Supplementary for Geometric Inductive Biases for Identifiable Unsupervised Learning of Disentangled Representations

1. Proofs and Analysis

In this section, we provide analysis on the impact of the latent space structure in our identifiability theory in Section 1.1, and provide proofs of α and β-identifiability in Section 1.2-1.3. We also provide analysis on the training objectives for our proposed GDRAE in Section 1.4, showing how both geometric and PCA inductive biases can be captured by the objectives.

1.1. Impact of the Latent Space Structure

We show that the latent space structure, *i.e.*, the latent space $\mathcal{Z} = \mathcal{R}^K \subset \mathbb{R}^K$ is a Cartesian products of K closed intervals $\mathcal{R} \triangleq [0, r] \subset \mathbb{R}$, plays a vital role in resisting latent rotations, and lay the basis for proving α and β-identifiability. We firstly provide the following Proposition 1 and its Corollaries 1-2.

Proposition 1. *Let $f : \mathcal{U} \to \mathbb{R}^K$ be a C^2 bijection, where $\mathcal{U} \supset \mathcal{Z}$ is an open set and $f(\mathcal{Z}) = \mathcal{Z}$. For $z \in \mathcal{Z}$, we denote $z' = f(z)$, and denote α (*resp.*, β) as the number of elements of z (*resp.*, z') that do not equal to 0 or r. Then $\alpha = \beta$.*

Proof. Without loss of generality, assume that

$$z^1, z^2, \cdots, z^\alpha \in \mathcal{R} - \{0, r\}, z^{\alpha+1}, \cdots, z^K \in \{0, r\}, \tag{1}$$

$$z'^1, z'^2, \cdots, z'^\beta \in \mathcal{R} - \{0, r\}, z'^{\beta+1}, \cdots, z'^K \in \{0, r\}. \tag{2}$$

By denoting $z_{\|\alpha} \triangleq \left(z^1, z^2, \cdots, z^\alpha\right)$, we observe that $z_{\|\alpha}$ is an interior point of $\mathcal{R}^\alpha$, therefore we can take an open set $\mathcal{O}\left(z_{\|\alpha}\right) \subset \mathcal{R}^\alpha$ as a neighbourhood of $z_{\|\alpha}$. We define

$$\mu_z(x) \triangleq \left(x^1, x^2, \cdots, x^\alpha, z^{\alpha+1}, \cdots, z^K\right), \quad x \in \mathcal{R}^\alpha, \tag{3}$$

and denote $h \triangleq f \circ \mu_z : \mathcal{O}\left(z_{\|\alpha}\right) \to h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right)$. Since f is a C^2 mapping, the derivative of h at $z_{\|\alpha}$, *i.e.*, $\left\{\frac{\partial h}{\partial z^j}\right\}_{j=1}^{\alpha}$, exists and is continuous.

We firstly prove that $\frac{\partial h^l}{\partial z^j} = 0, l = \beta+1, \cdots, K$. We observe that $z' = h\left(z_{\|\alpha}\right)$, therefore $z' \in h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right)$. By denoting

$$d = \sup\left\{s \middle| \left\{y | y \in h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right) \wedge \|y - z'\| < s\right\} \subset h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right)\right\}, \tag{4}$$

we have that for $\delta = d / \left\|\frac{\partial h}{\partial z^j}\right\|$, $\left\|z' + t\frac{\partial h}{\partial z^j} - z'\right\| = \left\|t\frac{\partial h}{\partial z^j}\right\| < d$ when $t \in (-\delta, \delta)$, namely $z' + t\frac{\partial h}{\partial z^j} \in h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right)$. We then prove by contradictory. Assume that $\exists l = \beta+1, \cdots, K$ such that $\frac{\partial h^l}{\partial z^j} \neq 0$, then $\exists t \in (-\delta, \delta)$, such that

$$\left(z' + t\frac{\partial h}{\partial z^j}\right)^l = z'^l + t\frac{\partial h^l}{\partial z^j} \notin \mathcal{R}. \tag{5}$$

This is because $z'^l \in \{0, r\}$, so we only take t such that $t\frac{\partial h^l}{\partial z^j} < 0$ in the case of $z'^l = 0$, while take t such that $t\frac{\partial h^l}{\partial z^j} > 0$ in the case of $z'^l = r$. However, since $\mu_z\left(\mathcal{O}\left(z_{\|\alpha}\right)\right) \subset \mathcal{Z}$, we have that $h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right) \subset \mathcal{Z}$, namely $z' + t\frac{\partial h}{\partial z^j} \in h\left(\mathcal{O}\left(z_{\|\alpha}\right)\right) \subset \mathcal{Z}$, which contradicts Eq. (5). So we have that $\frac{\partial h^l}{\partial z^j} = 0, l = \beta+1, \cdots, K$.

We then prove that $\alpha \leqslant \beta$. Since h is a bijection, hence the Jacobian $J \triangleq \begin{bmatrix} \frac{\partial h^1}{\partial z^1} & \cdots & \frac{\partial h^1}{\partial z^\alpha} \\ \vdots & \ddots & \vdots \\ \frac{\partial h^K}{\partial z^1} & \cdots & \frac{\partial h^K}{\partial z^\alpha} \end{bmatrix} \in \mathbb{R}^{K \times \alpha}$ at $z_{\|\alpha}$ is full rank, *i.e.*, the rank is α. Since $\frac{\partial h^l}{\partial z^j} = 0, l = \beta+1, \cdots, K$, so the rank of the matrix formed by the first β rows of J is also α, therefore we have that $\beta \geqslant \alpha$.

We finally prove that $\alpha = \beta$. By taking $g = f^{-1}$, then $g : f(\mathcal{U}) \to \mathcal{U}$ is a C^2 bijection, where $f(\mathcal{U}) \supset \mathcal{Z}$ is an open set and $g(\mathcal{Z}) = \mathcal{Z}$. Hence by reusing the above analysis for g, we have that $\alpha \geqslant \beta$. Therefore all in all, we have that $\alpha = \beta$. □

Corollary 1. *Let $f : \mathcal{U} \to \mathbb{R}^K$ be a C^2 bijection, where $\mathcal{U} \supset \mathcal{Z}$ is an open set and $f(\mathcal{Z}) = \mathcal{Z}$. By denoting $\mathcal{B} \triangleq \left\{z \in \mathcal{Z} | x^l \in \{0, r\}\right\}$ and $\mathcal{L}_x^j \triangleq \left\{\left(x^1, \cdots, x^{j-1}, t, x^{j+1}, \cdots, x^K\right) | t \in \mathcal{R} \wedge x \in \mathcal{B}\right\}$, we have that $\exists x \in \mathcal{B}$ and a permutation π of $[K]$, such that*

$$f\left(\mathcal{L}_0^j\right) = \mathcal{L}_x^{\pi_j}, \quad j \in [K]. \tag{6}$$

Proof. From Proposition 1, we have that for $0 \triangleq (0, 0, \cdots, 0) \in \mathcal{Z}$, $\exists x \in \mathcal{B}$, such that $f(0) = x$.

We firstly prove that $\forall j \in [K]$, $\exists j' \in [K]$, such that $f\left(\mathcal{L}_0^j\right) = \mathcal{L}_x^{j'}$. From Proposition 1, for $t \in \mathcal{R} - \{0, r\}$, there only exists one element of $f(t_{\wedge j})$ such that it does not equal to 0 or r. Let us denote the index of this element as $j'(t)$. Obviously, $j'(t)$ is invariant to t, because it contradicts the continuity of f that j' changes at t. It also holds that $f^{l \neq j'}(t_{\wedge j})$ is invariant to t. This is because contradiction to Proposition 1 occurs if $f^{l \neq j'}(t_{\wedge j})$ changes to $\mathcal{R} - \{0, r\}$, while contradiction to the continuity of f occurs if $f^{l \neq j'}(t_{\wedge j})$ changes to $\{0, r\}$. Therefore we have that $\exists x \in \mathcal{B}$ such that

$$f(t_{\wedge j} | t \in \mathcal{R} - \{0, r\}) = \left\{\left(x^1, \cdots, x^{j'-1}, a, x^{j'+1}, \cdots, x^K\right) | a \in \mathcal{R} - \{0, r\}\right\}. \tag{7}$$

For $t \in \{0, r\}$, we learn that $f^{l \neq j'}(t_{\wedge j}) = x^l$ due to the continuity of f, while $f^{j'}(t_{\wedge j}) \in \{0, r\}$ due to Proposition 1. Therefore all in all, we have that

$$f\left(\mathcal{L}_0^j\right) = \mathcal{L}_x^{j'}. \tag{8}$$

We then prove that there exists a permutation π of $[K]$, such that $f\left(\mathcal{L}_0^j\right) = \mathcal{L}_x^{\pi_j}$. We only need to prove that $\forall j_1 \neq j_2$, it holds that $j_1' \neq j_2'$. We prove by using contradictory. Assume that $\exists j_1 \neq j_2$ such that $j_1' = j_2' = j'$. Without loss of generality, we assume that $x^{j'} = 0$. From Proposition 1, we have that $f^{j'}(r_{\wedge j_1}) = f^{j'}(r_{\wedge j_2}) = r$, namely $f(r_{\wedge j_1}) = f(r_{\wedge j_2})$. However, $r_{\wedge j_1} \neq r_{\wedge j_2}$ contradicts to that f if a bijection. Therefore we have that $j_1' \neq j_2'$. Therefore all in all, the conclusion of Corollary 1 now follows. □

Corollary 2. *Let $f, g : \mathcal{U} \to \mathbb{R}^D$ be C^2 bijections, where $\mathcal{U} \supset \mathcal{Z}$ is an open set and $f(\mathcal{Z}) = g(\mathcal{Z})$. $\exists P, \varphi$, such that*

$$g_{(0,j)}(t) = (f \circ \varphi \circ P)_{(0,j)}(t), \quad \forall j \in [K], t \in \mathcal{R}, \tag{9}$$

where we define $f_{(x,j)}(t) \triangleq f\left(x^1, \cdots, x^{j-1}, t, x^{j+1}, \cdots, x^K\right), x \in \mathcal{B}$, and

$$\varphi(z) = \left(\varphi_1\left(z^1\right), \cdots, \varphi_K\left(z^K\right)\right), \quad z \in \mathcal{Z} \tag{10}$$

with φ_j maintaining a bijective map from $\mathcal{R}$ to $\mathcal{R}$. $P \in \mathbb{R}^{K \times K}$ is a permutation matrix.

Proof. From Corollary 1, we have that there exists a permutation π of $[K]$, such that $f^{-1} \circ g$ maintains a bijective map between $\mathcal{L}_0^j$ and $\mathcal{L}_x^{\pi_j}$. By denoting

$$\varphi_{\pi_j} \triangleq \left(f_{(x,\pi_j)}\right)^{-1} \circ g_{(0,j)}, \tag{11}$$

we have that φ_j maintains a bijective map from $\mathcal{R}$ to $\mathcal{R}$. Let $P \in \mathbb{R}^{K \times K}$ be the permutation matrix corresponding to π, then for $\forall t \in \mathcal{R}$, we have that

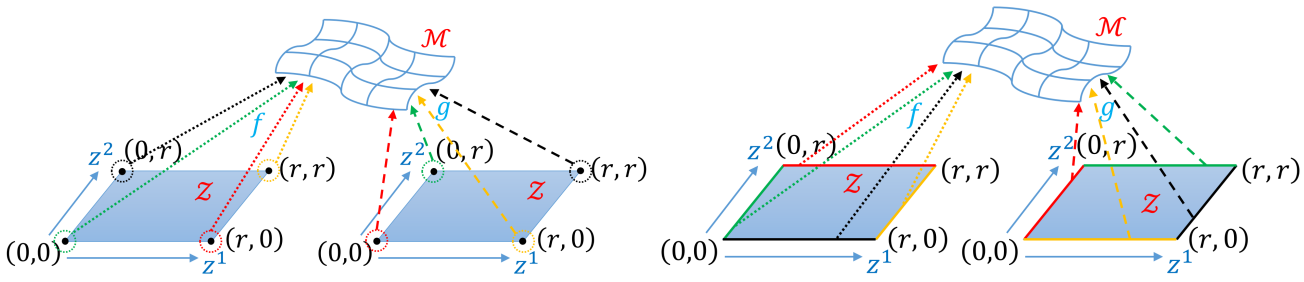
$$(f \circ \varphi \circ P)_{(0,j)}(t) = f \circ \varphi \circ P(t_{\wedge j}) = f \circ \varphi\left(t_{\wedge \pi_j}\right) \tag{12}$$

$$= f\left(\varphi_1(0), \cdots, \varphi_{\pi_j - 1}(0), \varphi_{\pi_j}(t), \varphi_{\pi_j + 1}(0), \cdots, \varphi_K(0)\right) \tag{13}$$

$$= f\left(x^1, \cdots, x^{\pi_j - 1}, \varphi_{\pi_j}(t), x^{\pi_j + 1}, \cdots, x^K\right) = f_{(x,\pi_j)}\left(\varphi_{\pi_j}(t)\right) = g_{(0,j)}(t), \tag{14}$$

therefore the conclusion of Corollary 2 now follows. □

We will show that the conclusion of Corollary 2 lays the basis for proving the α and β-identifiability as in Section 1.2-1.3. The conclusion of Corollary 2 is derived on the basis of Proposition 1 and Corollary 1, hence we provide illustrative diagrams in Fig. 1 for intuitively understanding conclusions of Proposition 1 and Corollary 1.



(a) Illustration of Proposition 1.

