1. Overview

This supplement contains the following additional material:

- **Sec. 2:** Detailed derivations and proofs omitted from the main paper.
- **Sec. 3:** Details about datasets, our experimental setups, and the three disentanglement metrics used for quantitative performance assessment.
- **Sec. 4:** Additional experimental results.
- **Sec. 5:** Model architectures used in the experiments.

2. Derivations and Proofs

2.1. Invariance of KL Divergence under Nonlinear Invertible Transformation

**Theorem 1.** Let $x$ and $z$ be two random variables related as $z = \text{enc}(x)$ for a nonlinear invertible function $\text{enc}(\cdot)$. Then,

$$KL(p_d(x)||p(x)) = KL(q(z)||p(z)).$$

(1)

**Proof.** We let $\text{dec}(\cdot) := \text{enc}^{-1}(\cdot)$, thus $x = \text{dec}(z)$ and $z = \text{dec}^{-1}(x)$. From the theorem of density of function of random variable, we can express $p(x)$ in terms of $p(z)$ as follows:

$$p(x) = p(\text{dec}^{-1}(x)) \cdot |\det \nabla_x \text{dec}^{-1}(x)|$$

(2)

$$= p(z) \cdot \frac{1}{|\det \nabla_z \text{dec}(z)|}$$

(3)

In (2), $\nabla$ is the Jacobian operator (yielding a square invertible matrix), and (3) is from the theorem of derivative of inverse function, namely $\nabla_x \text{dec}^{-1}(x) = (\nabla_z \text{dec}(z))^{-1}$ where the latter $(\cdot)^{-1}$ is the matrix inversion. Similarly $p_d(x)$ can be written as:

$$p_d(x) = q(z) \cdot \frac{1}{|\det \nabla_z \text{dec}(z)|}$$

(4)

Now the following completes the proof:

$$KL(p_d(x)||p(x)) = \int p_d(x) \cdot \log \frac{p_d(x)}{p(x)} \, dx$$

(5)

$$= \int q(z) \cdot \frac{1}{|\det \nabla_z \text{dec}(z)|} \cdot \log \frac{q(z)}{p(z)} \, dx$$

(6)

$$= \int q(z) \cdot \log \frac{q(z)}{p(z)} \, dz$$

(7)

$$= KL(q(z)||p(z)),$$

(8)

where in the third equality, we used $dx = |\det \nabla_x \text{dec}(z)|dz$. □
2.2. Decomposition of KL-z

**Theorem 2.** Let $R$ and $N$ be the sets of relevant and nuisance variables, respectively, in the latent space $z \in Z \subseteq \mathbb{R}^d$ such that $R \cup N = \{1, \ldots, d\}$ and $R \cap N = \emptyset$. The KL divergence in the latent space $Z$ can then be decomposed as

$$KL_z := KL \left( q(z) \middle| \prod_{j=1}^{d} p(z_j) \right) = TC + \sum_{j \in R} KL(q(z_j) || p(z_j)) + \sum_{j \in N} KL(q(z_j) || p(z_j)), \quad (9)$$

where

$$TC := KL \left( q(z) || \prod_{j=1}^{d} q(z_j) \right). \quad (10)$$

**Proof.**

$$KL_z := KL \left( q(z) || \prod_{j=1}^{d} p(z_j) \right) \quad (11)$$

$$= \mathbb{E}_{q(z)} \left[ \log \frac{q(z)}{\prod_{j=1}^{d} p(z_j)} \right] = \mathbb{E}_{q(z)} \left[ \log \frac{q(z) \prod_{j=1}^{d} q(z_j)}{\prod_{j=1}^{d} p(z_j) \prod_{j'=1}^{d} q(z_{j'})} \right] \quad (12)$$

$$= KL \left( q(z) || \prod_{j=1}^{d} q(z_j) \right) + \mathbb{E}_{q(z)} \left[ \log \prod_{j=1}^{d} q(z_j) \right] \quad (13)$$

$$= TC + \sum_{j=1}^{d} \mathbb{E}_{q(z)} \left[ \log \frac{q(z_j)}{p(z_j)} \right] = TC + \sum_{j=1}^{d} KL(q(z_j) || p(z_j)) \quad (14)$$

$$= TC + \sum_{j \in R} KL(q(z_j) || p(z_j)) + \sum_{j \in N} KL(q(z_j) || p(z_j)) \quad (15)$$

\[ \square \]

2.3. Bound on Factorized Divergence $KL(q(z_j)||p(z_j))$

**Theorem 3.** Let $q(z_j)$ and $p(z_j)$ denote the approximate and prior marginal densities, respectively, of factor $j$ in latent space $Z$. Let $q(z_j|x)$ be the posterior (encoder driven) density of the latent factor conditioned on input $x$ with density $p_d(x)$. Then

$$KL(q(z_j)||p(z_j)) \leq \mathbb{E}_{p_d(x)}[KL(q(z_j|x)||p(z_j))]. \quad (16)$$

