}} = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})].$$

Utilizing the conditional probability identity $\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) - \log p(\mathbf{x})$, we can decompose the reconstruction term as:

$$\mathcal{L}_{\mathrm{recon}} = \underbrace{-\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})\right]}_{\text{Joint Term}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\log p(\mathbf{x})\right]}_{\text{Prior Term}}. \tag{15}$$

Using the decoder mapping $\hat{\mathbf{y}} = \mathbf{A}\mathbf{x} + \mathbf{z}$, the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$ is given by $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, where the covariance and its inverse are:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{I}_m & \mathbf{A}^\top \\ \mathbf{A} & \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_\mathbf{z} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} = \begin{bmatrix} \mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A} & -\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1} \\ -\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_\mathbf{z}^{-1} \end{bmatrix}.$$

The joint term in (15) is expressed using the multivariate Gaussian log-likelihood as follows:

$$\begin{aligned}
&-\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})\right] \\
&= \frac{m+n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right] \\
&= \frac{m+n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^\top\right]\right) \\
&= \frac{m+n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_q\right) \\
&= \frac{m+n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_\mathbf{z}| + \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} \mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A} & -\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1} \\ -\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_\mathbf{z}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{\Sigma}_\mathbf{x} & \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} \\ \boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top & \boldsymbol{\Sigma}_\mathbf{y} \end{bmatrix}\right) \\
&= \frac{m+n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_\mathbf{z}| + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_\mathbf{x} + \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x} - \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top - \boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} + \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\right).
\end{aligned}$$

With the isotropic Gaussian prior $p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, the prior term in (15) is:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\log p(\mathbf{x})\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[-\frac{m}{2}\log(2\pi) - \frac{1}{2}\mathbf{x}^\top\mathbf{x}\right] = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}_\mathbf{x}).$$

Combining the joint and prior terms yields the final reconstruction term:

$$\mathcal{L}_{\text{recon}} = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_\mathbf{z}| + \frac{1}{2}\mathrm{Tr}\left(\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x} - \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top - \boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} + \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\right),$$

where $\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w}$. □

A.3 L_2 Reconstruction Penalty

Proof. Under the encoder mapping $\mathbf{x} = \mathbf{B}\mathbf{y} + \mathbf{w}$, the L_2 reconstruction penalty is evaluated by substituting the inference relation directly into the quadratic form:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\|\mathbf{y} - \hat{\mathbf{y}}\|^2\right]$$

$$\begin{aligned}
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\|\mathbf{y} - \mathbf{A}(\mathbf{B}\mathbf{y} + \mathbf{w})\|^2] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\|(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y} - \mathbf{A}\mathbf{w}\|^2] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y} - \mathbf{A}\mathbf{w}]^\top[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y} - \mathbf{A}\mathbf{w}]] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\mathrm{Tr}[[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y} - \mathbf{A}\mathbf{w}][(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y} - \mathbf{A}\mathbf{w}]^\top]] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\mathrm{Tr}[[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y} - \mathbf{A}\mathbf{w}][\mathbf{y}^\top(\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top - \mathbf{w}^\top\mathbf{A}^\top]]] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}[\mathrm{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y}\mathbf{y}^\top(\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top - (\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{y}\mathbf{w}^\top\mathbf{A}^\top - \mathbf{A}\mathbf{w}\mathbf{y}^\top(\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top + \mathbf{A}\mathbf{w}\mathbf{w}^\top\mathbf{A}^\top]] \\
&= \mathrm{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbb{E}[\mathbf{y}\mathbf{y}^\top](\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top - (\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbb{E}[\mathbf{y}\mathbf{w}^\top]\mathbf{A}^\top - \mathbf{A}\mathbb{E}[\mathbf{w}\mathbf{y}^\top](\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top + \mathbf{A}\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\mathbf{A}^\top].
\end{aligned}$$

The cross-covariance terms $\mathbb{E}[\mathbf{y}\mathbf{w}^\top]$ and $\mathbb{E}[\mathbf{w}\mathbf{y}^\top]$ vanish under the assumption that the observation $\mathbf{y}$ and the encoder noise $\mathbf{w}$ are independent and zero-mean, such that $\mathbb{E}[\mathbf{y}\mathbf{w}^\top] = \mathbb{E}[\mathbf{y}]\mathbb{E}[\mathbf{w}]^\top = \mathbf{0}$. Substituting the respective covariance matrices $\boldsymbol{\Sigma}_\mathbf{y}$ and $\boldsymbol{\Sigma}_\mathbf{w}$, we arrive at the final analytical form:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{y})}\left[\|\mathbf{y} - \hat{\mathbf{y}}\|^2\right] = \mathrm{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{\Sigma}_\mathbf{y}(\mathbf{I}_n - \mathbf{A}\mathbf{B})^\top + \mathbf{A}\boldsymbol{\Sigma}_\mathbf{w}\mathbf{A}^\top].$$