(b) Illustration of Corollary 1.

Figure 1. We provide illustrative diagrams for Proposition 1 and Corollary 1. We use a case where the dimensionality of latent space $\mathcal{Z}$ is 2 for intuitive observation, and consider two bijections $f, g : \mathcal{Z} \to \mathcal{M}$, where mappings f and g are drawn by using dotted lines and dashed lines, respectively. For bijection $h \triangleq g^{-1} \circ f : \mathcal{Z} \to \mathcal{Z}$, Proposition 1 essentially reads that h maps "corner points" of $\mathcal{Z}$ to "corner points" of $\mathcal{Z}$, while Corollary 1 essentially reads that h maps "edge lines" of $\mathcal{Z}$ to "edge lines" of $\mathcal{Z}$, where the mapping correspondences are indicated by colors.

1.2. Proof of α-identifiability

We now provide proof of α-identifiability as stated in Theorem 1 in main text, which we restate in the following Theorem 3. We firstly provide the following Lemma 1.

Lemma 1. *Let $f, g \in \Theta_\alpha^{\mathcal{M}}$. If $\forall j \in [K], t \in \mathcal{R}$, $\frac{\partial f}{\partial z^j}|_{f(t_{\wedge j})} = \frac{\partial g}{\partial z^j}|_{g(t_{\wedge j})}$ and $f(0) = g(0)$, then $f(z) = g(z), z \in \mathcal{Z}$.*

Proof. Since $f, g \in \Theta_\alpha^{\mathcal{M}}$, we have that

$$\left.\frac{\partial f}{\partial z^j}\right|_p = \left.\frac{\partial f}{\partial z^j}\right|_{f\left(z^j_{\wedge j}(p)\right)} = \left.\frac{\partial g}{\partial z^j}\right|_{g\left(z^j_{\wedge j}(p)\right)} = \left.\frac{\partial g}{\partial z^j}\right|_p, \quad j \in [K], z \in \mathcal{Z}, \tag{15}$$

namely the Jacobian of f and g are equal at everywhere, therefore we have that $f(z) = g(z) + c, \forall z \in \mathcal{Z}$, where c is constant. Since $f(0) = g(0)$, we have that $c = 0$, therefore $f(z) = g(z), \forall z \in \mathcal{Z}$, and the conclusion follows. □

Theorem 3 (Theorem 1 restated)**.** *Given an α-structure manifold $\mathcal{M}$, we denote $\Theta_\alpha^{\mathcal{M}} \triangleq \{f | f$ is α-related to $\mathcal{M}\}$. Then $\Theta_\alpha^{\mathcal{M}}$ is identifiable.*

Proof. From Corollary 2, we have that for $f, g \in \Theta_\alpha^{\mathcal{M}}$, $\exists P, \varphi$, such that

$$g_{(0,j)}(t) = (f \circ \varphi \circ P)_{(0,j)}(t), \quad \forall j \in [K], t \in \mathcal{R}, \tag{16}$$

where $P \in \mathbb{R}^{K \times K}$ is a permutation matrix, and $\varphi(z) = \left(\varphi_1\left(z^1\right), \cdots, \varphi_K\left(z^K\right)\right), z \in \mathcal{Z}$ with φ_j maintaining a bijective map from $\mathcal{R}$ to $\mathcal{R}$. Therefore we have that

$$\left.\frac{\partial g}{\partial z^j}\right|_{g(a_{\wedge j})} = \left.\frac{\partial g_{(0,j)}}{\partial t}\right|_{f(a)} = \left.\frac{\partial (f \circ \varphi \circ P)_{(0,j)}}{\partial t}\right|_{f(a)} = \left.\frac{\partial f \circ \varphi \circ P}{\partial z^j}\right|_{f(a_{\wedge j})}, \quad \forall j \in [K], a \in \mathcal{R}, \tag{17}$$

and $g(0) = f \circ \varphi \circ P(0)$. Therefore from Lemma 1, the conclusion of Theorem 3 follows. □

1.3. Proof of β-identifiability

We now provide proof of β-identifiability as stated in Theorem 2, which we restate in the following Theorem 4.

Theorem 4 (Theorem 2 restated)**.** *Given an β-structure manifold $\mathcal{M}$, we denote $\Theta_\beta^{\mathcal{M}} \triangleq \{f | f$ is β-related to $\mathcal{M}\}$. Then $\Theta_\beta^{\mathcal{M}}$ is identifiable.*

Proof. Given $f \in \Theta_\beta^{\mathcal{M}}$, we know that $\left(\mathcal{M}, f^{-1}\right)$ is a global coordinate chart of $\mathcal{M}$, where the local coordinate of p is given by $f^{-1}(p) = \left(z^1(p), z^2(p), \cdots, z^K(p)\right) \in \mathbb{R}^K$. Let $(\mathcal{M}, \widetilde{\varphi})$ be another global coordinate chart of $\mathcal{M}$, where the local

coordinate of $p \in \mathcal{M}$ is given by $\widetilde{\varphi}(p) = \left(\widetilde{z}^1(p), \widetilde{z}^2(p), \cdots, \widetilde{z}^K(p)\right) \in \mathbb{R}^K$. Let $\frac{\partial}{\partial z^l} = X_l^i \frac{\partial}{\partial \widetilde{z}^i} \triangleq X_l^i \partial_i$[1], we have that

$$\nabla_{\frac{\partial}{\partial z^l}} \frac{\partial}{\partial z^m} = \nabla_{X_l^i \partial_i} X_m^j \partial_j = X_l^i \left(\frac{\partial X_m^k}{\partial \widetilde{z}^i} + X_m^j \Gamma_{ji}^k \right) \partial_k \tag{18}$$

$$= \left(X_l^i \frac{\partial X_m^k}{\partial \widetilde{z}^i} + X_l^i X_m^j \Gamma_{ji}^k \right) \partial_k = \left(\frac{\partial \widetilde{z}^i}{\partial z^l} \frac{\partial X_m^k}{\partial \widetilde{z}^i} + X_l^i X_m^j \Gamma_{ji}^k \right) \partial_k \tag{19}$$

$$= \left(\frac{\partial X_m^k}{\partial z^l} + X_l^i X_m^j \Gamma_{ji}^k \right) \partial_k, \tag{20}$$

where we use $\frac{\partial \widetilde{z}^i}{\partial z^l} = \frac{\partial}{\partial z^l}\left(\widetilde{z}^i\right) = \left(X_l^i \frac{\partial}{\partial \widetilde{z}^i}\right)\left(\widetilde{z}^i\right) = X_l^i$, and Γ_{ji}^k is the coefficient of connection ∇ under $(\mathcal{M}, \widetilde{\varphi})$. By letting $\nabla_{\frac{\partial}{\partial z^l}} \frac{\partial}{\partial z^m} \equiv 0$ for $l \neq m$, we obtain the following Differential Equation (DE)

$$\frac{\partial X_m^k}{\partial z^l} + X_l^i X_m^j \Gamma_{ji}^k = 0, \quad 1 \leqslant k \leqslant K, \quad l \neq m. \tag{21}$$

Given initial condition $X_m^k(0)$, the Picard-Lindelöf theorem guarantees the existence and uniqueness of the solution. On the other hand, from Corollary 2, we have that for $f, g \in \Theta_\beta^{\mathcal{M}}$, $\exists P, \varphi$, such that

$$g_{(0,j)}(t) = (f \circ \varphi \circ P)_{(0,j)}(t), \quad \forall j \in [K], t \in \mathcal{R}, \tag{22}$$

where $P \in \mathbb{R}^{K \times K}$ is a permutation matrix, and $\varphi(z) = \left(\varphi_1\left(z^1\right), \cdots, \varphi_K\left(z^K\right)\right), z \in \mathcal{Z}$ with φ_j maintaining a bijective map from $\mathcal{R}$ to $\mathcal{R}$. Since we learn from Eq. (22) that g and $f \circ \varphi \circ P$ gives the same initial condition to Eq. (21), therefore we further have that $g = f \circ \varphi \circ P$, and the conclusion of Theorem 4 follows. □

1.4. Analysis on Model Objectives

We firstly prove Proposition 3 in main text (restated in Proposition 2) that combines constraining $J(z)$ to be orthogonal and regression of predicted singular values by constraining $\widetilde{J}(z)$ to be orthogonal.

Proposition 2 (Proposition 3 restated)**.** *For $z \in \mathcal{Z}$, $\widetilde{J}^\top(z)\widetilde{J}(z) = I \Rightarrow J(z)$ is orthogonal and $s_j(z) = \sigma_j, j \in [K]$.*

Proof. Since $\widehat{J}^\top \widetilde{J} = I$, we have that $S^\top J^\top J S = I$, where $S \triangleq \operatorname{diag}\left(\frac{1}{s_1}, \cdots, \frac{1}{s_K}\right)$, namely

$$J^\top J = \operatorname{diag}\left(s_1^2, \cdots, s_K^2\right), \tag{23}$$

which means that J is orthogonal. Let the SVD of J be $J = U \Sigma V^\top$, where $\Sigma \triangleq \operatorname{diag}(\sigma_1, \cdots, \sigma_K)$, then we have that

$$J^\top J = V \Sigma^\top \Sigma V^\top = V \operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_K^2\right) V^\top = \operatorname{diag}\left(s_1^2, \cdots, s_K^2\right). \tag{24}$$

Note that Eq. (24) is a SVD of $\operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_K^2\right)$, hence up to permutation we have that $\sigma_j^2 = s_j^2$, namely $s_j = \sigma_j$. □

We then show how our proposed model objectives can capture both geometric and PCA inductive biases as in the following Section 1.4.1-1.4.2.

1.4.1 Geometric Inductive Biases

We prove Proposition 2 in main text (restated in Proposition 3) showing that our model can capture the β-inductive biases.

Proposition 3 (Proposition 2 restated)**.** *Let $J(z) \in \mathbb{R}^{D \times K}$ be the Jacobian of the decoder g at $z \in \mathcal{Z}$, and $\{\sigma_j(z)\}_{j=1}^K$ be K singular values of $J(z)$. $\nabla_{\frac{\partial}{\partial z^i}} \frac{\partial}{\partial z^j}, \forall i \neq j \in [K]$ holds, if $\forall z \in \mathcal{Z}$, $J(z)$ is orthogonal and $\frac{\partial \sigma_j}{\partial z^i} = 0, \forall i \neq j \in [K]$.*

[1] The Einstein summation convention is used, namely $X_l^i \frac{\partial}{\partial \widetilde{z}^i} \triangleq \sum_{i=1}^K X_l^i \frac{\partial}{\partial \widetilde{z}^i}$.

Proof. It is known that the coefficient Γ_{ij}^k of the Levi-Civita connection, or the Riemannian connection ∇, is given by the *Christoffel symbol*

$$\Gamma_{ij}^k = \frac{1}{2} g^{kl} \left(\frac{\partial g_{il}}{\partial z^j} + \frac{\partial g_{jl}}{\partial z^i} - \frac{\partial g_{ij}}{\partial z^l} \right), \tag{25}$$

where g_{ij} and g^{ij} are the covariant and contravariant components of the Riemannian metric $\mathcal{G}$, respectively. The matrix form of $\mathcal{G}$ is induced by $J^\top J$, namely $\left(J^\top J\right)_{ij} = g_{ij}$, where $\left(J^\top J\right)_{ij}$ denotes the (i,j)-th element of $J^\top J$. Since J is orthogonal, we have that $g_{ij} = 0$ for $i \neq j$. Hence we further have that

$$\Gamma_{ij}^k = \frac{1}{2} g^{ki} \frac{\partial g_{ii}}{\partial z^j} + \frac{1}{2} g^{kj} \frac{\partial g_{jj}}{\partial z^i}, \quad i \neq j. \tag{26}$$

Since $\frac{\partial \sigma_j}{\partial z^i} = 0, i \neq j$, we have that $\frac{\partial g_{ii}}{\partial z^j} = 0$ and $\frac{\partial g_{jj}}{\partial z^i} = 0$ for $i \neq j$, and hence

$$\Gamma_{ij}^k = 0, \quad i \neq j. \tag{27}$$

Therefore we further have that

$$\nabla_{\frac{\partial}{\partial z^i}} \frac{\partial}{\partial z^j} = \Gamma_{ji}^k \frac{\partial}{\partial z^k} \equiv 0, \quad i \neq j, \tag{28}$$

namely the conclusion of Proposition 3 now follows. □

We proof Proposition 1 in main text (restated in Proposition 4) that gives a prototype for α-structure models.