**Proof.**

$$\mathbb{E}_{p_d(x)}[KL(q(z_j|x)||p(z_j))] - KL(q(z_j)||p(z_j)) \quad (17)$$

$$= \mathbb{E}_{p_d(x)}[\mathbb{E}_{q(z_j|x)}[\log q(z_j|x) - \log p(z_j)]] - \mathbb{E}_{q(z_j)}[\mathbb{E}_{q(z_j|x)}[\log q(z_j) - \log p(z_j)]] \quad (18)$$

$$= \mathbb{E}_{p_d(x)}[\mathbb{E}_{q(z_j|x)}[\log q(z_j|x) - \log q(z_j)]] \quad (19)$$

$$= \mathbb{E}_{p_d(x)}[KL(q(z_j|x)||q(z_j))] \geq 0. \quad (20)$$

\[ \square \]

2.4. Justification of the Prior Choice in BF-VAE-0

Recall that in BF-VAE-0, we adopted a 0-mean, free-precision Gaussian as a prior model $p(z_j)$, namely $p(z_j) = \mathcal{N}(z_j; 0, \alpha_j^{-1})$, and asserted that it is as flexible as a free-mean, free-precision Gaussian $\mathcal{N}(z_j; \mu_j, \alpha_j^{-1})$, in terms of matching $p(z_j)$ and $q(z_j)$, i.e., minimizing $\mathbb{E}_{p_d(x)}[KL(q(z_j|x)||p(z_j))]$, the upper bound of $KL(q(z_j)||p(z_j))$.

We prove this assertion by showing that the following minimization problem wrt $p(z_j) = \mathcal{N}(z_j; \mu_j, \alpha_j^{-1})$ attains the optimum $\mu_j = 0$ with arbitrary $\alpha_j$ (i.e., either $\alpha_j > 1$ or $\alpha_j < 1$ or even $\alpha_j = 1$).

$$\min_{\mu_j, \alpha_j > 0} \mathbb{E}_{p_d(x)}[KL(q(z_j|x)||\mathcal{N}(z_j; \mu_j, \alpha_j^{-1}))], \quad (21)$$
where \( q(z_j|x) = \mathcal{N}(z_j; m_j(x), s_j(x)^2) \).

Note that the objective of (21) is essentially an average of KL divergences over different \( x \)'s, and we can write (21) as (dropping the subscript \( j \) for simplicity):

\[
\min_{\mu, \alpha > 0} \mathbb{E}_{p(m, s^2)}[\text{KL}(\mathcal{N}(m, s^2)||\mathcal{N}(\mu, \alpha^{-1}))],
\]

(22)

for some distribution \( p(m, s^2) \). Since the KL term admits a closed-form, (22) is equivalent to (up to constant):

\[
\min_{\mu, \alpha > 0} \mathbb{E}_{p(m, s^2)}[\alpha((m - \mu)^2 + s^2)] - \log \alpha,
\]

(23)

Letting \( \bar{\sigma}^2 = \mathbb{E}[s^2] \) leads to:

\[
\min_{\mu, \alpha > 0} \alpha \mathbb{E}_p(m) [((m - \mu)^2 + \alpha \bar{\sigma}^2) - \log \alpha].
\]

(24)

Without loss of generality, we can assume that \( p(m) \) is symmetric (i.e., \( p(m) = p(-m) \)). Then (24) can be written as (using \( \mathbb{E}[m] = 0 \)):

\[
\min_{\mu, \alpha > 0} \alpha (\bar{m}^2 + \mu^2 + \bar{\sigma}^2) - \log \alpha,
\]

(25)

where we let \( \bar{m}^2 = \mathbb{E}[m^2] \).

From (25), it is obvious that the optimal \( \mu^* = 0 \) and \( \alpha^* = (\bar{m}^2 + \bar{\sigma}^2)^{-1} \), where the latter can take arbitrary value depending on \( \bar{m} \) and \( \bar{\sigma} \). This justifies our choice of 0-mean and free-precision Gaussians as latent prior distributions.

### 2.5. Bound on Average Marginal Data Log-Likelihood in BF-VAE-1

**Theorem 4.** Let \( \{x^n\}_{n=1}^N \) be the set of iid ambient observations. Then, the negative log likelihood of these observations can be upper bounded by the BF-VAE-1 model as

\[
- \log p(\{x^n\}_{n=1}^N) \leq \text{Rec}(\theta, \nu) + \frac{1}{N} \text{KL}(q(\alpha)||p(\alpha)) + \mathbb{E}_{p(\theta|\nu)} \mathbb{E}_{q(\alpha)} \left[ \text{KL}(q(z|x)||p(z|x)) \right].
\]

(26)

**Proof.** We begin with the KL divergence between the true posterior distribution and our variational approximation.

\[
\text{KL} \left( q(\alpha) \prod_{n=1}^N q(z^n|x^n) \bigg\| p(\alpha, \{z^n\}_{n=1}^N | \{x^n\}_{n=1}^N) \right)
\]