□

B Proof of Lemma 3.1

Proof. Following the matrix differentiation identities in Petersen & Pedersen (2006), the gradients of the objective $\mathcal{L}_\beta(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{w})$ are computed as follows:

$$\begin{aligned}
\frac{\partial\mathcal{L}_\beta}{\partial\mathbf{A}} &= -\frac{1}{2}\left[2\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top - 2\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w})\right] \\
&= \boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w}) - \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top, \\
\frac{\partial\mathcal{L}_\beta}{\partial\mathbf{B}} &= \frac{\beta}{2}(2\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}) - \frac{1}{2}(2\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y} - 2\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}) \\
&= \beta\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y} - \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y} + \mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}, \\
\frac{\partial\mathcal{L}_\beta}{\partial\boldsymbol{\Sigma}_\mathbf{w}} &= \frac{\beta}{2}(\mathbf{I}_m - \boldsymbol{\Sigma}_\mathbf{w}^{-1}) + \frac{1}{2}\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}, \\
\frac{\partial\mathcal{L}_\beta}{\partial\boldsymbol{\Sigma}_\mathbf{z}} &= \frac{1}{2}\left[\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{z}^{-1} + \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1} - \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\Sigma}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{z}^{-1}\right. \\
&\qquad \left.-\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{B}^\top + \boldsymbol{\Sigma}_\mathbf{w})\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{z}^{-1} + \boldsymbol{\Sigma}_\mathbf{z}^{-1}\right].
\end{aligned}$$

Setting $\nabla \mathcal{L}_\beta = \mathbf{0}$ yields the following first-order optimality conditions:

$$
\begin{aligned}
\mathbf{0} &= \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}}) - \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top \\
&\iff \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}}) = \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top \\
&\iff \mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})^{-1}, \qquad (17) \\
\mathbf{0} &= \beta\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} - \mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}} + \mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} \\
&\iff (\beta\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A})\mathbf{B} = \mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \\
&\iff \mathbf{B} = [\mathbf{I}_m + \mathbf{A}^\top(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}/\beta)\mathbf{A}]^{-1}\mathbf{A}^\top(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}/\beta), \qquad (18) \\
\mathbf{0} &= \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top\mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})\mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \\
&\iff \mathbf{0} = \mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top\mathbf{A}^\top - \boldsymbol{\Sigma}_{\mathbf{y}} - \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})\mathbf{A}^\top + \boldsymbol{\Sigma}_{\mathbf{z}} \\
&\iff \boldsymbol{\Sigma}_{\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{y}} - \mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top\mathbf{A}^\top + \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})\mathbf{A}^\top, \qquad (19) \\
\mathbf{0} &= \frac{\beta}{2}(\mathbf{I}_m - \boldsymbol{\Sigma}_{\mathbf{w}}^{-1}) + \frac{1}{2}\mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A} \\
&\iff \beta\boldsymbol{\Sigma}_{\mathbf{w}}^{-1} = \beta\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{A} \\
&\iff \boldsymbol{\Sigma}_{\mathbf{w}} = [\mathbf{I}_m + \mathbf{A}^\top(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}/\beta)\mathbf{A}]^{-1}. \qquad (20)
\end{aligned}
$$

Substituting (17) into (19) and utilizing the Woodbury matrix identity (Henderson & Searle, 1981), we obtain:

$$
\boldsymbol{\Sigma}_{\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} = (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})^{-1}. \qquad (21)
$$

Finally, we verify that the decoder gain $\mathbf{A}$ can be reformulated as $\mathbf{A} = (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}$. This form is established by verifying the following matrix identity:

$$
\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})^{-1} = (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}. \qquad (22)
$$

To prove (22), we pre-multiply both sides by $(\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})$ and post-multiply by $(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})$, yielding:

$$
\begin{aligned}
\text{LHS}: &\quad (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top = \mathbf{B}^\top + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top, \\
\text{RHS}: &\quad \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}}) = \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\Sigma}_{\mathbf{w}} = \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \mathbf{B}^\top.
\end{aligned}
$$