Proposition 4 (Proposition 1 restated)**.** *$\forall f \in \Theta_\alpha^{\mathcal{M}} \Longleftrightarrow \exists p \in \mathcal{M}$, C^2 bijective mappings $\left\{ f_j : \mathcal{U} \to \mathbb{R}^D \right\}_{j=1}^K$ with $\mathcal{U} \supset \mathcal{Z}$ being an open set, such that*

$$f(z) = p + \sum_{j=1}^K f_j \left(z_{\wedge j}^j \right), \quad \forall z \in \mathcal{Z}. \tag{29}$$

Proof. We firstly prove $\Longrightarrow$. For $z \in \mathbb{R}^K$, we denote $z_{\sim j} \triangleq \left(z^1, \cdots, z^j, 0, \cdots, 0\right)$. Given $f \in \Theta_\alpha^{\mathcal{M}}$, we have that $\forall z \in \mathcal{Z}$,

$$f(z) = f(0) + \underbrace{f(z_{\sim 1}) - f(0)}_{f_1(z)} + \underbrace{f(z_{\sim 2}) - f(z_{\sim 1})}_{f_2(z)} + \cdots + \underbrace{f(z_{\sim K}) - f(z_{\sim K-1})}_{f_K(z)}, \tag{30}$$

where

$$f_j(z) = f(z_{\sim j}) - f(z_{\sim j-1}) = \int_0^{z^j} \frac{\partial f\left(z^1, \cdots, z^{j-1}, t, 0, \cdots, 0\right)}{\partial t} \mathrm{d}t. \tag{31}$$

Since $f \in \Theta_\alpha^{\mathcal{M}}$, we have that $\frac{\partial f\left(z^1, \cdots, z^{j-1}, t, 0, \cdots, 0\right)}{\partial t} = \frac{f(0, \cdots, 0, t, 0, \cdots, 0)}{\partial t} = \frac{\partial f(t_{\wedge j})}{\partial t}$, hence we further have that

$$f_j(z) = \int_0^{z^j} \frac{\partial f(t_{\wedge j})}{\partial t} \mathrm{d}t = f\left(z_{\wedge j}^j\right) - f(0). \tag{32}$$

Since f is a C^2 bijection, $f_j(z)$ is also a C^2 bijection. Therefore by combing Eq. (30) and Eq. (32), we learn that $\exists p = f(0)$ and $f_j(z) = f\left(z_{\wedge j}^j\right) - f(0)$, such that $f(z) = p + \sum_{j=1}^K f_j\left(z_{\wedge j}^j\right), \forall z \in \mathcal{Z}$.

We then prove $\Longleftarrow$. Obviously, given Eq. (29), we have that f is a C^2 bijection with

$$\frac{\partial f}{\partial z^j}\Big|_p = \frac{\partial f_j}{\partial z^j}\Big|_p = \frac{\partial f_j}{\partial z^j}\Big|_{f\left(z_{\wedge j}^j(p)\right)} = \frac{\partial f}{\partial z^j}\Big|_{f\left(z_{\wedge j}^j(p)\right)}. \tag{33}$$

Hence $f \in \Theta_\alpha^{\mathcal{M}}$, and the conclusion of Proposition 4 now follows. □

1.4.2 PCA Inductive Biases

We provide similar analysis as in [19].

Proposition 5 (Theorem 1, [19])**.** *Given N samples $X \in \mathbb{R}^{N\times D}$, consider a perfect decoder $J \in \mathbb{R}^{D\times K}$, namely the corresponding reconstructions $\widehat{X} = ZJ^\top$ are the same as X, where $Z \in \mathbb{R}^{N\times K}$ are the latent codes encoded from X. Assume that J is an orthogonal matrix with K distinct nonzero singular values $\{\sigma_j\}_{j=1}^K$, and Z is bounded, then the solution to the following objective*

$$U^*, \Sigma^*, V^* = \underset{U,\Sigma,V}{\operatorname{argmin}} \sum_{j=1}^{K} \sigma_j^2 \tag{34}$$

fulfils 1) V^* is a signed permutation matrix, and 2) $U^* = V_X$, where $J = U\Sigma V^\top$ and $X = U_X\Sigma_X V_X^\top$ are SVDs of J and X, respectively.

Proof. We borrow the proof strategy from [19]. From Proposition 1 of [15], we learn that for an orthogonal J with K distinct nonzero singular values, V is a signed permutation matrix. Without loss of generality, assume that $V = I$, where I is the identity matrix. Therefore $X = \widehat{X} = ZJ^\top = ZV\Sigma^\top U^\top = Z\Sigma^\top U^\top$, and

$$U^\top X^\top XU = U^\top U\Sigma Z^\top Z\Sigma^\top U^\top U = \Sigma Z^\top Z\Sigma^\top. \tag{35}$$

Since $X = U_X\Sigma_X V_X^\top$, by denoting $\Lambda_X \triangleq \Sigma_X^\top\Sigma_X$ and $W \triangleq U^\top V_X$ we further arrive at

$$U^\top X^\top XU = U^\top V_X\Lambda_X V_X^\top U = W\Lambda_X W^\top = \Sigma Z^\top Z\Sigma^\top. \tag{36}$$

Let the SVD of Z be $Z = U_Z\Sigma_Z V_Z^\top$, then by denoting $\Lambda_Z = \Sigma_Z^\top\Sigma_Z$, we have that $Z^\top Z = V_Z\Lambda_Z V_Z^\top$. Then from Eq. (36), we can derive that

$$\left(Z^\top Z\right)^{-1} = V_Z\Lambda_Z^{-1}V_Z^\top = \Sigma^\top W\Lambda_X^{-1}W^\top\Sigma. \tag{37}$$

Let $\cdot_{ij}$ denote the (i,j)-th element of a matrix, and $\widetilde{\Lambda}_Z \triangleq \operatorname{diag}\left(\left(Z^\top Z\right)_{11}, \cdots, \left(Z^\top Z\right)_{KK}\right)$. Since Z is bounded, the variance of each latent components (*i.e.*, $\left(\widetilde{\Lambda}_Z\right)_{jj}$) are upper bounded. Let $\widehat{\Lambda}_Z \triangleq \operatorname{diag}\left(\left(Z^\top Z\right)_{11}^{-1}, \cdots, \left(Z^\top Z\right)_{KK}^{-1}\right)$, we show that $\left(\widehat{\Lambda}_Z\right)_{jj}$ are lower bounded. Assume that $\left(\widetilde{\Lambda}_Z\right)_{jj} \leqslant \varepsilon_j$, we have that

$$\operatorname{tr}\left(Z^\top Z\right) = \operatorname{tr}\left(\Lambda_Z\right) = \sum_{j=1}^{K}\left(\Lambda_Z\right)_{jj} = \operatorname{tr}\left(\widetilde{\Lambda}_Z\right) = \sum_{j=1}^{K}\left(\widetilde{\Lambda}_Z\right)_{jj} \leqslant \sum_{j=1}^{K}\varepsilon_j \triangleq \epsilon, \tag{38}$$

therefore we have that $\left(\Lambda_Z\right)_{jj} \leqslant \epsilon$, and consequently

$$\left(\widehat{\Lambda}_Z\right)_{jj} = \sum_{k=1}^{K}\frac{\left(V_Z\right)_{jk}^2}{\left(\Lambda_Z\right)_{kk}} \geqslant \frac{1}{\epsilon}\sum_{k=1}^{K}\left(V_Z\right)_{jk}^2 = \frac{1}{\epsilon}, \tag{39}$$

which shows that $\left(\widehat{\Lambda}_Z\right)_{jj}$ are lower bounded, and hence we assume that $\left(\widehat{\Lambda}_Z\right)_{jj} \geqslant \xi_j$. From Eq. (37), we observe that

$$\left(\widehat{\Lambda}_Z\right)_{jj} = \sigma_j^2\left(W\Lambda_X^{-1}W^\top\right)_{jj}, \tag{40}$$

therefore we further have that

$$\operatorname{tr}\left(\Lambda_X^{-1}\right) = \operatorname{tr}\left(W\Lambda_X^{-1}W^\top\right) = \sum_{j=1}^{K}\frac{\left(\widehat{\Lambda}_Z\right)_{jj}}{\sigma_j^2} \tag{41}$$

Note that $\operatorname{tr}\left(\Lambda_X^{-1}\right)$ is constant. Since $\left(\widehat{\Lambda}_Z\right)_{jj} \geqslant \xi_j$ and we are free to optimize each σ_j^2 separately, therefore we have that original objective (Eq. (34)) is optimized $\Longleftrightarrow \left(\widehat{\Lambda}_Z\right)_{jj} = \xi_j$ for each j. By denoting $\Gamma_Z \triangleq \operatorname{diag}\left(\frac{\left(\widehat{\Lambda}_Z\right)_{11}}{\xi_1}, \cdots, \frac{\left(\widehat{\Lambda}_Z\right)_{KK}}{\xi_K}\right)$ and $\widetilde{\Sigma} \triangleq \operatorname{diag}\left(\frac{\sigma_1}{\sqrt{\xi_1}}, \cdots, \frac{\sigma_K}{\sqrt{\xi_K}}\right) \in \mathbb{R}^{D\times K}$, from Eq. (40) we have that

$$\left(\Gamma_Z\right)_{jj} = \widetilde{\Sigma}_{jj}^2\left(W\Lambda_X^{-1}W^\top\right)_{jj} \tag{42}$$

and consequently

$$\operatorname{tr}\left(\Gamma_Z\right) = \operatorname{tr}\left(\widetilde{\Sigma}^\top W \Lambda_X^{-1} W^\top \widetilde{\Sigma}\right). \tag{43}$$

Since $\operatorname{tr}\left(\Gamma_Z\right)$ is minimized $\Longleftrightarrow \left(\widehat{\Lambda}_Z\right)_{jj} = \xi_j \Longleftrightarrow$ Eq. (34) is optimized, we learn that the original objective (Eq. (34)) is minimized as $\operatorname{tr}\left(\widetilde{\Sigma}^\top W \Lambda_X^{-1} W^\top \widetilde{\Sigma}\right)$ is minimized. By utilizing the trace inequality (Proposition 1 of [19]), we have that

$$\operatorname{tr}\left(\widetilde{\Sigma}^\top W \Lambda_X^{-1} W^\top \widetilde{\Sigma}\right) \geqslant K \det\left(\widetilde{\Sigma}^\top W \Lambda_X^{-1} W^\top \widetilde{\Sigma}\right)^{1/K} \tag{44}$$

with the equality if and only if

$$\widetilde{\Sigma}^\top W \Lambda_X^{-1} W^\top \widetilde{\Sigma} = \lambda I \tag{45}$$

for some $\lambda \geqslant 0$. Note that this solution can be taken, since in such a case (Eq. (45)), by combining Eq. (42) we have that $\Gamma_Z = \lambda I$, which is coincident with $\Gamma_Z = I$ gives at $\left(\widehat{\Lambda}_Z\right)_{jj} = \xi_j$ when Eq. (34) is minimized. From Eq. (45) we have that

$$W \Lambda_X^{-1} W^\top = \widetilde{\Gamma}^{-\top} \widetilde{\Gamma}^{-1}, \tag{46}$$

where $\widetilde{\Gamma}^{-\top} = \operatorname{diag}\left(\frac{\sqrt{\xi_1}}{\sigma_1}, \cdots, \frac{\sqrt{\xi_K}}{\sigma_K}\right) \in \mathbb{R}^{D \times K}$. Note that the left-hand side of Eq. (46) gives an SVD decomposition of the diagonal matrix $\widetilde{\Gamma}^{-\top} \widetilde{\Gamma}^{-1}$, hence $W = U^\top V_X = I$ and $U = V_X$. The conclusion of Proposition 5 now follows. □

As analyzed by [19], VAE-based architectures exploit the property $\operatorname{diag}\left(Z^\top Z\right) = I$ induced by the KL divergence loss to capture the PCA inductive biases (see Appendix A of [19]). Different from VAE-based architectures, our framework does not involve a KL divergence loss but constrain latent codes to be distributed in a closed latent space. In such a case, we exploit the property that $\operatorname{diag}\left(Z^\top Z\right)$ are upper bounded instead.

2. Implementation Details and Supplemental Experiments

In this section, we provide supplemental experimental results on both the synthetic *Sector2D* manifold proposed in main text and natural benchmark datasets for unsupervised disentangling with ground-truth generative factors: 1) **3DFaces** [13]: 127,050 grey-scale 64×64 images of 3D faces with generative factors: *face id*[50], *azimuth*[21], *elevation*[11], *lighting*[11], 2) **3DShapes** [1]: 480,000 RGB $64 \times 64 \times 3$ images of 3D shapes with generative factors: *shape*[4], *scale*[8], *orientation*[15], *floor hue*[10], *wall hue*[10], *object hue*[10], and 3) **3DCars** [14]: 17,568 RGB $64 \times 64 \times 3$ images of 3D cars with generative factors: *elevation*[4], *azimuth*[24], *object type*[183]. As we mentioned in main text, our experiments focus on: 1) showing that the proposed geometric inductive biases can be exploited for unsupervised disentangling, and 2) showing that our proposed model is able to capture the geometric inductive biases. We firstly introduce implementation details of our experiments, then provide experimental results for the aforementioned datasets along with analysis.

Architectural Details Our experiments involve a β-VAE model and our proposed GDRAE model. We provide architecture details for both models. In terms of the latent dimensionality K, we use $K = 2$ for the *Sector2D* manifold and use $K = 10$ for the real disentangling datasets, following previous unsupervised disentangling work [2, 7]. For the *Sector2D* manifold, the sample dimensionality D is 2, and for the real datasets, the sample dimensionality D is $64 \times 64 \times c$, where $c = 1$ for grey-scale images and $c = 3$ for RGB images. For β-VAE, we follow the standard settings, using an encoder and decoder that both have plain architectures as shown in Table 1. For our GDRAE model, the decoder is designed to be *Normalizing Flow* [3, 4, 9] (NF)-based models, because NFs are intrinsic bijective mappings with explicit inverted mappings and tractable computation of Jacobian determinant, which not only satisfies the bijection prerequisites of our decoder as we claimed in our theories, but also can be exploited to implemented the $\mathcal{L}_{\text{ortho}}$ regularization in a sample-efficient manner based on the recently proposed *Linearized Transpose* [11] (LT) technique. Refer to [11] for a review on NF modules and the sample-efficient $\mathcal{L}_{\text{ortho}}$ regularization for constraining the Jacobian to be orthonormal. The architectural details for our NF-based decoders are shown in Table 2. In addition to the encoder and the decoder, our proposed model also involves a singular value predictor, which is a fully-connected architecture of "FC. 1024, ReLU, FC. 1024, ReLU, FC. K".