(27)

\[
= - \mathbb{E}_{q(\alpha)} \prod_{n=1}^N q(z^n|x^n) \left[ \log \frac{q(\alpha) \prod_{n=1}^N q(z^n|x^n)}{p(\alpha, \{z^n\}_{n=1}^N | \{x^n\}_{n=1}^N)} \right]
\]

(28)

\[
= - \mathbb{E}_{q(\alpha)} \prod_{n=1}^N q(z^n|x^n) \left[ \log \frac{q(\alpha) \prod_{n=1}^N q(z^n|x^n)}{p(\alpha, \{z^n\}_{n=1}^N | \{x^n\}_{n=1}^N)} \right]
\]

(29)

\[
= \mathbb{E}_{q(\alpha)} \prod_{n=1}^N q(z^n|x^n) \left[ \log p(\{x^n\}_{n=1}^N) + \log \frac{q(\alpha)}{p(\alpha)} + \log \prod_{n=1}^N q(z^n|x^n) - \log p(x^n|z^n) \right]
\]

(30)

\[
= \log p(\{x^n\}_{n=1}^N) + \text{KL}(q(\alpha)||p(\alpha)) + \mathbb{E}_{q(\alpha)} \left[ \sum_{n=1}^N \left\{ \text{KL}(q(z^n|x^n)||p(z^n|\alpha)) - \mathbb{E}_{q(z^n|x^n)} \left[ \log p(x^n|z^n) \right] \right\} \right]
\]

(31)

Using the fact that KL divergence (27) is non-negative, we have the following inequality:

\[
- \log p(\{x^n\}_{n=1}^N) \leq \text{KL}(q(\alpha)||p(\alpha)) + \mathbb{E}_{q(\alpha)} \left[ \sum_{n=1}^N \left\{ \text{KL}(q(z^n|x^n)||p(z^n|\alpha)) - \mathbb{E}_{q(z^n|x^n)} \left[ \log p(x^n|z^n) \right] \right\} \right]
\]

(32)

Dividing both sides by \( N \) leads to our upper bound \( \mathcal{U}_1 \) in (26).
2.6. Explicit Parametric Forms of KL Divergences for BF-VAE-1 Likelihood Bound

Theorem 5. Let \( q(\alpha) \) and \( p(\alpha) \) be parameterized as Gamma densities:

\[
q(\alpha) = \prod_j G(\alpha_j; \hat{a}_j, \hat{b}_j), \tag{34}
\]

\[
p(\alpha) = \prod_j G(\alpha_j; a_j, \beta_j), \tag{35}
\]

where \( G(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by} \), and \( \Gamma(\cdot) \) is the gamma function. Then,

\[
KL(q(\alpha) || p(\alpha)) = \sum_{j=1}^d \left( \hat{a}_j - a_j \right) \psi(\hat{a}_j) + a_j \log \frac{\hat{b}_j}{b_j} + \log \frac{\Gamma(\hat{a}_j)}{\Gamma(a)} + \frac{\hat{a}_j (\beta_j - \hat{b}_j)}{b_j} \right). \tag{36}
\]

\[
\\
Proof. \text{We will use the following well-known fact of the expectations of Gamma and log-Gamma random variables. For } y \sim G(y; a, b), \\
\[
E[y] = \frac{a}{b}, \quad E[\log y] = \psi(a) - \log b,
\]

where \( \psi(\cdot) \) is the digamma function (i.e., \( \psi(y) = \frac{\Gamma'(y)}{\Gamma(y)} \)).

First we tackle \( KL(q(\alpha) || p(\alpha)) \), the second term in \( \mathcal{U}_1 \) in (26). It essentially involves in the KL divergence between two univariate Gamma distributions, which can be derived as:

\[
KL(G(y; \hat{a}, \hat{b}) || G(y; a, b)) = E_{G(y; \hat{a}, \hat{b})} \left[ \log \frac{G(y; \hat{a}, \hat{b})}{G(y; a, b)} \right] 
\]

\[
= E_{G(y; \hat{a}, \hat{b})} \left[ \hat{a} \log \hat{b} - \log \Gamma(\hat{a}) + (\hat{a} - 1) \log y - \hat{b}y \right] - \left( a \log b - \log \Gamma(a) + (a - 1) \log y - by \right) 
\]

\[
= \hat{a} \log \hat{b} - a \log b - \log \Gamma(\hat{a}) + \log \Gamma(a) + (\hat{a} - a)(\psi(\hat{a}) - \log \hat{b}) + (b - \hat{b})\frac{\hat{a}}{b} 
\]

\[
= (\hat{a} - a)\psi(\hat{a}) + a \log \frac{\hat{b}}{b} + \log \frac{\Gamma(a)}{\Gamma(\hat{a})} + \frac{\hat{a} - a}{\hat{b}} 
\]