The equivalence of the transformed expressions confirms the reformulation. □

C Proof of Lemma 4.1

Proof. The stationary points of the $\lambda\beta$-VAE objective $\mathcal{L}_{\lambda\beta}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{w}})$ are found by setting the gradients with respect to each parameter to zero. The gradients are derived as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{\lambda\beta}}{\partial \mathbf{A}} &= (\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}}) - (\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + 2\lambda\mathbf{I}_n)\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top, \qquad (23a) \\
\frac{\partial \mathcal{L}_{\lambda\beta}}{\partial \mathbf{B}} &= \beta\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} + \mathbf{A}^\top(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} - \mathbf{A}^\top(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + 2\lambda\mathbf{I}_n)\boldsymbol{\Sigma}_{\mathbf{y}}, \qquad (23b) \\
\frac{\partial \mathcal{L}_{\lambda\beta}}{\partial \boldsymbol{\Sigma}_{\mathbf{w}}} &= \frac{\beta}{2}(\mathbf{I}_m - \boldsymbol{\Sigma}_{\mathbf{w}}^{-1}) + \frac{1}{2}\mathbf{A}^\top(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}, \qquad (23c) \\
\frac{\partial \mathcal{L}_{\lambda\beta}}{\partial \boldsymbol{\Sigma}_{\mathbf{z}}} &= \frac{1}{2}\left[\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{y}} - \mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top\mathbf{A}^\top + \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})\mathbf{A}^\top\right)\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\right]. \qquad (23d)
\end{aligned}
$$

Setting $\nabla \mathcal{L}_{\lambda\beta} = \mathbf{0}$ and defining the operator $\mathbf{M} \coloneqq (\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} + 2\lambda\mathbf{I}_n)/\beta$, we obtain the first-order optimality conditions. From (23a) and (23d), the decoder gain $\mathbf{A}$ and noise covariance $\boldsymbol{\Sigma}_{\mathbf{z}}$ are determined by:

$$
\begin{aligned}
\mathbf{A} &= \boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\mathbf{w}})^{-1} = (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}, \\
\boldsymbol{\Sigma}_{\mathbf{z}} &= (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})^{-1}.
\end{aligned}
$$

Notably, these conditions are structurally identical to those of the standard β-VAE. From (23b) and (23c), the encoder gain $\mathbf{B}$ and noise covariance $\mathbf{\Sigma_w}$ are coupled via the operator $\mathbf{M}$:

$$\begin{aligned}\mathbf{B} &= [\beta\mathbf{I}_m + \mathbf{A}^\top(\mathbf{\Sigma_z}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}]^{-1}\mathbf{A}^\top(\mathbf{\Sigma_z}^{-1} + 2\lambda\mathbf{I}_n) \\ &= (\mathbf{I}_m + \mathbf{A}^\top\mathbf{M}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{M}, \\ \mathbf{\Sigma_w} &= \beta[\beta\mathbf{I}_m + \mathbf{A}^\top(\mathbf{\Sigma_z}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}]^{-1} \\ &= (\mathbf{I}_m + \mathbf{A}^\top\mathbf{M}\mathbf{A})^{-1}.\end{aligned}$$

□

D Proof of Theorem 3.2

This appendix provides the formal derivation of informational collapse in the β-VAE within a linear-Gaussian regime. Throughout this analysis, $\|\cdot\|_2$ denotes the spectral norm (the largest singular value), which provides the primary analytical framework for evaluating the contraction of the encoder gain. The stationarity conditions for the β-VAE objective (Lemma 3.1) define a system of coupled equations where the optimal encoder $(\mathbf{B}, \mathbf{\Sigma_w})$ and decoder $(\mathbf{A}, \mathbf{\Sigma_z})$ are mutually dependent. We resolve this system via the alternating optimization procedure detailed in Algorithm 1, which iteratively updates each parameter block using its closed-form stationarity condition while holding others fixed.

We demonstrate that for $\beta > 1$, this iterative sequence induces a spectral contraction of the encoder gain toward the trivial fixed point $\mathbf{B} = \mathbf{0}$. This global convergence ensures that the latent signal vanishes regardless of the initial randomized state. While we monitor the optimization process using the Frobenius norm $\|\cdot\|_F$ in our numerical implementation for computational efficiency, the equivalence of norms in finite-dimensional spaces ensures that the observed numerical behavior directly implies the spectral convergence established in our theoretical analysis.