Sector2D		Real Datasets	
Encoder	**Decoder**	**Encoder**	**Decoder**
Input 2-dimensional vector	Input $\in \mathbb{R}^2$	Input $64 \times 64 \times c$ image	Input $\in \mathbb{R}^{10}$
FC. 1024, ReLU	FC. 1024, ReLU	$4 \times 4@(2,1)$ conv. 32, ReLU	$1 \times 1@(1,0)$ conv. 256, ReLU
FC. 1024, ReLU	FC. 1024, ReLU	$4 \times 4@(2,1)$ conv. 32, ReLU	$4 \times 4@(1,0)$ upconv. 64, ReLU
FC. 2×2	FC. 1024, ReLU	$4 \times 4@(2,1)$ conv. 64, ReLU	$4 \times 4@(2,1)$ upconv. 64, ReLU
	FC. 2	$4 \times 4@(2,1)$ conv. 64, ReLU	$4 \times 4@(2,1)$ upconv. 32, ReLU
		$4 \times 4@(1,0)$ conv. 256, ReLU	$4 \times 4@(2,1)$ upconv. 32, ReLU
		$1 \times 1@(1,0)$ conv. 2×10	$4 \times 4@(2,1)$ upconv. c

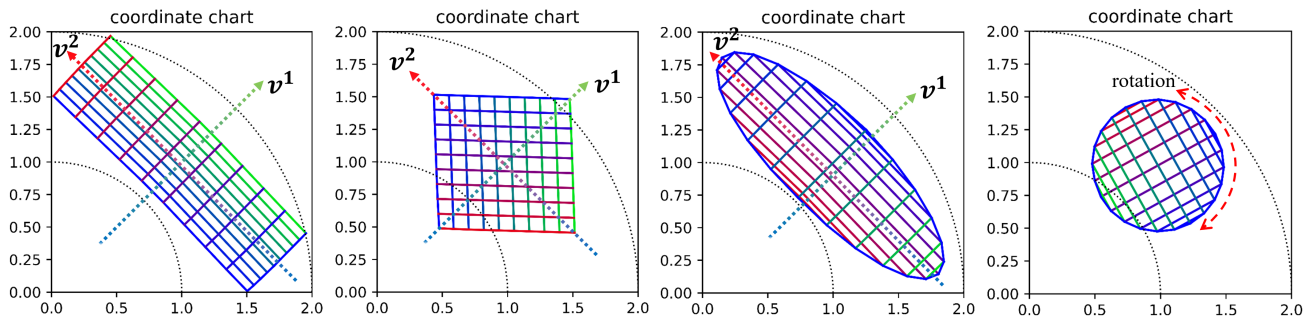
Table 1. The β-VAE model architectures for the toy *Sector2D* manifold and real datasets. We use "$k \times k@(s,p)$ conv. n" to represent a convolution layer with output channel n, kernel size k, stride s and padding p. The representation of a transposed convolution layer (denoted as upconv.) is similar. "FC. n" represents a fully-connected layer with n output channels.

Sector2D		Real Datasets	
Encoder	**Decoder**	**Encoder**	**Decoder**
Input 2-dimensional vector	Input $\in \mathbb{R}^2$	Input $64 \times 64 \times c$ image	Input $\in \mathbb{R}^{10}$
FC. 1024, ReLU	RT. 1024, CP. 1024	$4 \times 4@(2,1)$ conv. 32, ReLU	$\rightarrow (128, 4)$, RT. 128, CP. 256
FC. 1024, ReLU	RT. 1024, CP. 1024	$4 \times 4@(2,1)$ conv. 32, ReLU	$\rightarrow (64, 8)$, RT. 64, CP. 128
FC. 2	RT. 1024, CP. 1024	$4 \times 4@(2,1)$ conv. 64, ReLU	$\rightarrow (32, 16)$, RT. 32, CP. 64
		$4 \times 4@(2,1)$ conv. 64, ReLU	$\rightarrow (16, 32)$, RT. 16, CP. 32
		$4 \times 4@(1,0)$ conv. 256, ReLU	$\rightarrow (4, 64)$, RT. 4, CP. 16
		$1 \times 1@(1,0)$ conv. 10	

Table 2. The proposed model architectures for the toy *Sector2D* manifold and real datasets. For all datasets, we use "RT. n" to denote a 1×1 convolution with n output channels. For the *Sector2D* manifold (the case of $K = D \equiv 2$), the "CP. n" denotes an *additive coupling* layer with its nonlinear module being a fully-connected architecture of "FC. n, ReLU, FC. $\frac{l}{2}$", where l is the number of channels of input to the "CP. n" layer. For real datasets (the case of $10 = K < D = 64 \times 64 \times c$), the "CP. n" denotes an *additive coupling* layer with its nonlinear module being a convolutional architecture of "$4 \times 4@(2,1)$ conv. n, $4 \times 4@(2,1)$ upconv. $\frac{n}{2}$, ReLU, $1 \times 1@(1,0)$ conv. $\frac{l}{2}$", where l is the number of channels of input to the "CP. n" layer. We use $\rightarrow (n, s)$ to denote a layer that boosts the dimensionality of input features to $n \times s \times s$ by padding zeros at the end of the input features. Note that for the decoders of our proposed model, we increase the dimensionality of the intermediate features (*e.g.*, 1024 for the *Sector2D* manifold) to be higher than the dimensionality of the sample to improve the expressive power of the decoder, and obtain the final decoded sample by dropping redundant channels of the output feature.

Implementation Details We briefly introduce the principle of the Jacobian orthonormality regularization $\mathcal{L}_{\text{ortho}}$, see [11] for a detailed introduction. For the case $K = D$, the Jacobian J_g of our NF-based decoder g is a squared matrix and its determinant $\det J_g$ is tractable thanks to the favorable property of NFs. Given the LT technique that can efficiently estimate the spectral norm σ_g^* of J_g, $\mathcal{L}_{\text{ortho}}$ is implemented by constraining $\sigma_g^* = \sqrt[K]{\det J_g}$. This is because the spectral norm is essentially the maximum singular value, and $\det J_g$ is equal to the product of all singular values of J_g. Therefore, all singular values of J_g are equal and hence J_g is orthonormal, if the maximum singular value σ_g^* is equal to $\sqrt[K]{\det J_g}$. For the case $K < D$, the Jacobian J_g is no longer a squared matrix and hence $\det J_g$ becomes intractable. However, we can exploit the invertibility of NFs, obtaining the pseudo-inverse $g^\dagger$ of the decoder. Further analysis (see [11]) shows that constraining $\sigma_g^* = \frac{1}{\sigma_{g^\dagger}^*}$ leads to J_g being orthonormal, where $\sigma_{g^\dagger}^*$ is the spectral norm of $J_{g^\dagger}$.

Experimental Settings In terms of training settings, we train $100,000$ iterations with batch size 64, using an Adam optimizer [8] with fixed learning rate $1e^{-4}$. All models are train from scratch in an end-to-end manner by using PyTorch [12]. For our propose training objective (Eq. 9 in main text), we use hyper-parameters to balance different terms. Specifically, for $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{bound}}$ and $\mathcal{L}_{\text{ortho}}$, we set the corresponding hyper-parameters λ_{recon}, λ_{bound} and λ_{ortho} to be 1000, namely encouraging each objective to be strongly satisfied since they are not affected by other objectives. For the $\mathcal{L}_{\text{s\_norm}}$ term, we set $\lambda_{\text{s\_norm}}$ to be 10, and for β-VAE models, we set β to be 10. For settings of other hyper-parameters corresponding to specific experiments, we introduce along with the detailed experiments in the following Sections 2.1-2.2. All our experiments are done by using a single NVIDIA RTX-2080Ti GPU.



(a) Cubic $\mathcal{Z}$, constant σ. (b) Cubic $\mathcal{Z}$, isotropic σ. (c) Spherical $\mathcal{Z}$, constant σ. (d) Spherical $\mathcal{Z}$, isotropic σ.

Figure 2. Recovered manifold using our proposed model under different experimental settings. For better comparison, the outlines of the *Sector2D* manifold is drawn by using black dotted lines. The global linear primary and secondary directions are along v^2 and v^1.

2.1. Results on the *Sector2D* Manifold

We provide further analysis on the *Sector2D* manifold $\mathcal{M}$ proposed in main text. In main text, we show that our proposed model is capable of faithfully recovering $\mathcal{M}$ while β-VAE fails to do so. As we mentioned in main text, we know that the microscopic structure around $\forall p \in \mathcal{M}$ does not induce local PCA inductive biases, therefore for models that successfully disentangles generative factors of $\mathcal{M}$, inductive biases other than PCA should be exploited. We show in this section that the inductive biases that our model exploits is the geometric inductive biases.

Analysis on the Geometric Inductive Biases of $\mathcal{M}$ As we mentioned in main text, the *Sector2D* manifold $\mathcal{M}$ is a β-structure manifold under mild distortion. It is known that the matrix form of the Riemannian metric $\mathcal{G}$ on $\mathcal{M}$ is given as $\operatorname{diag}\left(1, \left(v^1\right)^2\right)$, which is induced by the polar transformation. Hence the Levi-Civita connection coefficient Γ_{ij}^k is given as (see Eq. (25)) $\Gamma_{11}^k = \Gamma_{22}^k = 0$ and $\Gamma_{12}^k = \Gamma_{21}^k = g^{k2}v^1$, where $\Gamma_{12}^1 = \Gamma_{21}^1 = 0$ and $\Gamma_{12}^2 = \Gamma_{21}^2 = \frac{1}{v^1}$. Hence for $\mathcal{M}$ we have that $\nabla_{\frac{\partial}{\partial v^1}} \frac{\partial}{\partial v^2} = \Gamma_{21}^k \frac{\partial}{\partial v^k} = \frac{1}{v^1}\frac{\partial}{\partial v^2}, \nabla_{\frac{\partial}{\partial v^2}} \frac{\partial}{\partial v^1} = \Gamma_{12}^k \frac{\partial}{\partial v^k} = \frac{1}{v^1}\frac{\partial}{\partial v^2}$. Therefore the distortion with respect to a β-structure manifold is induced by $\Gamma_{21}^2 = \Gamma_{12}^2 = \frac{1}{v^1}$, since $\Gamma_{ij}^k = 0$ corresponds to a β-structure manifold, as discussed in proof of Proposition 3. The distortion results in a manifold with global nonlinear structure (*i.e.*, a sector instead of a plain rectangular manifold), which allows us to provide more ablation studies on different settings to show how our proposed model exploits the geometric inductive biases to disentangle.

Ablation Study on the Singular Value Structure We provide ablation study for our proposed model on the structure of the singular values σ of the decoder Jacobian. In main text, the recovered manifold in Fig. 6(a) is obtained under a cubic latent space by using our model where the singular values σ is predicted by the singular value predictor s, *i.e.*, $(\sigma_1, \sigma_2) = s(z)$ for $z \in \mathcal{Z}$, and the structure of the latent space $\mathcal{Z}$ is cubic, *i.e.*, $\mathcal{Z} \triangleq [-1, 1]^2$. We firstly alter σ_1, σ_2 to be predicted by s depending on $z \in \mathcal{Z}$ to fixed (*i.e.*, independent of $z \in \mathcal{Z}$) but trainable σ_1, σ_2, and refer to such an ablation setting as "constant σ". The recovered manifold is shown in Fig. 2(a). We observe that the global nonlinear structure can not be captured under such a setting, but a global linear approximation to $\mathcal{M}$ is captured, which indicates that the singular value structure is a key factor in capturing the $\mathcal{M}$ structure and thus successfully disentangling the two nonlinear principal directions of $\mathcal{M}$. Since the singular value structure essentially specifies the global geometric structure of the manifold $\mathcal{M}$, the results also indicates that our proposed model indeed exploits the geometric inductive biases (*i.e.*, the singular value structure) to disentangle.

Ablation Study on the Latent Space Structure We then provide ablation study for our proposed model on the structure of the latent space $\mathcal{Z}$ by altering the cubic latent space $\mathcal{Z} \triangleq [-1, 1]^2$ to a spherical latent space $\mathcal{Z} \triangleq \left\{(x, y) \,|x^2 + y^2 \leqslant 1\right\}$. Using a spherical $\mathcal{Z}$, the recovered manifold (*i.e.*, $\widetilde{\mathcal{M}} = \{g(z) \,|z \in \mathcal{Z}\}$ where g is the decoder) under the "constant σ" is shown in Fig. 2(c). We see that similarly to Fig. 2(a), the two principal directions along v^1 and v^2 of the linear approximation to $\mathcal{M}$ can still be captured. We then alter the "constant σ" setting to the "isotropic σ" setting, where the predicted singular values σ_1, σ_2 are not only fixed (*i.e.*, independent of $z \in \mathcal{Z}$) and trainable, but also satisfy $\sigma_1 = \sigma_2$. The recovered manifold under the "isotropic σ" setting with a spherical latent space $\mathcal{Z}$ is shown in Fig. 2(d). By running experiments under such a setting multiple times, we do not observe significant alignment between latent and the manifold principal directions. Indeed,

the combination of a spherical $\mathcal{Z}$ and isotropic σ leads to an unidentifiable model, *i.e.*, the latent space after arbitrary rotation is equivalently suitable for reconstruction [15]. Comparing Fig. 2(c) and Fig. 2(d), we learn that the structure of the singular values indeed induce geometric inductive biases that can be exploited for directional alignment. We also exam how the model performs under the "isotropic σ" setting with a cubic latent space $\mathcal{Z}$, and the recovered manifold is shown in Fig. 2(b). By running experiments under such a setting multiple times, we observe that though the latent dimensions z^1, z^2 do not align with the manifold principal directions v^1 and v^2, the recovered rectangular manifold, however, always aligns its diagonal directions with v^1 and v^2 respectively, as shown in Fig. 2(b), which indicates that a cubic latent space structure indeed plays a vital role in resisting latent rotations, as we mentioned in main text.