Since \( KL(q(\alpha) || p(\alpha)) = \sum_{j=1}^d KL(G(\alpha_j; \hat{a}_j, \hat{b}_j) || G(\alpha_j; a_j, \beta_j)) \), we directly arrive at (36). \( \square \)

Theorem 6. Let \( q(\alpha) \) and \( p(\alpha) \) be parameterized as Gamma densities in (34) and (35). Then,

\[
E_{q(\alpha)} \left[ KL(q(\mathbf{z} || p(\mathbf{z})) \right] = \frac{1}{2} \sum_{j=1}^d \left( -\psi(\hat{a}_j) + \log \frac{\hat{b}_j}{s_j(\mathbf{x})^2} + \frac{\hat{a}_j (m_j(\mathbf{x})^2 + s_j(\mathbf{x})^2)}{\hat{b}_j} - 1 \right). \tag{42}
\]

\[
\\
Proof. \text{Here we derive } E_{p_d(x)} E_{q(\alpha)} \left[ KL(q(\mathbf{z} || p(\mathbf{z})) \right], \text{ the third term in } \mathcal{U}_1 \text{ in (26). The outer expectation over } p_d(x) \text{ is computed by Monte Carlo estimation with minibatches, while for a given } x, \text{ the inner expectation of the Gaussian KL over the Gamma also admits a closed form as follows:}
\]

\[
E_{q(\alpha)} \left[ KL(q(\mathbf{z} || p(\mathbf{z})) \right] = \sum_{j=1}^d E_{q(\alpha_j)} \left[ KL\left( \mathcal{N}(m_j(\mathbf{x}), s_j(\mathbf{x})^2) \ || \ \mathcal{N}(0, \alpha_j^{-1}) \right) \right] 
\]

\[
= \sum_{j=1}^d E_{q(\alpha_j)} \left[ \frac{1}{2} \alpha_j (m_j(\mathbf{x})^2 + s_j(\mathbf{x})^2) - 1 - \log s_j(\mathbf{x})^2 - \log \alpha_j \right] 
\]

\[
= \frac{1}{2} \sum_{j=1}^d \left( -\psi(\hat{a}_j) + \log \frac{\hat{b}_j}{s_j(\mathbf{x})^2} + \frac{\hat{a}_j (m_j(\mathbf{x})^2 + s_j(\mathbf{x})^2)}{\hat{b}_j} - 1 \right). \tag{45}
\]

\( \square \)
2.7. Decomposition of Approximate Marginal \( q(z) \) in BF-VAE-2

**Theorem 7.** Let \( q(z) \) be the marginal density of given by encoder \( q(z|x) \), where the ambient data is distributed according to \( p_d(x) \). Then,

\[
q(z) = q(z_R) \cdot \prod_{j \in N} q(z_j),
\]

where \( R \) and \( N \) are the sets of relevant and nuisance variables, respectively.

**Proof.** From the definition of the marginal \( q(z) \) we get:

\[
q(z) = \int q(z|x)p_d(x)dx = \int \prod_{j=1}^d q(z_j|x)p_d(x)dx
\]

\[
= \int \prod_{j \in R} q(z_j|x) \prod_{j \in N} q(z_j|x)p_d(x)dx
\]

\[
= \int \prod_{j \in R} q(z_j|x) \prod_{j \in N} q(z_j)p_d(x)dx
\]

\[
= \left( \int \prod_{j \in R} q(z_j|x)p_d(x)dx \right) \cdot \prod_{j \in N} q(z_j)
\]

\[
= q(z_R) \cdot \prod_{j \in N} q(z_j).
\]

3. Details of Datasets, Experimental Setups, and Disentanglement Metrics

3.1. Datasets

In this work, we consider the following datasets:

- **3D-Face** [7]. This dataset contains synthetic face images rendered from a 3D face model, with four factors of variation (number of modes): subject ID (50), azimuth (21), elevation (11), and lighting along the horizon (11).

- **Sprites** [6]. The dataset consists of 737,280 binary images of three shapes (oval, square, and heart), undergoing variations in four geometric factors: scale (6 variation modes), rotation (40), and \( X, Y \) translation (32 modes each), resulting in five factors total.

- **Color-Extended Sprites (C-Spr)** [5]. The dataset extends Sprites by filling the sprites with some randomly chosen colors in the images. Since the color values are not fully explored for every configuration of the factors, we regard it as simple noise, not a factor, as is also intended by the dataset creators.

- **Oval Shape Subset (O-Spr)**. From the original dataset, we select the oval shape images, yielding four independent factors of variation across 245,760 images.

- **Teapots** [1]. The dataset contains 200,000 images of a teapot across five different sources of variation: azimuth, elevation, and the color of the teapot object. Since the factor labels provided in the dataset are real and uniquely valued, to evaluate metrics I and II, we discretized them into ten bins.

- **Celeb-A** [4]. This popular dataset captures variations in 40 binary attributes including azimuth, hair color, bang, gender, mustache, brightness, etc. The is a challenging dataset because the labels (attributes) are biased\(^1\) and noisy.