D.1 Uniform Spectral Bound

Lemma D.1 (Uniform Spectral Bound)**.** *For all iterations $t \geq 1$ of the alternating optimization procedure, the encoder noise covariance $\mathbf{\Sigma}_\mathbf{w}^{(t)}$ is spectrally bounded by unity:*

$$\|\mathbf{\Sigma}_\mathbf{w}^{(t)}\|_2 \leq 1.$$

Proof. At any iteration $t \geq 0$, the stationarity conditions established in Lemma 3.1 define the update for the subsequent state as:

$$\mathbf{\Sigma}_\mathbf{w}^{(t+1)} = \left[\mathbf{I}_m + \mathbf{A}^{(t)\top}\left(\mathbf{\Sigma}_\mathbf{z}^{(t)-1}/\beta\right)\mathbf{A}^{(t)}\right]^{-1}.$$

Given that $\mathbf{\Sigma}_\mathbf{z}^{(t)} \succ \mathbf{0}$ and $\beta > 1$, the operator $\mathbf{M} \coloneqq \mathbf{A}^{(t)\top}(\mathbf{\Sigma}_\mathbf{z}^{(t)-1}/\beta)\mathbf{A}^{(t)}$ is positive semi-definite ($\mathbf{M} \succeq \mathbf{0}$). It follows that $(\mathbf{I}_m + \mathbf{M}) \succeq \mathbf{I}_m$. Under the Loewner order, the matrix inverse operation reverses the inequality, yielding:

$$\mathbf{\Sigma}_\mathbf{w}^{(t+1)} = (\mathbf{I}_m + \mathbf{M})^{-1} \preceq \mathbf{I}_m.$$

This ordering implies that all eigenvalues $\lambda_i(\mathbf{\Sigma}_\mathbf{w}^{(t+1)})$ are contained within the interval $(0, 1]$. Consequently, the spectral norm satisfies:

$$\|\mathbf{\Sigma}_\mathbf{w}^{(t+1)}\|_2 = \lambda_{\max}(\mathbf{\Sigma}_\mathbf{w}^{(t+1)}) \leq 1.$$

While the initial state $\mathbf{\Sigma}_\mathbf{w}^{(0)}$ depends on the specific randomized initialization, the alternating update ensures that the sequence remains spectrally bounded for all $t \geq 1$. □

Algorithm 1 Alternating Optimization for β-VAE and $\lambda\beta$-VAE

Require: Data covariance $\mathbf{\Sigma_y}$, parameters β, λ, tolerance ϵ, max iterations T
Ensure: Optimized parameters $\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_z}, \mathbf{\Sigma_w}$

Initialize $\mathbf{B}, \mathbf{\Sigma_w}$ randomly
flag $\leftarrow$ True
for $t = 0$ to $T - 1$ **do**
  if *flag* **then**
    $\mathbf{\Sigma_z} \leftarrow (\mathbf{\Sigma_y}^{-1} + \mathbf{B}^\top \mathbf{\Sigma_w}^{-1} \mathbf{B})^{-1}$
    $\mathbf{A} \leftarrow \mathbf{\Sigma_z} \mathbf{B}^\top \mathbf{\Sigma_w}^{-1}$
    if $t > 0$ and $\|\mathbf{A} - \mathbf{A}_{prev}\|_F \leq \epsilon$ and $\|\mathbf{\Sigma_z} - \mathbf{\Sigma}_{\mathbf{z},prev}\|_F \leq \epsilon$ **then**
      break
    end if
    $(\mathbf{A}_{prev}, \mathbf{\Sigma}_{\mathbf{z},prev}) \leftarrow (\mathbf{A}, \mathbf{\Sigma_z})$
    flag $\leftarrow$ False
  else
    if $\lambda = 0$ **then**
      $\mathbf{M} \leftarrow \mathbf{\Sigma_z}^{-1}/\beta$ ▷ Standard β-VAE
    else
      $\mathbf{M} \leftarrow (\mathbf{\Sigma_z}^{-1} + 2\lambda \mathbf{I}_n)/\beta$ ▷ $\lambda\beta$-VAE
    end if
    $\mathbf{\Sigma_w} \leftarrow (\mathbf{I}_m + \mathbf{A}^\top \mathbf{M} \mathbf{A})^{-1}$
    $\mathbf{B} \leftarrow \mathbf{\Sigma_w} \mathbf{A}^\top \mathbf{M}$
    if $\|\mathbf{B} - \mathbf{B}_{prev}\|_F \leq \epsilon$ and $\|\mathbf{\Sigma_w} - \mathbf{\Sigma}_{\mathbf{w},prev}\|_F \leq \epsilon$ **then**
      break
    end if
    $(\mathbf{B}_{prev}, \mathbf{\Sigma}_{\mathbf{w},prev}) \leftarrow (\mathbf{B}, \mathbf{\Sigma_w})$
    flag $\leftarrow$ True
  end if
end for