2.2. Results on Natural Disentangling Datasets

In this section, we provide experiments results on the aforementioned real disentangling datasets. We provide both qualitative and quantitative results. In terms of quantitatively evaluating disentanglement performance, several metrics are proposed recently. To provide convincing results, we measure disentangling performance for all experiments on 8 quantitative metrics: MIG [2] (*Mutual Information Gap*), JEMMIG [5] (*Joint Entropy Minus Mutual Information Gap*), DCI [6] (*Disentanglement, Completeness and Informativeness*), DCIMIG [16], IRS [17] (*Interventional Robustness Score*), SAP [10] (*Attribute Predictability Score*), see [18] for a review of metrics for quantitatively measuring disentanglement performance. Following previous work [19, 18], we run each experiment 10 times and use the violin plots to visualize the distribution of metric scores.

2.2.1 Results of the α-Structure Model

We show more results by using the α-structure model. In main text, we construct a 3DShapes [1] subset that contains only three generative factors: *floor hue*, *wall hue* and *object hue*, and see that an α-structure model is able to disentangle the three generative factors. However, it is not clear what inductive biases the model exploits to disentangle. We now provide further ablation studies showing that the model indeed exploits the α-structure inductive biases. We also provide evidence showing that the constructed subset is indeed an α-structure dataset, which shows the applicability of the α-structure in real scenarios. In addition to the constructed subset, we also exam how the α-structure model performs on the other disentangling datasets.

Applicability of the α-Structure in Real Data We start by introducing how the 3DShapes subset is constructed. For the 3DShapes dataset with generative factors: *shape*, *scale*, *orientation*, *floor hue*, *wall hue* and *object hue*, let

$$f : S_{\text{shape}} \times S_{\text{scale}} \times S_{\text{orientation}} \times S_{\text{floor\_hue}} \times S_{\text{wall\_hue}} \times S_{\text{object\_hue}} \to \mathbb{R}^D \tag{47}$$

denote the ground-truth generative process of images, where S_{k} is the value set of the generative factor k. To construct the subset, we first randomly sampling $v^*_{\text{shape}} \in S_{\text{shape}}, v^*_{\text{scale}} \in S_{\text{scale}}, v^*_{\text{orientation}} \in S_{\text{orientation}}$, and then obtain the subset as

$$\mathcal{D} = \left\{ f\left(v^*_{\text{shape}}, v^*_{\text{scale}}, v^*_{\text{orientation}}, v_1, v_2, v_3\right) | v_1 \in S_{\text{floor\_hue}}, v_2 \in S_{\text{wall\_hue}}, v_3 \in S_{\text{object\_hue}} \right\}. \tag{48}$$

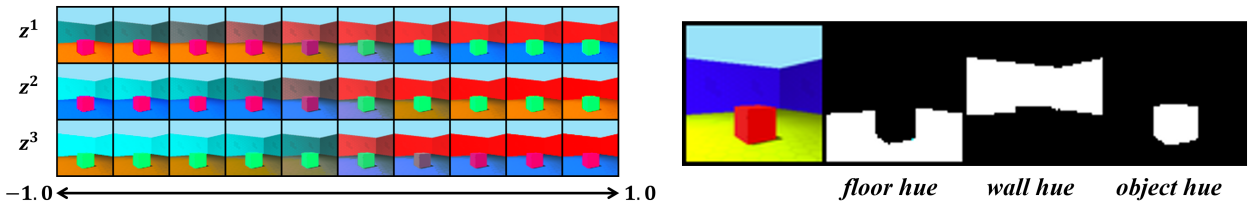
We firstly show that for $\mathcal{D} \subset \mathbb{R}^{64\times64\times3}$, different generative factors affect disjoint spatial regions of images and hence the α-structure can be applied. To do so, we firstly obtain global variances $V_{\text{floor\_hue}}, V_{\text{wall\_hue}}, V_{\text{object\_hue}} \in \mathbb{R}^{64\times64}$ with respect to spatial regions of the three generative factors as

$$V_{\text{floor\_hue}} = \mathbb{E}_{\text{RGB}} \mathbb{E}_{v \in S_{\text{floor\_hue}}} \text{Var}\left(\left\{ f^*\left(v, v_2, v_3\right) | v_2 \in S_{\text{wall\_hue}}, v_3 \in S_{\text{object\_hue}} \right\}\right) \in \mathbb{R}^{64\times64}, \tag{49}$$

where $\mathbb{E}_{\text{RGB}}$ averages RGB channels, and $f^*\left(v_1, v_2, v_3\right) \triangleq f\left(v^*_{\text{shape}}, v^*_{\text{scale}}, v^*_{\text{orientation}}, v_1, v_2, v_3\right)$. We use $V_{\text{floor\_hue}}$ as an example, and $V_{\text{wall\_hue}}, V_{\text{object\_hue}}$ can be similarly obtained. We then obtain region masks $M_k \in \{0,1\}^{64\times64}$ as

$$\left(M_{\mathrm{k}}\right)_{ij} = \mathbb{1}_{\left(V_{\mathrm{k}}\right)_{ij} > t}, \quad \mathrm{k} \in \left\{\text{floor\_hue}, \text{wall\_hue}, \text{object\_hue}\right\}, \tag{50}$$

where $\left(M_{\mathrm{k}}\right)_{ij}$ denotes the (i,j)-th element of M_{k}, and t is a threshold. Note that each element of an RGB image from $\mathcal{D}$ is an integer in $[255]$, and we set $t = 10$. The obtained region masks of global variances are shown in Fig. 3(b). Comparing masks with the leftmost images from $\mathcal{D}$, we see that different generative factors indeed affect disjoint spatial regions of images for the constructed subset $\mathcal{D}$, which shows the applicability of the α-structure in dataset $\mathcal{D}$.



(a) Latent traversals of an α'-structure model.

(b) Region masks of global variances of generative factors.

Figure 3. We show visualization of latent traversal of an α'-structure model trained on a 3DShapes subset with three generative factors: *object hue*, *wall hue* and *floor hue*. The latent traversal in Fig. 3(a) is performed in the same manner as in Fig. 7 in main text. In Fig. 3(b), we visualize region masks of global variances of generative factors, showing that different generative factors indeed affect disjoint spatial regions of images and hence the α-structure is applicable to such a dataset.

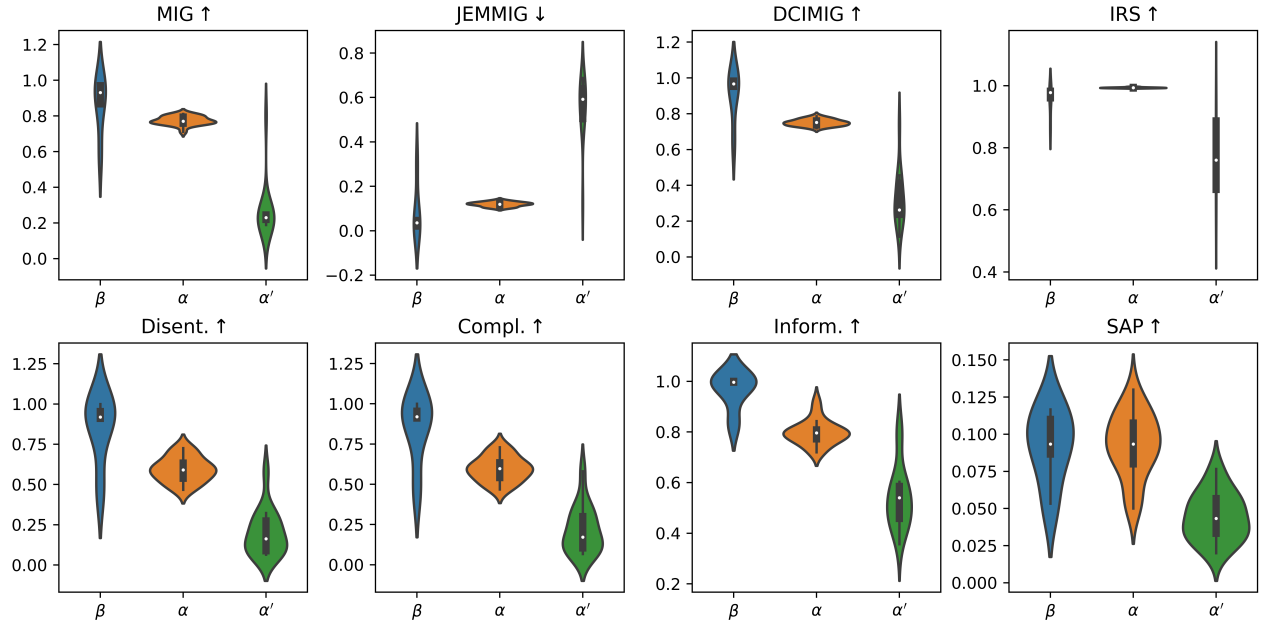


Figure 4. Quantitative results of disentangling metrics on dataset $\mathcal{D}$ achieved by β-VAE, g and g'. For β-VAE model, we use $\beta = 10$, and we use α and α' to denote g and g', respectively. For disentangling metrics, ↑ (*resp.*, ↓) means that higher (*resp.*, lower) scores are better.

Results on the constructed subset of the 3DShapes dataset We then provide ablations studies showing that for the model successfully disentangling generative factors of $\mathcal{D}$ in main text, the α-structure inductive biases is exploited. We introduce how to construct the α-structure model that obtains Fig. 7 in main text based on Proposition 4. The prototype $p \in \mathbb{R}^{64\times64\times3}$ is set to be a trainable parameter, and we set the subdecoders f_i to be $f_i\left(z^i\right) = \left(k_i z^i + b_i\right) p_i, i \in [3]$, where $k_i, b_i \in \mathbb{R}$ and $p_i \in \mathbb{R}^{64\times64\times3}$ are trainable parameters, and the α-structure decoder g is composed as $g\left(z\right) = p + \sum_{i=1}^{3} f_i\left(z^i\right)$. Note that the computation $k_i z^i + b_i$ is implemented by a trivial fully-connected layer "FC. 1". Fig. 7 in main text shows that even such a linear α-structure model g is capable of capturing the structure of $\mathcal{D}$. To show that the independence between subdecoders f_i is a key factor in capturing the α-structure, we provide a variant version of g by prepending a "FC. 3" layer $w : \mathbb{R}^3 \to \mathbb{R}^3$ that "mixes" latent dimensions to the decoder g, obtaining $g' = g \circ w$. Hence for g' the subdecoders are no longer independent with each other, and by Proposition 4 we know that g' is no longer guaranteed to be an α-structure model. We refer to g' as an α'-structure model, and the visualization of latent traversal of g' is provided in Fig. 3(a). We can observe that g' does not align latent dimensions with generative factors. In addition to qualitative comparisons, we also provides quantitative results of the aforementioned disentangling metrics as in Fig. 4. From Fig. 4, we learn that g achieves a disentangling performance that slightly lower than the performance of a β-VAE model but much higher than the performance of g'. Hence we see that the independence of subdecoders indeed plays a vital role in disentangling generative factors of $\mathcal{D}$, and by Proposition 4 we learn that g indeed exploits the α-structure inductive biases.

Results on other datasets We finally provide quantitative and qualitative results obtained by using an α-structure model on the aforementioned real disentangling datasets. Qualitative and quantitative results are provided in Fig. 17 and Fig. 5-7, respectively. In terms of quantitative comparisons, we employ an isotropic β-VAE model (*i.e.*, the encoder produces a single posterior Gaussian variance that is shared across all latent dimensions) for better comparison, since an isotropic β-VAE model does not induce local PCA inductive biases [19, 15] and we empirically found through experiments that it achieves negligible disentangling performance, see Fig. 11-13. From quantitative results in Fig. 5-7, we see that though the α-structure model g achieves slightly better disentangling performance than the performance of an isotropic β-VAE model, it achieves lower disentangling performance than a β-VAE model. We argue that this is because the α-structure model is insufficiently flexible to capture the ground-truth structure of these real datasets.

2.2.2 Results of the β-Structure Model

In this section, we provide qualitative and quantitative results on the aforementioned real disentangling datasets obtained by using our propose β-structure model, showing that the β-structure inductive biases can be exploited for disentangling and that our model is able to capture the geometric inductive biases.