\(^1\)Celeb-A attributes are not fully disentangled. For instance, the attributes male and bald (or mustache) are highly correlated, and the dataset does not fully/uniformly cover the entire joint factor space (e.g., \((\text{male} = 0, \text{bald} = 1)\) is rare).
3.2. Experimental Setup

The latent dimension $d$ for all models is set to 10 for all datasets except $d = 20$ for Celeb-A. For the trade-off parameters, we have used the optimal values reported in the publications: i) $\beta$-VAE: $\beta = 8.0$ for 3D-Face, $\beta = 4.0$ forSprites, C-Spr, and O-Spr, and $\beta = 6.0$ for Teaots, ii) Factor-VAE: $\gamma = 10.0$ for 3D-Face and Teaots, and $\gamma = 35.0$ for Sprites, C-Spr, and O-Spr. For our BF-VAE-0 and BF-VAE-2 models, the sparseness trade-off parameters ($\eta$'s) are obtained through grid search to yield the best performance. They are: i) BF-VAE-0: $\eta = 1.0$ (3D-Face), $\eta = 0.05$ (Sprites), $\eta = 0.01$ (C-Spr), $\eta = 0.05$ (O-Spr), and $\eta = 0.03$ (Teaots), ii) BF-VAE-2: $\eta_S = \eta_H = 1.0$ (3D-Face), $\eta_S = 0.05$, $\eta_H = 0.1$ (Sprites), $\eta_S = 0.01$, $\eta_H = 0.1$ (C-Spr), $\eta_S = \eta_H = 0.1$ (O-Spr and Teaots).

3.3. Disentanglement Metrics

**Metric I** [2] - One factor fixed. The goal of this metric is to assess the variability of the discovered latent factors as a function of a ground-truth factor. Let $v$ be the vector of ground-truth factors for $x$. For each factor index $j$, a set of $L$ samples $v^{(i)} = (v_j, v_{-j})$, $i = 1, \ldots, L$, is collected, corresponding to clamping of factor $j$ (no variance) and free variation of the remaining factors. The existence of latent factor(s) $z_u$ with similar, vanishing variation would then indicate the discovery of known factor $j$. In practice, this is accomplished by evaluating the encoder's outputs, $z^{(i)} \sim q(z|x^{(i)})$ for images $x^{(i)}$ corresponding to samples $v^{(i)}$. The (sample) variance $V(z_j)$ is used to find the index $u$ of the factor with the smallest variance,

$$u := \arg \min_{1 \leq j \leq d} V(z_j). \quad (52)$$

Then serves as the covariate for predicting the true factor index $j$: the metric is defined to be the accuracy of a simple classifier that predicts $j$ from $u$, among all $j = 1, \ldots, d$. If a model achieves strong disentanglement, we can expect $u = j$ (up to a permutation), making the classification easy. In particular, since $u, j \in \{1, \ldots, d\}$, the data pairs $\{(u, j)\}$ can be represented as a simple contingency table, in which a majority vote classifier is used for prediction. We use $L = 100$ samples to form each pair $(u, j)$, and collect 800 pairs to compute the accuracies of the majority vote classifiers. Since the metric is based on random samples, we repeat the evaluation ten times randomly to report the means and standard deviations.

**Metric II** - One factor varied. Following the notion of disentanglement, another reasonable approach was proposed in [3] to collect samples with *only one factor varied*, instead of one factor fixed as in Metric I. That is, we collect images with $v^{(i)} = (v_j^{(i)}, v_{-j})$ for $i = 1, \ldots, L$. (52) is then modified to $\arg \max$, and we can use the same majority vote classification to report the accuracy. The evaluation results in [3] demonstrate that this metric shows higher agreement with qualitative assessment of disentanglement than Metric I. However, note that to compute Metric II the dataset needs to contain dense joint variations in all true factors, typically a reasonable assumption for large, diverse datasets.

**Metric III** [1] proposed three metrics: 1) Disentanglement, 2) Completeness, and 3) Informativeness. These scores are regression-prediction based, using the latent vector $z$ as the covariate for individual ground-truth factors $v_j$. Specifically, $D$ measures the degree of dedication of each latent variable $z_k$ in predicting $v_j$ against others $v_{-j}$ (the higher, the better), $C$ captures the degree of exclusive contribution of $z_k$ in predicting $v_j$ against others $z_{-k}$ (the higher, the better), and $I$ measures the prediction error (the smaller, the better). For the regressors, both LASSO and Random Forests are used.

4. Additional Experimental Results

As we stated in the main paper, we place additional experimental results in this section: They are: i) Qualitative results (latent space traversal) on all datasets, and ii) Comparison with high-capacity prior models, the mixtures of Gaussians (Sec. 4.2).