D.2 Informational Collapse

Proof. Let $(\mathbf{B}^{(0)}, \mathbf{\Sigma}_{\mathbf{w}}^{(0)})$ denote the initial randomized state of the encoder. Following the stationarity conditions in Lemma 3.1, the corresponding optimal decoder parameters $(\mathbf{A}^{(0)}, \mathbf{\Sigma}_{\mathbf{z}}^{(0)})$ are given by:

$$\mathbf{A}^{(0)} = \left[\mathbf{\Sigma_y}^{-1} + \mathbf{B}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{w}}^{(0)})^{-1} \mathbf{B}^{(0)}\right]^{-1} \mathbf{B}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{w}}^{(0)})^{-1},$$
$$\mathbf{\Sigma}_{\mathbf{z}}^{(0)} = \left[\mathbf{\Sigma_y}^{-1} + \mathbf{B}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{w}}^{(0)})^{-1} \mathbf{B}^{(0)}\right]^{-1}.$$

Direct comparison of these expressions yields the fundamental coupling relation:

$$\mathbf{A}^{(0)} = \mathbf{\Sigma}_{\mathbf{z}}^{(0)} \mathbf{B}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{w}}^{(0)})^{-1}. \tag{24}$$

The subsequent update for the encoder gain $\mathbf{B}^{(1)}$ in the alternating optimization procedure is:

$$\mathbf{B}^{(1)} = \left[\mathbf{I}_m + \mathbf{A}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{z}}^{(0)-1}/\beta) \mathbf{A}^{(0)}\right]^{-1} \mathbf{A}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{z}}^{(0)-1}/\beta).$$

By identifying $\mathbf{\Sigma}_{\mathbf{w}}^{(1)} = [\mathbf{I}_m + \mathbf{A}^{(0)\top} (\mathbf{\Sigma}_{\mathbf{z}}^{(0)-1}/\beta) \mathbf{A}^{(0)}]^{-1}$ and substituting the transpose of (24) into the expression for $\mathbf{B}^{(1)}$, we derive the recursive mapping:

$$\begin{aligned}
\mathbf{B}^{(1)} &= \beta^{-1} \mathbf{\Sigma}_{\mathbf{w}}^{(1)} \mathbf{A}^{(0)\top} \mathbf{\Sigma}_{\mathbf{z}}^{(0)-1} \\
&= \beta^{-1} \mathbf{\Sigma}_{\mathbf{w}}^{(1)} \left[(\mathbf{\Sigma}_{\mathbf{w}}^{(0)})^{-1} \mathbf{B}^{(0)} \mathbf{\Sigma}_{\mathbf{z}}^{(0)}\right] \mathbf{\Sigma}_{\mathbf{z}}^{(0)-1} \\
&= \beta^{-1} \mathbf{\Sigma}_{\mathbf{w}}^{(1)} (\mathbf{\Sigma}_{\mathbf{w}}^{(0)})^{-1} \mathbf{B}^{(0)}.
\end{aligned} \tag{25}$$

By induction, after t iterations, we establish the following recursive identity for the encoder gain $\mathbf{B}^{(t)}$:

$$\mathbf{B}^{(t)} = \beta^{-t}\boldsymbol{\Sigma}_{\mathbf{w}}^{(t)}(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}. \tag{26}$$

Taking the spectral norm and utilizing the uniform bound $\|\boldsymbol{\Sigma}_{\mathbf{w}}^{(t)}\|_2 \leq 1$ for $t \geq 1$ from Lemma D.1, we obtain:

$$\|\mathbf{B}^{(t)}\|_2 \leq \beta^{-t}\|\boldsymbol{\Sigma}_{\mathbf{w}}^{(t)}\|_2 \cdot \|(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}\|_2 \leq \beta^{-t} \cdot C_0, \tag{27}$$

where $C_0 = \|(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}\|_2$ is a finite constant determined by the initialization. For $\beta > 1$, this inequality establishes that the encoder gain undergoes a spectral contraction. Consequently, the dynamics drive the encoder gain globally toward zero:

$$\|\mathbf{B}^{(t)}\|_2 \to 0 \quad \text{as } t \to \infty,$$

identifying the trivial fixed point $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{w}}) = (\mathbf{0}, \mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}}, \mathbf{I}_m)$ as the unique solution of the system.