Ablation Study by Removing the PCA Inductive Biases We perform ablation study using a similar strategy as in Section 2.1. We provide comparison of quantitative results of disentangling performances on real datasets between our model and a β-VAE model under two ablation settings: the original setting and the setting where the PCA inductive biases of models are removed. For β-VAE, we remove the PCA inductive biases of the model by using an isotropic β-VAE model (see Section 2.2.1), since the PCA inductive biases can not be induced given that latent channel capacities (*i.e.*, posterior Gaussian variances) are the same [19, 15]. For our model, we remove the PCA inductive biases by choosing the "constant σ" setting as introduced in Section 2.2.1, since Proposition 5 as well as [19, 15] show that distinct singular values of the decoder Jacobian is necessary for capturing the PCA inductive biases. The quantitative comparisons are provided in Fig. 8-13. We see that while under the original setting, our model achieves similar disentangling performance to that of the β-VAE model, the performances under the setting removing the PCA inductive biases are quite different. Specifically, compared to the original models, while the isotropic β-VAE model consistently achieves low performance (see Fig. 11-13), our model is able to maintain a relatively high performance (see Fig. 8-10) compared to isotropic β-VAE. This indicates that the PCA inductive biases is the only inductive biases that β-VAE exploits to disentangle. We are also aware of that for our variant model without PCA inductive biases, though the disentangling performance is maintained at a higher level than the performance of isotropic β-VAE, the reduction of the performance can be significantly observed compared to our original model. We argue that the reduction of the performance is because the model under the "constant σ" setting is not as flexible as the original model, which implies that the singular value structure indeed affects the disentangling performance. However, the "constant σ" allows the model to recover a trivial approximation to the data manifold (similar to the linear approximation to the *Sector2D* manifold in Fig. 2(a)), which may explain why the model under the "constant σ" setting still achieves a disentangling performance that is higher than the isotropic β-VAE but lower than the original model. The above analysis implies that our model captures the global geometric inductive biases to disentangle. In terms of qualitative results, we provide visualization of latent traversals of our original model in the same manner as we done in main text, see Fig. 17.

Ablation Study on the Latent Space Structure Similar to ablation studies in Section 2.1, we also alter the cubic latent space $\mathcal{Z} \triangleq [-1, 1]^{10}$ to a spherical latent space $\mathcal{Z} \triangleq \left\{z \in \mathbb{R}^{10} | \left\|z\right\|_2 \leqslant 1\right\}$ to explore how the latent space structure affects the disentangling performance on real datasets. The results are shown in Fig. 14-16. We observe that compared to a cubic latent space, the model with a spherical latent space consistently achieves much lower disentangling performances, which is coincident with our observation on the *Sector2D* manifold as in Section 2.1 and indicates that a cubic latent space structure also plays a vital role in achieving high disentangling performances on real datasets.

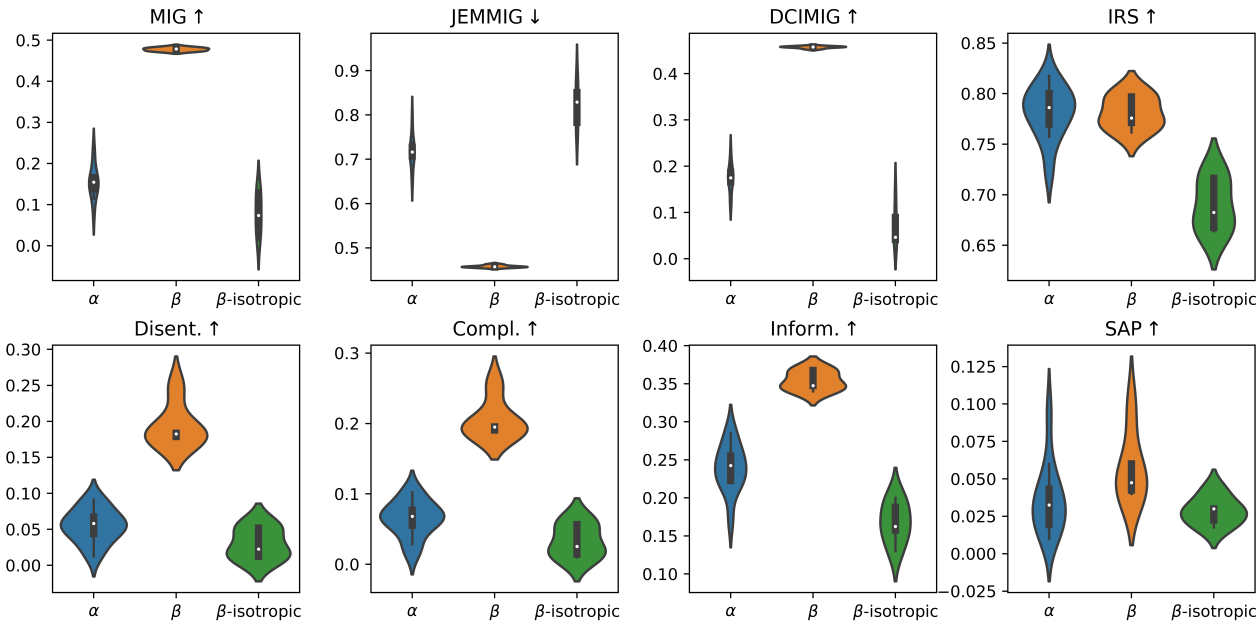


Figure 5. Quantitative results of disentangling metrics on 3DFaces achieved by g (denoted as α), a β-VAE model (denoted as β), and a isotropic β-VAE model (denoted as β-isotropic). The isotropic β-VAE model is served as a baseline method for better comparison.

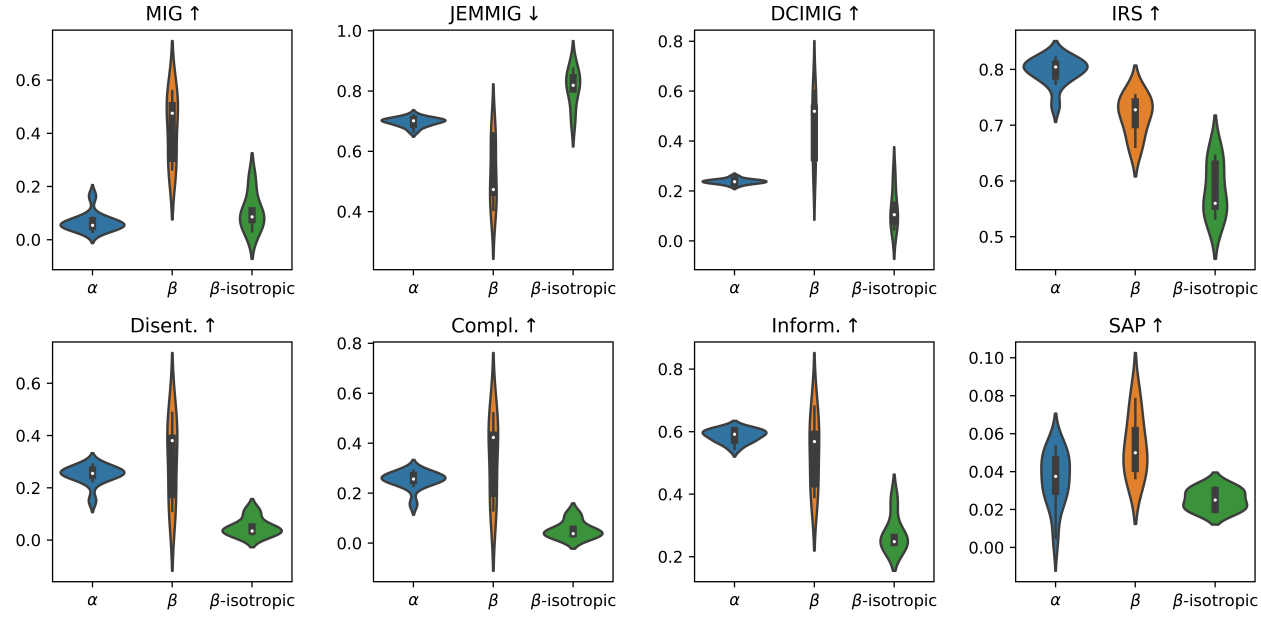


Figure 6. Quantitative results of disentangling metrics on 3DShapes achieved by g (denoted as α), a β-VAE model (denoted as β), and a isotropic β-VAE model (denoted as β-isotropic). The isotropic β-VAE model is served as a baseline method for better comparison.

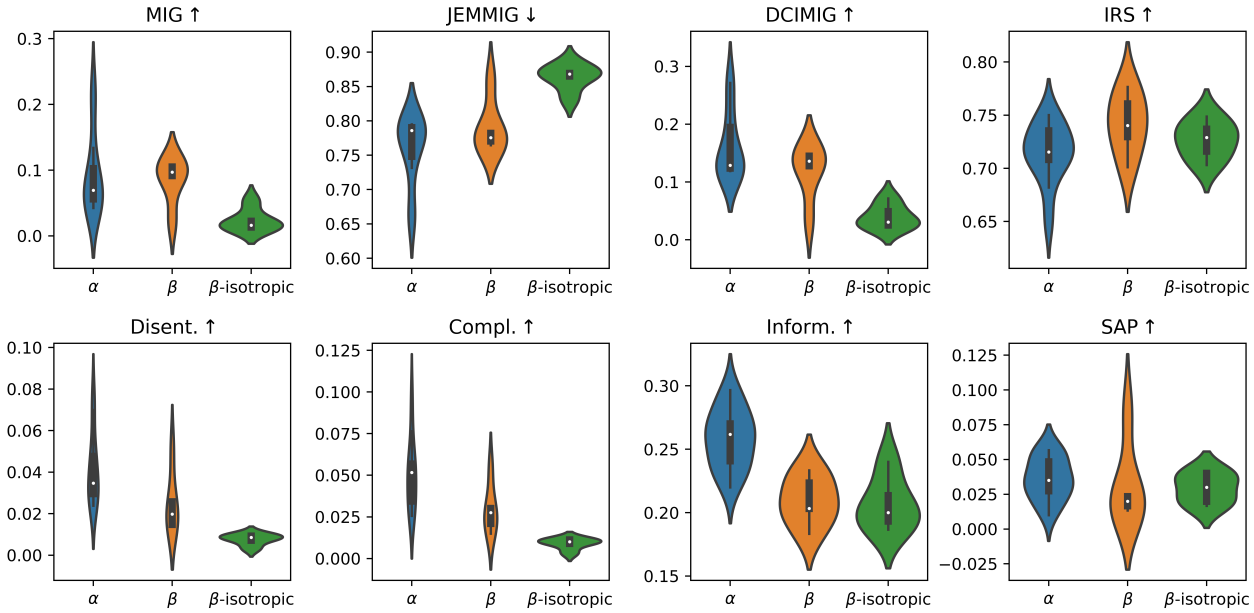


Figure 7. Quantitative results of disentangling metrics on 3DCars achieved by g (denoted as α), a β-VAE model (denoted as β), and a isotropic β-VAE model (denoted as β-isotropic). The isotropic β-VAE model is served as a baseline method for better comparison.

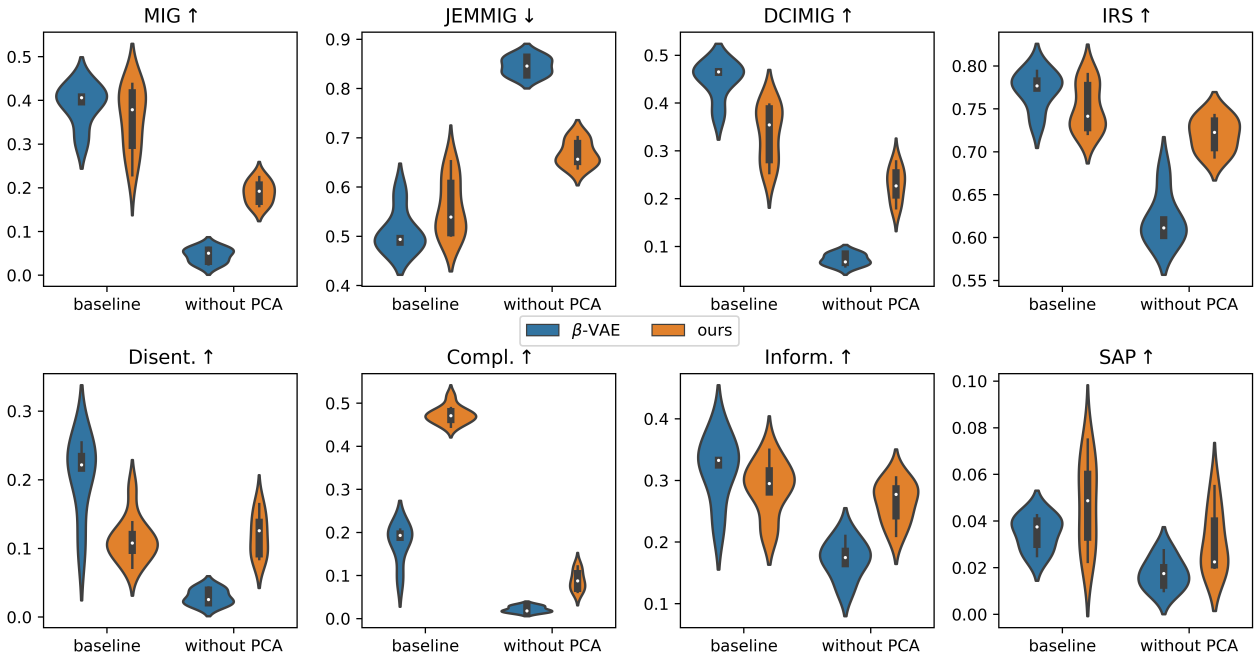


Figure 8. Comparison on 3DFaces dataset between our model and a β-VAE model under two ablation settings: the original setting (denoted as "baseline") and the setting where the PCA inductive biases of models are removed (denoted as "without PCA").