4.1. Qualitative Results: Latent Space Traversal and Learned Relevance Indicators

As we did in the main paper, we depict images synthesized by traversing a single latent variable at a time and fixing the rest, while showing the accuracy of variable relevance indicator. Recall that our models have implicit/explicit indicators that point to relevant and nuisance variables. Specifically, i) BF-VAE-0 (learned $\alpha_j$): $j$ relevant if $\alpha_j$ is away from 1, while $j$ is nuisance if $\alpha_j \approx 1$, ii) BF-VAE-1 (DOF of the corrected prior $P(z_j)$, equal to $\tilde{a}_j$): $j$ is relevant if $\tilde{a}_j$ is small (distant from Gaussian), and vice versa, iii) BF-VAE-2 (learned relevance indicator variable $r_j$): $j$ is relevant if $r_j$ is large, and vice versa.

As the relevance indicators have different interpretation across the three models, we define proxy indicators $\hat{r}$ for BF-VAE-0 and BF-VAE-1, which have consistent interpretation as that of BF-VAE-2, i.e., high/low $r_j$ or $\hat{r}_j$ indicates relevance/nuisance of the variable $z_j$. Specifically, we define: $\hat{r}_j = |1 - \alpha_j^{-1}|$ for BF-VAE-0. For BF-VAE-1, we normalize the DOF ($2\tilde{a}_j$) values to $[0, 1]$-scale by a linear transformation.

The results are shown, with the discussions in the captions, for:
• 3D-Face in Fig. 1, Fig. 2, Fig. 3
• Sprites in Fig. 4, Fig. 5, Fig. 6
• C-Spr in Fig. 7, Fig. 8, Fig. 9
• O-Spr in Fig. 10, Fig. 11, Fig. 12
• Teapots in Fig. 13, Fig. 14, Fig. 15
• Celeb-A in Fig. 16, Fig. 17, Fig. 18.

Also, the enlarged version of the latent traversal images for BF-VAE-2 in the main paper is shown in Fig. 19. As detailed in the caption, adopting large $\eta$ leads only strong factors to be detected, while having small $\eta$ allows many weak factors identified.

### 4.2. Comparison with High Capacity Prior Model (Mixture of Gaussians)

The results on all labeled datasets are shown in Fig. 20. As evident, the MoG overfits as $K$ increases for all datasets, implying that overly flexible prior model can be detrimental to the disentanglement performance. Equally significantly, MoG underperforms across the entire range of $K$, compared to our BF-VAE models.

### 5. Model Architectures and Optimization

In this section, we provide the detailed specifications of the structures of different deep VAE models used in our experiments as well as the optimization strategies.

#### 5.1. Model Architectures

We adopt the model architectures and optimization parameters similar to those in [2]. The encoders consist of 5-layer conv-nets followed by two fully connected layers, and the decoders are 4-layer deconv-nets after two fully connected layers. We apply $(4 \times 4)$ filters for the convolution and the transposed convolution (deconv) in both models. For the adversarial discriminator $D$ used for optimizing the TC loss, we use a 6-layer MLP model with 1000 hidden units per layer and the leaky ReLU nonlinearity. Specifically,

- Model architecture for 3D Faces and Sprites(Sprites, C-Spr, and O-Spr) is specified in Tab. 1.
- Model architecture for Teapots and Celeb-A is listed in Tab. 2.

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INPUT 64 x 64 GREYSCALE IMAGE (3D-FACES)</strong></td>
<td><strong>INPUT $\in \mathbb{R}^{10}$</strong></td>
</tr>
<tr>
<td>4 x 4 conv. 32 ReLU, stride 2</td>
<td>FC. 256 ReLU (3D-FACES)</td>
</tr>
<tr>
<td>4 x 4 conv. 32 ReLU, stride 2</td>
<td>FC. 128 ReLU (Sprites)</td>
</tr>
<tr>
<td>4 x 4 conv. 64 ReLU, stride 2</td>
<td>4 x 4 upconv. 64 ReLU, stride 2</td>
</tr>
<tr>
<td>4 x 4 conv. 64 ReLU, stride 2</td>
<td>4 x 4 upconv. 32 ReLU, stride 2</td>
</tr>
<tr>
<td>FC. 256 ReLU. FC. 2 x 10 (3D-FACES)</td>
<td>FC. 128 ReLU. FC. 2 x 10 (Sprites)</td>
</tr>
<tr>
<td>4 x 4 upconv. 1. stride 2</td>
<td></td>
</tr>
</tbody>
</table>

#### 5.2. Optimization

The optimization parameters are chosen similarly as those in [2]. We use Adam with the batch size 64. We run $3 \times 10^5$ batch iterations with learning rates either $10^{-4}$ or $10^{-5}$. In RF-VAE [3] and our BF-VAE-2, the relevance vector $r$ is initialized as all-0.5 vector.
Figure 1. Latent space traversal in BF-VAE-0 on the 3D-Face dataset. The learned prior variances $\alpha^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The four visually evident dimensions of variability ($z_4, z_5, z_8, z_9$) are highlighted within colored boxes, where each exactly matches one of the four ground-truth factors ($z_4 =$ azimuth, $z_5 =$ lighting, $z_8 =$ elevation, and $z_9 =$ subject ID). The learned $\alpha_j$ for all these four dims are away from 1, and the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0). Note that $z_6$ is emphasized as relevant by our model, but the traversal does not seem to depict significant changes. Upon further investigation, it is revealed that this factor correlates with the subject ID, but less so than $z_9$. The chosen traversal range suppressed visual variability of $z_6$, but not $z_9$, suggesting the need to carefully determine the range bounds.
Tab. 2. Encoder and Decoder architecture for Teapots and Celeb-A dataset.

