α-TCVAE: On the relationship between Disentanglement and Diversity

Cristian Meo[*] TUDelft, NL.

Louis Mahon University of Edinburgh, UK.

Anirudh Goyal Google DeepMind, UK.

Justin Dauwels TUDelft, NL.

Abstract

Understanding and developing optimal representations has long been foundational in machine learning (ML). While disentangled representations have shown promise in generative modeling and representation learning, their downstream usefulness remains debated. Recent studies re-defined disentanglement through a formal connection to symmetries, emphasizing the ability to reduce latent domains (i.e., ML problem spaces) and consequently enhance data efficiency and generative capabilities. However, from an information theory viewpoint, assigning a complex attribute (i.e., features) to a specific latent variable may be infeasible, limiting the applicability of disentangled representations to simple datasets. In this work, we introduce α-TCVAE, a variational autoencoder optimized using a novel total correlation (TC) lower bound that maximizes disentanglement and latent variables informativeness. The proposed TC bound is grounded in information theory constructs, generalizes the β-VAE lower bound, and can be reduced to a convex combination of the known variational information bottleneck (VIB) and conditional entropy bottleneck (CEB) terms. Moreover, we present quantitative analyses and correlation studies that support the idea that smaller latent domains (i.e., disentangled representations) lead to better generative capabilities and diversity. Additionally, we perform downstream task experiments from both representation and RL domains to assess our questions from a broader ML perspective. Our results demonstrate that α-TCVAE consistently learns more disentangled representations than baselines and generates more diverse observations without sacrificing visual fidelity. Notably, α-TCVAE exhibits marked improvements on MPI3D-Real, the most realistic disentangled dataset in our study, confirming its ability to represent complex datasets when maximizing the informativeness of individual variables. Finally, testing the proposed model off-the-shelf on a state-of-the-art model-based RL agent, Director, significantly shows α-TCVAE downstream usefulness on the loconav Ant Maze task. Implementation available at `https://github.com/Cmeo97/Alpha-TCVAE`

1 Introduction

The efficacy of machine learning (ML) algorithms is intrinsically tied to the quality of data representation (Bengio et al., 2013). Such representations are useful not only for standard downstream tasks such as supervised learning (Alemi et al., 2017) and reinforcement learning (RL) (Li, 2017), but also for tasks such as transfer learning (Zhuang et al., 2020) and zero-shot learning (Sun et al., 2021). Unsupervised representation learning aims to identify semantically meaningful representations of data without supervision, by capturing the generative factors of variations that describe the structure of the data (Radford et al., 2016; Locatello et al., 2019b). According to Bengio et al. (2013), disentanglement learning holds the key to understanding the world from observations, generalizing knowledge across different tasks and domains while learning and generating compositional representations (Higgins et al., 2016; Kim & Mnih, 2018).

Problem Formulation. The goal of disentanglement learning is to identify a set of independent generative factors $\boldsymbol{z}$ that give rise to the observations $\boldsymbol{x}$ via $p(\boldsymbol{x}|\boldsymbol{z})$. However, from an information

[*] Work done while doing a research internship at Mila, Quebec AI Institute.

theory perspective, the amount of information retained by every latent variable may be insufficient to represent realistic generative factors (Kirsch et al., 2021; Do & Tran, 2020), limiting the applicability of disentangled representations to simple problems. What is more, Friedman & Dieng (2022) recently introduced the Vendi score, a new metric for gauging generative diversity, showing that entangled generative models, such as the Very Deep VAE (Child, 2020), consistently produce samples with less diversity compared to ground truth. This is indicative of their limited representational and generative prowess. In contrast, Higgins et al. (2019; 2022) re-defined disentangled representations through the lens of symmetries, linking disentanglement to computational problem spaces (e.g., disentangled representations inherently reduce the problem space (Arora & Barak, 2009)), suggesting that disentangled models should be able to explore and traverse the latent space more efficiently, leading to enhanced generative diversity.

Previous Work. Most existing disentangled models optimize lower bounds that only impose an information bottleneck on the latent variables, and while this can result in factorized representations (Higgins et al., 2016), it does not directly optimize latent variable informativeness (Do & Tran, 2020). As a result, while several approaches have been proposed to learn disentangled representations by optimizing different bounds (Chen et al., 2018; Kim & Mnih, 2018), imposing sparsity priors (Mathieu et al., 2019), or isolating source of variance (Rolinek et al., 2019), none of the proposed models successfully learned disentangled representations of realistic datasets. Moreover, to the best of our knowledge, no systematic and quantitative analyses have been proposed to assess to what extent disentanglement and generative diversity (Friedman & Dieng, 2022) are correlated.