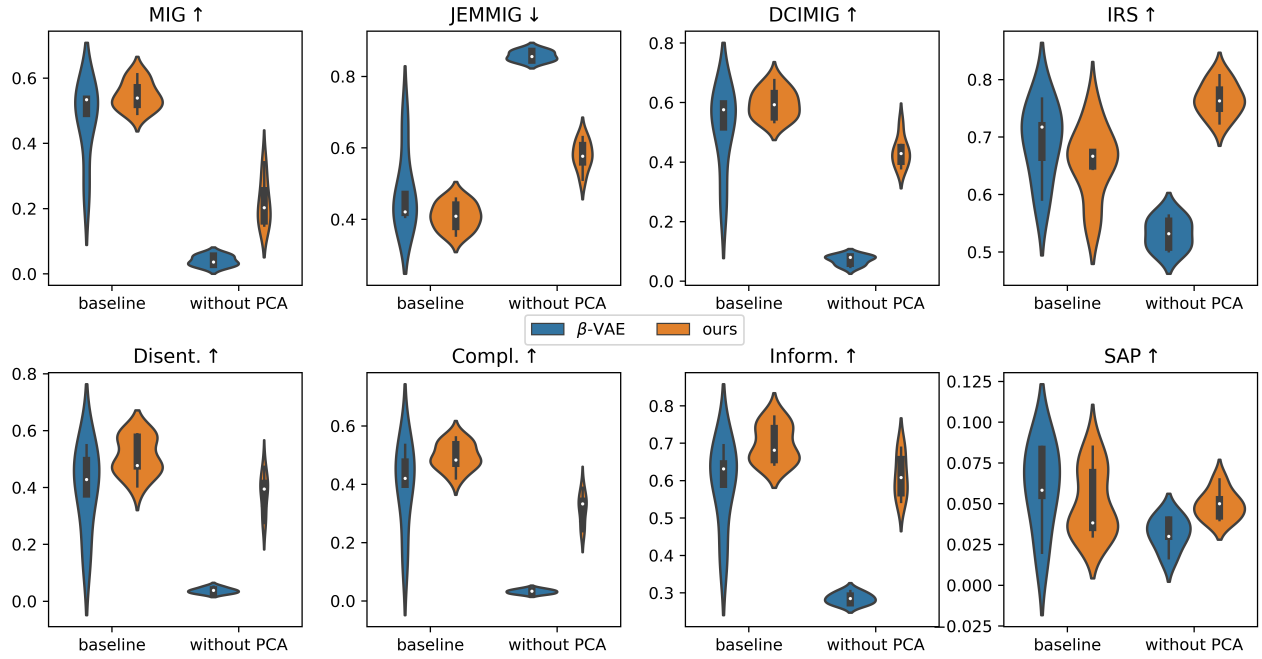


Figure 9. Comparison on 3DShapes dataset between our model and a β-VAE model under two ablation settings: the original setting (denoted as "baseline") and the setting where the PCA inductive biases of models are removed (denoted as "without PCA").

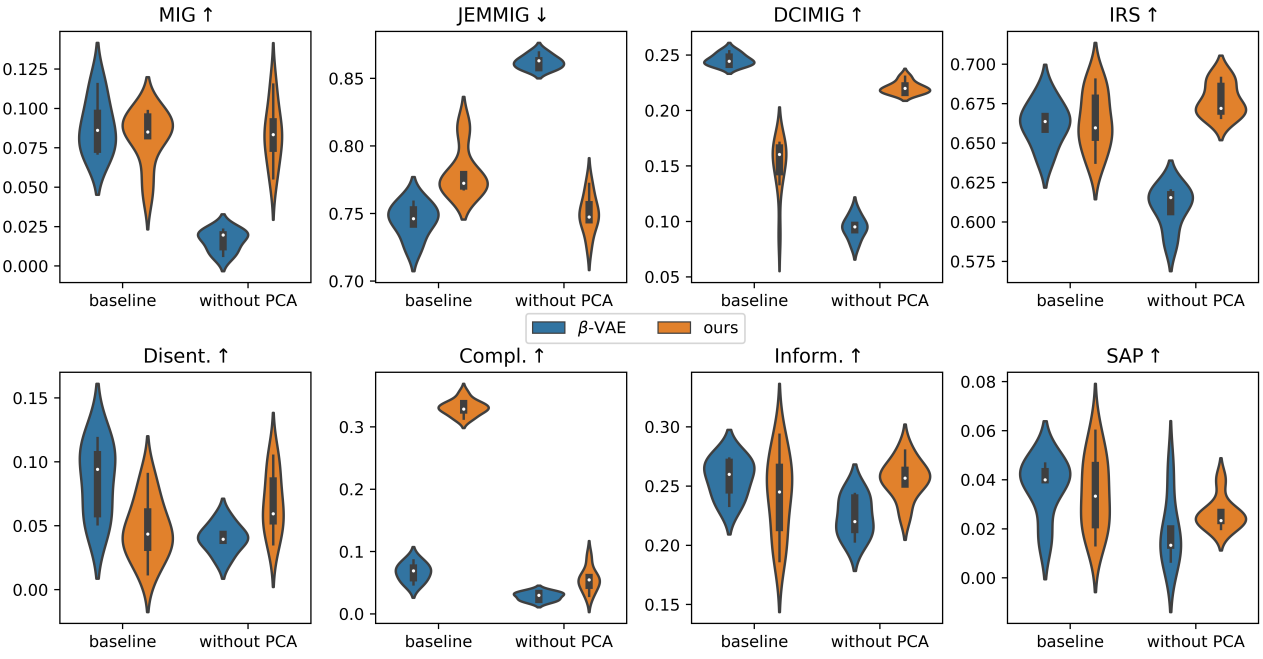


Figure 10. Comparison on 3DCars dataset between our model and a β-VAE model under two ablation settings: the original setting (denoted as "baseline") and the setting where the PCA inductive biases of models are removed (denoted as "without PCA").

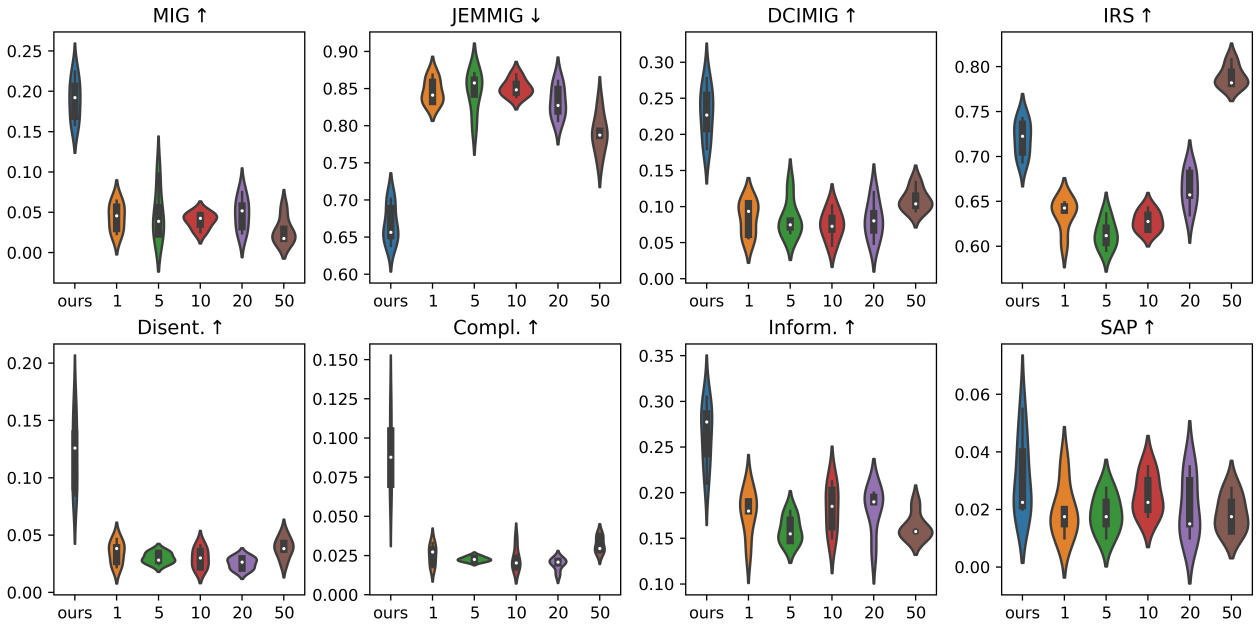


Figure 11. The isotropic β-VAE model consistently achieves low disentangling performance on 3DFaces for different hyper-parameters β, where for horizontal axes, "ours" denotes our model for comparison and numbers denote the value of β for the isotropic β-VAE model.

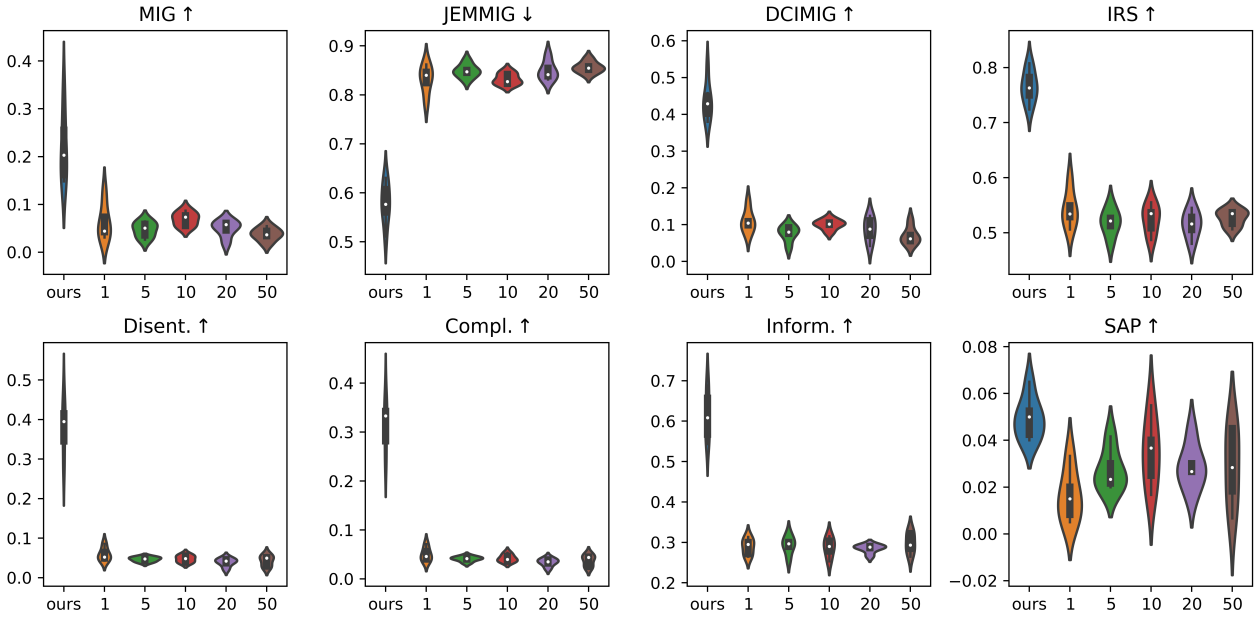


Figure 12. The isotropic β-VAE model consistently achieves low disentangling performance on 3DShapes for different hyper-parameters β, where for horizontal axes, "ours" denotes our model for comparison and numbers denote the value of β for the isotropic β-VAE model.

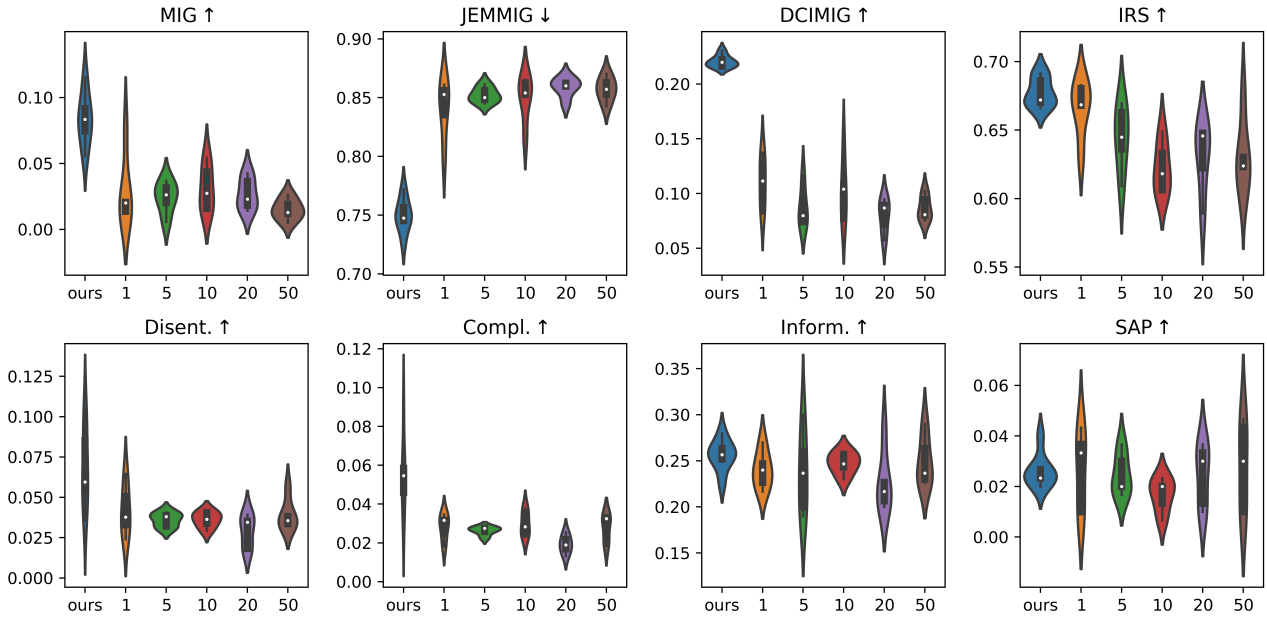


Figure 13. The isotropic β-VAE model consistently achieves low disentangling performance on 3DCars for different hyper-parameters β, where for horizontal axes, "ours" denotes our model for comparison and numbers denote the value of β for the isotropic β-VAE model.