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INPUT 64 x 64 x 3 RGB IMAGE</strong></td>
<td><strong>INPUT ( \in \mathbb{R}^{10} )</strong></td>
</tr>
<tr>
<td>4 x 4 conv. 32 ReLU. stride 2</td>
<td>FC. 256 ReLU.</td>
</tr>
<tr>
<td>4 x 4 conv. 32 ReLU. stride 2</td>
<td>FC. 2 x 2 x 256 ReLU.</td>
</tr>
<tr>
<td>4 x 4 conv. 64 ReLU. stride 2</td>
<td>4 x 4 upconv. 64 ReLU. stride 2</td>
</tr>
<tr>
<td>4 x 4 conv. 64 ReLU. stride 2</td>
<td>4 x 4 upconv. 64 ReLU. stride 2</td>
</tr>
<tr>
<td>4 x 4 conv. 256 ReLU. stride 2</td>
<td>4 x 4 upconv. 32 ReLU. stride 2</td>
</tr>
<tr>
<td>FC. 256 ReLU. FC. 2 x 10.</td>
<td>4 x 4 upconv. 3. stride 2</td>
</tr>
<tr>
<td>4 x 4 upconv. 3. stride 2</td>
<td></td>
</tr>
</tbody>
</table>

References

Figure 2. Latent space traversal in BF-VAE-1 on the 3D-Face dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\mathbb{P}(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. The four recovered, highlighted, dimensions match the ground-truth factors, and their $\mathbb{P}(z_j)$’s also have relatively small DOFs, as expected, while the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0).
Figure 3. Latent space traversal in BF-VAE-2 on the 3D-Face dataset. The learned relevance vector $r$ is shown at the bottom. Again the four factors are nearly correctly identified, corresponding to the high values in the indicator variables $r_j$'s.
Figure 4. Latent space traversal in BF-VAE-0 on the *Sprites* dataset. The learned prior variances $\alpha^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The five visually evident dimensions of variability ($z_1, z_3, z_4, z_7, z_9$) are highlighted within colored boxes, where each corresponds to: $z_1 = \text{scale}$, $z_3 = \text{shape and rotation}$, $z_4 = \text{Y-pos}$, $z_7 = \text{X-pos}$, and $z_9 = \text{shape and rotation}$. The learned $\alpha_j$ for all these five dims are away from 1, as we anticipated, while $\hat{r}_j$ are high for those dimensions.
Figure 5. Latent space traversal in BF-VAE-1 on the Sprites dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\tilde{p}(z_j)$ and the (proxy) relevance indicator $\tilde{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. The five recovered, highlighted, dimensions match the ground-truth factors.
Figure 6. Latent space traversal in BF-VAE-2 on the Sprites dataset. Again the five factors are nearly correctly identified except for $j = 4$, corresponding to the high values in the indicator variables $\tau_j$'s.
Figure 7. Latent space traversal in BF-VAE-0 on the C-Spr dataset. The learned prior variances $\alpha^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The five visually evident dimensions of variability ($z_3, z_5, z_6, z_9, z_{10}$) are highlighted within colored boxes, where each exactly matches one of the five ground-truth factors ($z_3 = X$-pos, $z_5 = Y$-pos, $z_6 = scale$, $z_9 = rotation$, and $z_{10} = shape$ and rotation). The learned $\alpha_j$ for all these five dims are away from 1, while $\hat{r}_j$ are high for those dimensions.
Figure 8. Latent space traversal in BF-VAE-1 on the C-Spr dataset. The DOF ($2\hat{a}_j$) of the corrected prior $p(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. The five recovered, highlighted, dimensions match the ground-truth factors, and their $p(z_j)$’s also have relatively small DOFs and large proxy indicator values, except for $j = 5$. 
Figure 9. Latent space traversal in BF-VAE-2 on the C-Spr dataset. Again the five factors are nearly correctly identified, corresponding to the high values in the indicator variables $r_j$’s, except for $j = 6$. 
Figure 10. Latent space traversal in BF-VAE-0 on the O-SPE dataset. The learned prior variances $\alpha^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. Those five highlighted dimensions of major variability ($z_1, z_2, z_3, z_6, z_9$), match the four ground-truth factors (scale, X-, Y-pos, rotation), while the rotation is spread across $z_6$ and $z_9$. These factors also exactly correspond to the learned $\alpha_j$’s that are distant from 1, as we anticipated, while $\hat{r}_j$ are high for those dimensions.
Figure 11. Latent space traversal in BF-VAE-1 on the $\hat{O}_{-Spr}$ dataset. The DOF ($2\hat{a}_j$) of the corrected prior $p(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. Similar to BF-VAE-0, it identifies five variables with the rotation spread across $z_1$ and $z_9$. These relevant variables, as expected, have small DOFs in $p(z_j)$'s, while the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0).