Proposed method. In this work, we propose α-TCVAE, a VAE optimized using a novel convex lower bound of the joint total correlation (TC) between the learned latent representation and the input data. The proposed bound, through a convex combination of the variational information bottleneck (VIB) Alemi et al. (2017) and the conditional entropy bottleneck (CEB) Fischer & Alemi (2020), maximizes the average latent variable informativeness, improving both representational and generative capabilities. The effectiveness of α-TCVAE is especially prominent in the MPI3D-Real Dataset (Gondal et al., 2019), the most realistic dataset in our study that is compositionally built upon distinct generative factors. Figure 1 illustrates a comparison of the latent traversals between α-TCVAE, Factor-VAE and β-VAE, showing that α-TCVAE leads to the best visual fidelity and generative diversity (i.e., higher Vendi Score). Interestingly, the proposed TC bound is grounded in information theory constructs, generalizes the β-VAE (Higgins et al., 2016) lower bound, and can be reduced to a convex combination of the known variational information bottleneck (VIB) (Alemi et al., 2017) and conditional entropy bottleneck (CEB) (Fischer & Alemi, 2020) terms.

Experimental Evaluation In order to determine the effectiveness of α-TCVAE and the downstream usefulness of the learned representations, we measure the diversity and quality of generated images and disentanglement of its latent representations. Then, we perform a correlation study between the considered downstream scores across all models, analyzing how generative diversity and disentanglement are related across different datasets. This analysis substantiates our claim that disentanglement leads to improved diversity. Finally, we conduct experiments to assess the downstream usefulness of the proposed method from a broader ML perspective. Notably, the proposed method consistently outperforms the related baselines, showing a significant improvement in the RL Ant Maze task when applied off-the-shelf in Director, a hierarchical model-based RL agent (Hafner et al., 2022).

2 Related Work

Generative Modelling and Disentanglement Recently Locatello et al. (2019b) demonstrated that unsupervised disentangled representation learning is theoretically impossible, nonetheless disentangled VAEs, acting as both representational and generative models, Kingma & Welling (2013); Higgins et al. (2016); Chen et al. (2018); Kim & Mnih (2018) achieve practical results by leveraging implicit biases within the data and learning dynamics Burgess & Kim (2018); Higgins et al. (2019); Mathieu et al. (2019). On the generation side, they have been widely used to generate data such as images (Chen et al., 2019), text (Shi et al., 2019), speech (Sun et al., 2020; Li et al., 2023) and music (Wang et al., 2020). Various extensions to the base VAE model have been presented to improve generation quality in terms of visual fidelity (Peng et al., 2021; Vahdat & Kautz, 2020; Razavi et al.,

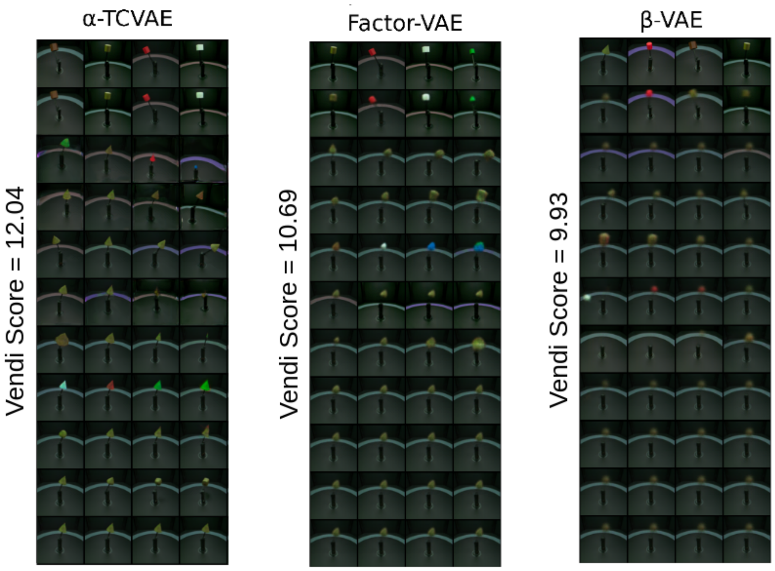


Figure 1: Ground truth (first row), reconstructions (second row) and latent traversals comparison of α-TCVAE, Factor-VAE, and β-VAE on the MPI3D-Real Dataset. Notably, α-TCVAE showcases superior visual fidelity and generative diversity, as indicated by a higher Vendi Score.