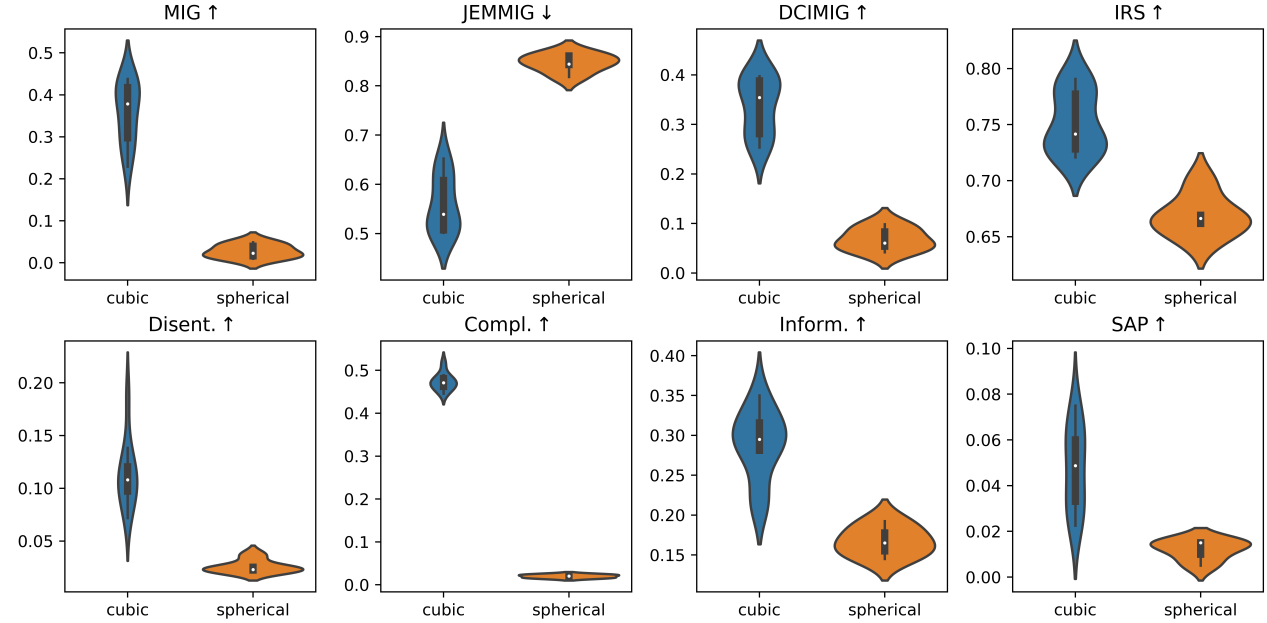


Figure 14. The model with a spherical latent space achieve much lower disentangling performances than the model with a cubic latent space on 3DFaces dataset. For horizontal axes, "cubic" and "spherical" denote latent space structures.

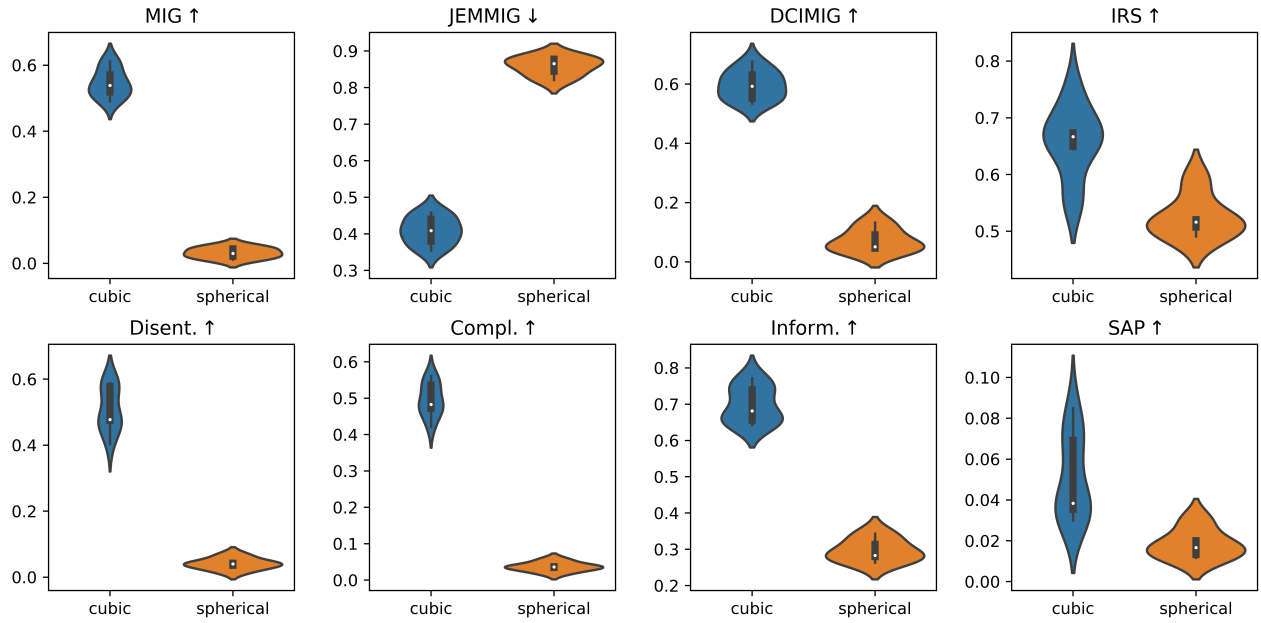


Figure 15. The model with a spherical latent space achieve much lower disentangling performances than the model with a cubic latent space on 3DShapes dataset. For horizontal axes, "cubic" and "spherical" denote latent space structures.

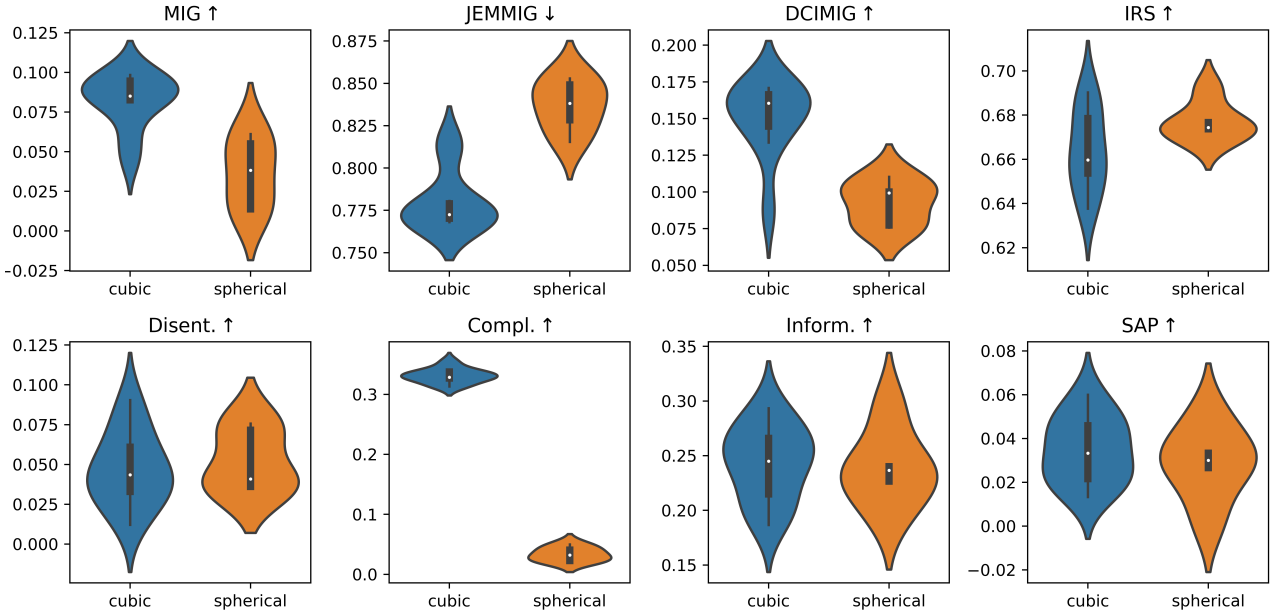


Figure 16. The model with a spherical latent space achieve much lower disentangling performances than the model with a cubic latent space on Cars3D dataset. For horizontal axes, "cubic" and "spherical" denote latent space structures.

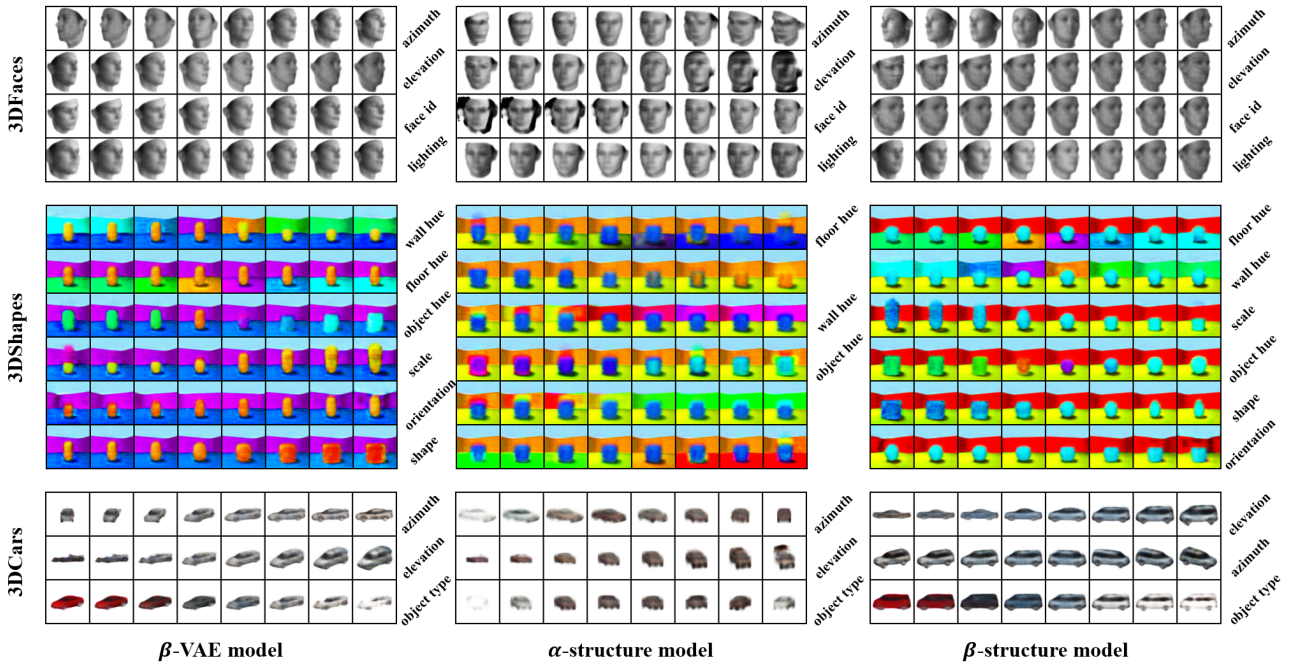


Figure 17. Visualization of latent traversal for different models on different datasets. The latent traversal is done by first randomly sample a latent code $z \in \mathcal{Z}$, then varying each latent dimension z^i from -1 to 1 for α and β-structure models and -3 to 3 for β-VAE model, while keeping other latent dimension $z^{j \neq i}$ fixed, and plot the corresponding reconstruction. For β-VAE model and β-structure model, the latent space dimensionality is set to 10 as we aforementioned, and we select the top K^* informative latent channels to visualize the corresponding latent traversals, where K^* is the ground-truth number of factors of variation of the dataset. The informativeness of latent channels are reflected by the posterior Gaussian variances for β-VAE, while for β-structure model the informativeness of latent channels are reflected by the magnitude of singular values. For α-structure model, the latent dimensionality is directly set to K^*. For traversals of each latent dimension, we subjectively judge the possible corresponding generative factor from semantic meaning of visualization.

References

[1] C. Burgess and H. Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

[2] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[3] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[4] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[5] K. Do and T. Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019.

[6] C. Eastwood and C. K. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[7] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] I. Kobyzev, S. Prince, and M. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[10] A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

[11] Z. Pan, L. Niu, and L. Zhang. Unigan: reducing mode collapse using a uniform generator. In *NeurIPS*, 2022.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[13] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[14] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28:1252–1260, 2015.

[15] M. Rolinek, D. Zietlow, and G. Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

[16] A. Sepliarskaia, J. Kiseleva, M. de Rijke, et al. Evaluating disentangled representations. *arXiv preprint arXiv:1910.05587*, 2019.

[17] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

[18] J. Zaidi, J. Boilard, G. Gagnon, and M.-A. Carbonneau. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*, 2020.

[19] D. Zietlow, M. Rolinek, and G. Martius. Demystifying inductive biases for β-vae based architectures. *arXiv preprint arXiv:2102.06822*, 2021.