To quantify the informational impact, we examine the mutual information $I(\mathbf{x}; \mathbf{v})$ using the joint covariance $\boldsymbol{\Sigma}$ defined in (4). Applying the Schur complement formula for determinants, the mutual information simplifies to:

$$I(\mathbf{x}; \mathbf{v}) = \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_{\mathbf{x}})}{\det(\boldsymbol{\Sigma}_{\mathbf{x}} - \mathbf{B}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{v}}\boldsymbol{\Gamma}^\top\mathbf{B}^\top)}. \tag{28}$$

As $\mathbf{B} \to \mathbf{0}$ for $\beta > 1$, the ratio inside the logarithm approaches unity, leading to $I(\mathbf{x}; \mathbf{v}) \to 0$. This demonstrates that spectral contraction enforces a complete informational collapse, leaving the latent representation asymptotically uninformative. □

E Proof of Informational Dynamics of $\lambda\beta$-VAE

Proof. Let $(\mathbf{B}^{(0)}, \boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})$ denote the initial randomized state of the encoder. Following the alternating optimization procedure, we determine the corresponding optimal decoder parameters $(\mathbf{A}^{(0)}, \boldsymbol{\Sigma}_{\mathbf{z}}^{(0)})$ using the stationarity conditions established in Lemma 3.1:

$$\mathbf{A}^{(0)} = \boldsymbol{\Sigma}_{\mathbf{z}}^{(0)}\mathbf{B}^{(0)\top}(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}.$$

The encoder is subsequently updated to state $(\mathbf{B}^{(1)}, \boldsymbol{\Sigma}_{\mathbf{w}}^{(1)})$ utilizing the augmented stationarity conditions from Lemma 4.1. Defining the operator $\mathbf{M}^{(0)} \coloneqq \beta^{-1}(\boldsymbol{\Sigma}_{\mathbf{z}}^{(0)-1} + 2\lambda\mathbf{I}_n)$, the updated encoder gain is given by:

$$\begin{aligned}
\mathbf{B}^{(1)} &= \boldsymbol{\Sigma}_{\mathbf{w}}^{(1)}\mathbf{A}^{(0)\top}\mathbf{M}^{(0)} \\
&= \beta^{-1}\boldsymbol{\Sigma}_{\mathbf{w}}^{(1)}\left[(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}\boldsymbol{\Sigma}_{\mathbf{z}}^{(0)}\right]\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{(0)-1} + 2\lambda\mathbf{I}_n\right) \\
&= \beta^{-1}\boldsymbol{\Sigma}_{\mathbf{w}}^{(1)}(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}\left(\mathbf{I}_n + 2\lambda\boldsymbol{\Sigma}_{\mathbf{z}}^{(0)}\right).
\end{aligned} \tag{29}$$

By induction, after t iterations, we establish the following recursive identity for the encoder gain:

$$\mathbf{B}^{(t)} = \beta^{-t}\boldsymbol{\Sigma}_{\mathbf{w}}^{(t)}(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}\prod_{k=0}^{t-1}\left(\mathbf{I}_n + 2\lambda\boldsymbol{\Sigma}_{\mathbf{z}}^{(k)}\right). \tag{30}$$

From Lemma D.1, we have $\|\boldsymbol{\Sigma}_{\mathbf{w}}^{(t)}\|_2 \leq 1$. From the optimality conditions, $\boldsymbol{\Sigma}_{\mathbf{z}} = (\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B})^{-1}$. Since $\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{B} \succeq \mathbf{0}$, it follows that $\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \succeq \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}$, implying $\boldsymbol{\Sigma}_{\mathbf{z}} \preceq \boldsymbol{\Sigma}_{\mathbf{y}}$ in the Loewner order. This provides the uniform spectral bound $\|\boldsymbol{\Sigma}_{\mathbf{z}}\|_2 \leq \|\boldsymbol{\Sigma}_{\mathbf{y}}\|_2$, which yields $\|\mathbf{I}_n + 2\lambda\boldsymbol{\Sigma}_{\mathbf{z}}\|_2 \leq 1 + 2\lambda\|\boldsymbol{\Sigma}_{\mathbf{y}}\|_2$.