2019). On the representational side, aiming for explainable and factorized representations, Higgins et al. (2016) proposed β-VAE, which inspired a number of following disentangled VAE-based models, such as Factor-VAE (Kim & Mnih, 2018), β-TCVAE Chen et al. (2018), and $\beta-$Annealed VAE (Burgess et al., 2018). Both β-VAE and Factor-VAE aim to learn disentangled representations by imposing a bottleneck on the information flowing through the latent space. While β-VAE does this by introducing a β hyperparameter that increases the strength of the information bottleneck, Factor-VAE introduces a TC regularization term. Chen et al. (2018) proposed β-TCVAE, which minimizes the total correlation of the latent variables using Monte-Carlo and importance sampling. Roth et al. (2023) proposed the Hausdorff Factorized Support (HFS) criterion, a relaxed disentanglement criterion that encourages only pairwise factorized support, rather than a factorial distribution, by minimizing a Hausdorff distance. This allows for arbitrary distributions of the factors over their support, including correlations between them. Our model, namely α-TCVAE is optimized by a TC lower bound as well, however we do not make use of any trick or expensive sampling strategy. In contrast, we derive a TC lower bound that does not require any extra network or sampling strategy and is theoretically grounded in the Deep Information Bottleneck framework Alemi et al. (2017).

Disentanglement and Deep Information Bottleneck In the last few years, a link between the latent space capacity and disentanglement of the learned variables (Bengio et al., 2013; Shwartz-Ziv & Tishby, 2017; Goyal et al., 2021) has been identified, showing that decreasing the capacity of a network induces disentanglement on the learned representations. This relationship has been explained by the information bottleneck (IB) principle, introduced by Tishby et al. (2001) as a regularization method to obtain minimal sufficient encoding by constraining the amount of information captured by the latent variables from the observed variable. Variational IB (VIB) (Alemi et al., 2017) has extended the IB framework by applying it to neural networks, which results in a simple yet effective method to learn representations that generalize well and are robust against adversarial attacks. Furthermore, (Alemi et al., 2017; Kirsch et al., 2021) outlined the relationship between VIB, VAE (Kingma & Welling, 2013) and β-VAE (Higgins et al., 2016), providing an information theoretical interpretation of the Kullback-Leibler (KL) divergence term used in these models as a regularizer. Despite the advantages introduced by the VIB framework, imposing independence between every latent variable may be too strong an assumption (Roth et al., 2023). For this reason, Fischer & Alemi (2020) introduced the conditional entropy bottleneck (CEB), which assumes conditional independence between the learned latent variables, providing the ability to learn more expressive and robust representations (Kirsch et al., 2021). Recently, a generalization of the mutual Information (MI), namely total correlation (TC), has been used to learn disentangled representations as well (Kim & Mnih, 2018). Following Hwang et al. (2021), who propose a similar TC bound for a multi-view setting, we derive a novel TC lower bound for the unsupervised representational learning setting. As a result, the proposed bound is able to learn expressive and disentangled representations.

Disentanglement and Diversity The ideal generative model learns a distribution that well explains the observed data, which can then be used to draw a diverse set of samples. Diversity is thus an important desirable property of generative models (Friedman & Dieng, 2022). We desire the ability to produce samples that are different from each other and from the samples we already have at train time, while still coming from the same underlying distribution. The benefits of diversity have been advocated in a number of different contexts, such as image synthesis (Mao et al., 2019), molecular design (Lin et al., 2021; Wu et al., 2021), natural language text (McCoy et al., 2023), and drug discovery (Kim & Mnih, 2018). Motivated by the benefits of generative diversity, several VAE-based models have aimed to show increased diversity in their generated samples (Razavi et al., 2019). Some works have also noted improvements in diversity due to disentanglement. Lee et al. (2018) adversarially disentangle style from content and show enhanced diversity of image-to-image translations. Kazemi et al. (2019) also perform style-content disentanglement, this time in the context of text generation, and again observe an increase in diversity. Li et al. (2020) shows that disentangling pose, shape, and texture leads to greater diversity in generated images. Collectively, these studies emphasize that diversity is often a valuable indicator of effectiveness in various applications, and suggest that diversity and disentanglement are intertwined aspects of generative models. Yet, to the best of our knowledge, no quantitative analyses that support this claim have been presented. In this work, we present a correlation study, showing how downstream metrics of disentanglement (e.g., DCI (Eastwood & Williams, 2018)) and diversity (e.g., Vendi Score (Friedman & Dieng, 2022)) are correlated across several models and datasets.