Figure 12. Latent space traversal in BF-VAE-2 on the O-Spr dataset. The learned relevance vector $r$ is shown at the bottom. The learned $r$ accurately indicates the relevant dimensions.
Figure 13. Latent space traversal in BF-VAE-0 on the Teapots dataset. The learned prior variances $\alpha^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\tilde{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The five variables that explain the major variability in images, ($z_2, z_4, z_6, z_7, z_8$), do not perfectly match the true factors one by one, and two or more factors are entangled in some variables (e.g., $z_8$ explains both color R and azimuth). Note that a similar failure was also observed in [1] with complex ResNet models.
Figure 14. Latent space traversal in BF-VAE-1 on the Teapots dataset. The DOF ($2 \alpha_j$) of the corrected prior $p(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. Overall similar behaviors as BF-VAE-0, but the DOFs of the corrected prior are small for those relevant dimensions correctly identifying the dimensions of major variability, while the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0).
Figure 15. Latent space traversal in BF-VAE-2 on the Teapots dataset. The learned relevance vector \( \mathbf{r} \) is shown at the bottom. The learned \( \mathbf{r} \) accurately indicates the relevant dimensions for majority of dimensions, with the exception of \( z_6 \). Upon careful examination, this was related to the choice of the traversal range, which was significantly different for this factor from the range of other factors.
Figure 16. Latent traversal in BF-VAE-0 on Celeb-A, where the learned prior variances $\alpha^{-1}$ shown at the bottom (the value 1.0 depicted as the red dotted line). The learned $\alpha_j$ for all $j$’s are away from 1, and the model identifies many relevant variables (in fact, all dimensions affect image observation visually). Some latent dimensions visually evidently correspond to specific factors, e.g. $z_1 =$ frontal hair, $z_3 =$ sunglasses, $z_4 =$ elevation, $z_5 =$ smiling, $z_6/z_{19} =$ azimuth, $z_9 =$ gender, $z_{11} =$ hue, $z_{15} =$ hair color, $z_{17} =$ brightness, and $z_{18} =$ lighting (azimuth).
Figure 17. Latent traversal in BF-VAE-1 on Celeb-A, where the DOF ($2\hat{u}_j$) of the corrected prior $p(z_j)$ shown at the bottom. Except dimensions $z_6$ and $z_{20}$, all dimensions correspond to some meaningful factors visually, and the DOFs are mostly relatively small, and most of the proxy indicators are active. Some latent dimensions visually evidently correspond to specific factors, e.g. $z_2$ = brightness, $z_4$ = elevation, $z_7$ = lighting (azimuth), $z_8$ = gender, $z_{10}$ = hair color, $z_{11}$ = hue, $z_{12}$ = azimuth, $z_{14}$ = frontal hair, $z_{16}$ = smiling, $z_{18}$ = sunglasses.
Figure 18. Latent traversal in BF-VAE-2 on Celeb-A, where the learned relevance vector $r$ shown at the bottom. The model is able to detect many relevant variables. Some latent dimensions visually evidently correspond to specific factors, e.g. $z_2/z_4 = \text{smiling}$, $z_5/z_{15} = \text{azimuth}$, $z_8/z_{17} = \text{hue}$, $z_9 = \text{baldness}$, $z_{12}/z_{13} = \text{gender}$, $z_{14} = \text{elevation}$, $z_{16} = \text{sunglasses}$, $z_{19} = \text{frontal hair}$, and $z_{20} = \text{brightness}$.
Figure 19. Latent space traversal in BF-VAE-2 on Celeb-A. We train two BF-VAE-2 models with two different $\eta$ values ($\eta = \eta_S = \eta_H$ large and small). **(Left panel: strong factors)** contains latent traversal results with four latent variables (two subjects for each) that are detected (according to high $r_j$) by both $\eta$ small and large models. They correspond to (from top to bottom): gender, frontal hair, azimuth, and brightness, which are considered as strong/major factors. **(Right panel: weak factors)** shows traversal with four other latent variables that are detected (according to high $r_j$) only by the small $\eta$ model. They correspond to: smiling, sunglasses, elevation, and baldness, which are considered as weak/minor factors.

Figure 20. Disentanglement performance (Metric II) of F-VAE with MoG prior (Blue/Dashed) with different mixture orders ($K$) vs. BF-VAE-2 (Red/Solid). Overall, the MoG overfits as $K$ increases for all datasets.