Applying the spectral norm and the sub-multiplicative property to (30) establishes an upper bound for the iterates:

$$\|\mathbf{B}^{(t)}\|_2 \leq \left(\frac{1 + 2\lambda\|\boldsymbol{\Sigma}_{\mathbf{y}}\|_2}{\beta}\right)^t C_0, \tag{31}$$

where $C_0 = \|(\boldsymbol{\Sigma}_{\mathbf{w}}^{(0)})^{-1}\mathbf{B}^{(0)}\|_2$ is a finite constant determined by the initialization. The term $(\mathbf{I}_n + 2\lambda\boldsymbol{\Sigma}_{\mathbf{z}})$ acts as a restorative mechanism against spectral decay. When $\beta < 1 + 2\lambda\|\boldsymbol{\Sigma}_{\mathbf{y}}\|_2$, the base of the exponent

can exceed unity, preventing the informational collapse observed in the standard β-VAE and ensuring the preservation of latent informativeness.

However, because this restorative effect is fundamentally constrained by the data covariance $\|\mathbf{\Sigma}_{\mathbf{y}}\|_2$, the global contraction factor β^{-1} eventually dominates as $\beta \to \infty$. In this asymptotic regime, $\|\mathbf{B}^{(t)}\|_2 \to 0$, driving the mutual information $I(\mathbf{x}; \mathbf{v}) \to 0$. Thus, while λ significantly extends the range of informational stability, it does not preclude signal decay under arbitrarily large regularization. □

F Additional Qualitative Results

This appendix provides supplementary qualitative evaluations for the three nonlinear datasets examined in the main text: dSprites, Shapes3D, and MPI3D-real. These visualizations complement the quantitative analysis reported in Section 5 by illustrating the reconstruction fidelity and latent–factor alignment under various (β, λ) configurations. For each dataset, we present:

1. **Reconstruction Fidelity:** Original observations compared with their respective reconstructions, demonstrating the restorative effect of λ on data fidelity within high-regularization regimes.
2. **Informational Coupling:** Heatmaps of the mutual information between latent dimensions and ground-truth generative factors, illustrating how λ preserves the statistical link between the learned representation and the underlying factors.

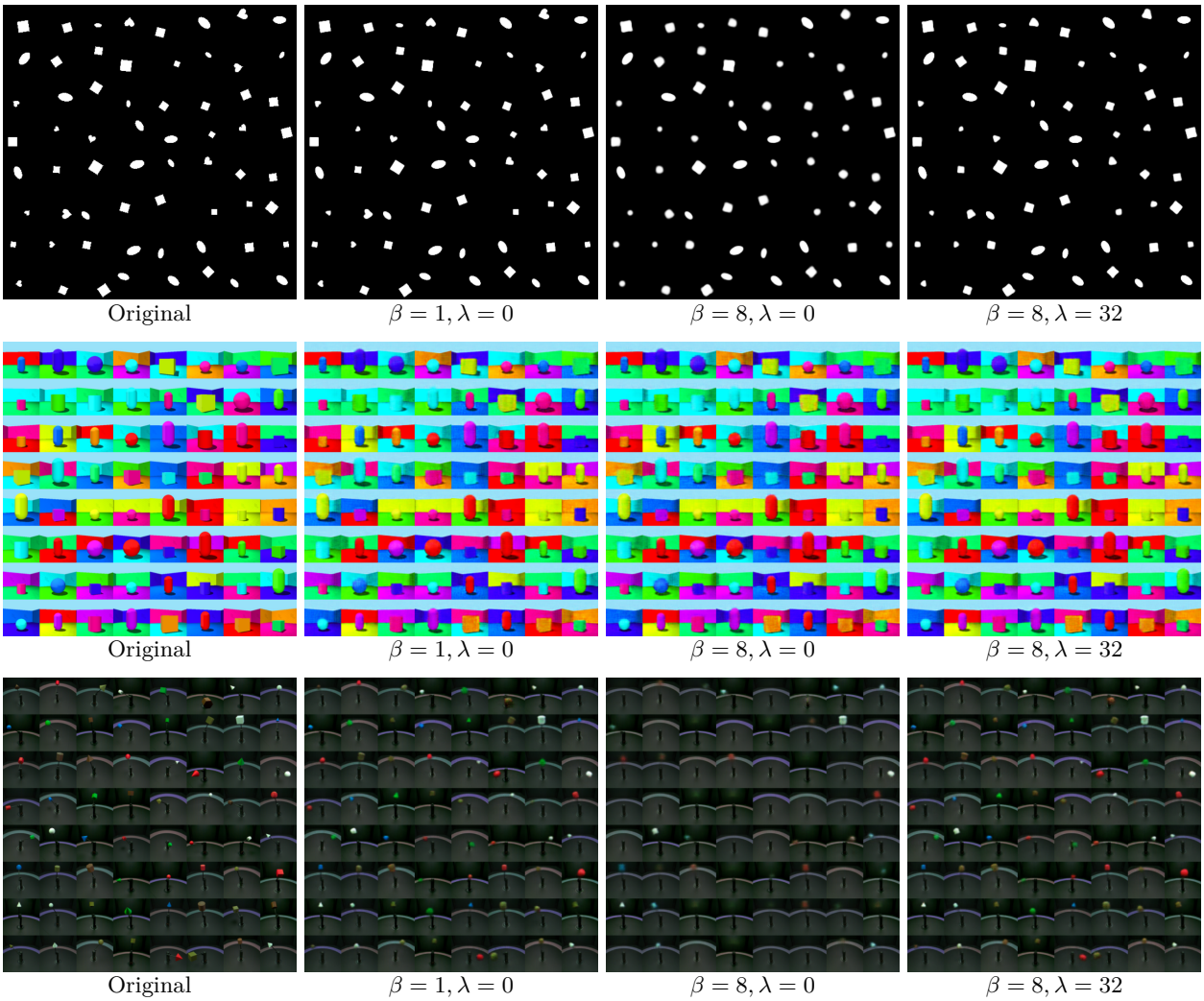
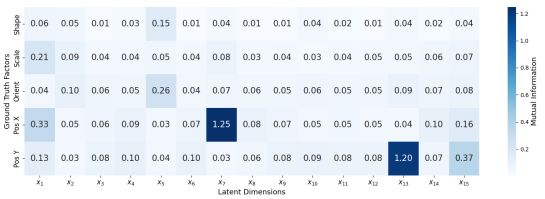