3 α-TCVAE Framework Derivation

Motivation. In contrast to most existing methods, which only impose an information bottleneck to learn disentangled representations, we seek to maximize the informativeness of individual latent variables as well. The total joint correlation (TC) can be explicitly expressed in terms of mutual information between the observed data and the latent generative factors, as shown in equation 4, allowing us to link disentanglement to latent variables informativeness. As a result, leveraging the TC formulation, we can derive a lower bound that not only promotes disentanglement but also maximizes the information retained by individual latent variables.

Derivation. In this section, we formally derive the novel TC bound. Let $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{V}\}$ be the ground-truth set that consists of images $\boldsymbol{x} \in \mathbb{R}^{N \times N}$, and a set of conditionally independent ground-truth data generative factors $\boldsymbol{v} \in \mathbb{R}^M$, where $\log p(\boldsymbol{v}|\boldsymbol{x}) = \sum_k \log p(v_k|\boldsymbol{x})$. The goal is to develop an unsupervised deep generative model that can learn the joint distribution of the data $\boldsymbol{x}$, while uncovering a set of generative latent factors $\boldsymbol{z} \in \mathbb{R}^K$, $K \geq M$, such that $\boldsymbol{z}$ can fully describe the data structure of $\boldsymbol{x}$ and generate data samples that follow the underlying ground-truth generative factors $\boldsymbol{v}$. Since directly optimizing the joint TC is intractable, we are going to maximize a lower bound of the joint total correlation $TC(\boldsymbol{z}, \boldsymbol{x})$ between the learned latent representations $\boldsymbol{z}$ and the input data $\boldsymbol{x}$, following the approach proposed by Hwang et al. (2021). The total correlation is defined as the KL divergence between the joint distribution and the factored marginals. In our case:

$$TC_\theta(\boldsymbol{z}) \triangleq D_{KL}\left[\int q_\theta(\boldsymbol{z}|\boldsymbol{x}) p_D(\boldsymbol{x}) d\boldsymbol{x} \| \prod_{k=1}^{K} q_\theta(\boldsymbol{z}_k)\right], \tag{1}$$

where the joint distribution is $q_\theta(\boldsymbol{z}) = \int q_\theta(\boldsymbol{z}|\boldsymbol{x}) p_D(\boldsymbol{x}) d\boldsymbol{x}$, $p_D(\boldsymbol{x})$ is the data distribution, $q_\theta(\boldsymbol{z}_k) = \int q_\theta(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}_{\neq k}$ and $\boldsymbol{z}_{\neq k}$ indicates that the k-th component of $\boldsymbol{z}$ is not considered. Since we aim to find the encoder $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ that disentangles the learned representations $\boldsymbol{z}$, we can formulate the following objective:

$$TC_\theta(\boldsymbol{z}, \boldsymbol{x}) \triangleq TC_\theta(\boldsymbol{z}) - TC_\theta(\boldsymbol{z}|\boldsymbol{x}), \tag{2}$$

where the conditional TC($\boldsymbol{z}|\boldsymbol{x}$) can be expressed as:

$$TC_\theta(\boldsymbol{z}|\boldsymbol{x}) \triangleq \mathbb{E}_{q_\theta(\boldsymbol{z})}\left[D_{KL}\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| \prod_{k=1}^{K} q_\theta(\boldsymbol{z}_k|\boldsymbol{x})\right]\right], \tag{3}$$