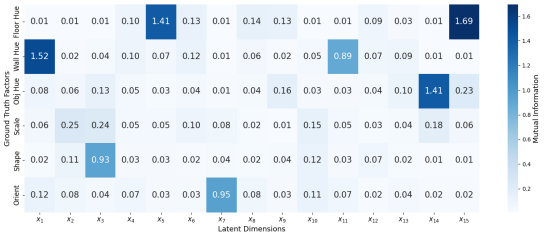
Figure 7: **Qualitative reconstructions (seed 0).** Each row shows original images alongside reconstructions for representative (β, λ) configurations. At $\lambda = 0$, increasing β leads to a visible degradation in reconstruction sharpness as the latent signal is progressively attenuated. Introducing $\lambda = 32$ restores structural integrity and high-frequency details.

dSprites: $\beta = 1, \lambda = 0$

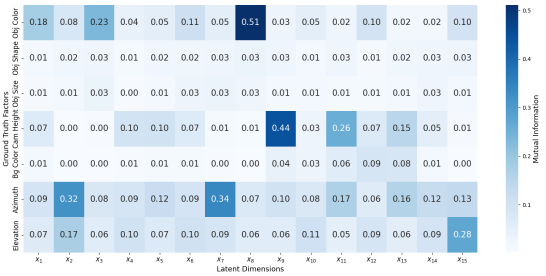


dSprites: $\beta = 8, \lambda = 32$

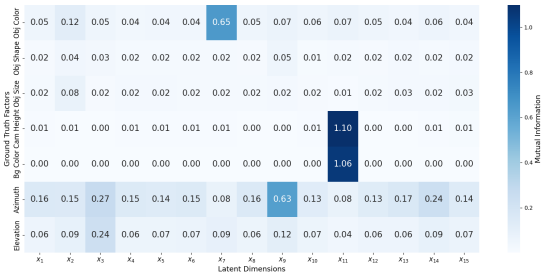
Shapes3D: $\beta = 1, \lambda = 0$



Shapes3D: $\beta = 8, \lambda = 32$



MPI3D-real: $\beta = 1, \lambda = 0$



MPI3D-real: $\beta = 8, \lambda = 32$

Figure 8: **Latent–factor mutual information heatmaps (seed 0).** Comparison between a low-regularization baseline ($\beta = 1, \lambda = 0$) and a high-regularization, information-preserving regime ($\beta = 8, \lambda = 32$). The $\lambda\beta$-VAE exhibits more localized and concentrated mutual information peaks, indicating reduced redundancy and enhanced disentanglement. This behavior demonstrates that the restorative mechanism induced by λ preserves latent informativeness, thereby enabling the factorization promoted by larger β without incurring informational collapse.