which is the expected KL divergence of the joint conditional from the factored conditionals. Intuitively, we can see that minimizing $TC_\theta(\boldsymbol{z}|\boldsymbol{x})$, $TC_\theta(\boldsymbol{z},\boldsymbol{x})$ will be maximized, enhancing the disentanglement of the learned representation $\boldsymbol{z}$. Moreover, decomposing equation 2 we can express the TC in terms of MI (Gao et al., 2019):

$$TC_\theta(\boldsymbol{z},\boldsymbol{x}) = \sum_{k=1}^{K} I_\theta(\boldsymbol{z}_k,\boldsymbol{x}) - I_\theta(\boldsymbol{z},\boldsymbol{x}), \tag{4}$$

where $I_\theta(\boldsymbol{z},\boldsymbol{x})$ is the mutual information between $\boldsymbol{z}$ and $\boldsymbol{x}$ and is known as the VIB term Alemi et al. (2017). Additionally, we can also express it in terms of Conditional MI:

$$TC_\theta(\boldsymbol{z},\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\left[(K-1)I_\theta(\boldsymbol{z}_k,\boldsymbol{x}) - I_\theta(\boldsymbol{z}_{\neq k},\boldsymbol{x}|\boldsymbol{z}_k)\right], \tag{5}$$

where $I_\theta(\boldsymbol{z}_{\neq k},\boldsymbol{x}|\boldsymbol{z}_k)$ is known as the CEB term (Fischer & Alemi, 2020). Equation 4 and equation 5 illustrate the link of the designed objective to both VIB and CEB frameworks. A complete derivation of them can be found in Appendices A.1 and A.2, respectively. While the VIB term promotes compression of the latent representation, the CEB term promotes balance between the information contained in each latent dimension. Since we want to promote both disentanglement and individual variable informativeness of the learned latent representation we propose a lower bound that convexly combines the found VIB and CEB terms. We define the bound as follows:

$$TC(\boldsymbol{z},\boldsymbol{x}) \geq \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})\right] - \underbrace{\frac{K\alpha}{K-\alpha} D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r_p(\boldsymbol{z}|\boldsymbol{x}))}_{\text{CEB}} - \frac{(1-\alpha)}{(1-\frac{\alpha}{K})}\underbrace{D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))}_{\text{VIB}}, \tag{6}$$

where α is a hyperparameter that trades off VIB and CEB terms. Following Hwang et al. (2021), we define $r_p(\boldsymbol{z}|\boldsymbol{x}) = N(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p\boldsymbol{I})$ and $r(\boldsymbol{z}) = N(\boldsymbol{0},\boldsymbol{I})$, respectively, where $\boldsymbol{\sigma}_p \triangleq \left(\sum_{k=1}^{K}\frac{1}{\boldsymbol{\sigma}_k^2}\right)^{-1}$ and $\boldsymbol{\mu}_p \triangleq \boldsymbol{\sigma}_p \cdot \sum_{k=1}^{K}\frac{\boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k^2}$ while $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ are the mean and standard deviation used to compute $\boldsymbol{z}_k$ using the reparametrization trick as in Kingma & Welling (2013). A full derivation of the bound defined in equation 6 can be found in Appendix A.

Practical Implications. Disentangled models with M generative factors and K latent dimensions usually have $(K-M)$ noisy latent dimensions Do & Tran (2020), but our CEB term induces an inductive bias on the information flowing through every individual latent variable, pushing otherwise noisy dimensions to be informative. The derived TC lower bound generalizes the structure of the widely used β-VAE (Higgins et al., 2016) bound. Indeed, for $\alpha = 0$, the TC bound reduces to β-VAE one. A comparison of α-TCVAE, β-VAE, β-TCVAE, HFS and Factor-VAE lower bounds can be found in Tab. A.3.

4 EXPERIMENTS

In this section, we design empirical experiments to understand the performance of α-TCVAE and its potential limitations by exploring the following questions: (1) Does maximising the informativeness of latent variables consistently lead to an increase in representational power and generative diversity? (2) Do disentangled representations inherently present higher diversity than entangled ones? (3) How are they correlated with other downstream metrics (i.e., FID (Heusel et al., 2017) and unfairness (Locatello et al., 2019a))? (4) To what extent does maximising the latent variables' informativeness in disentangled representations improve their downstream usefulness?

Experimental Setup. In order to assess the performance of both proposed and baseline models, we validate the considered models on the following datasets. **Teapots** (Moreno et al., 2016) contains $200,000$ images of teapots with features: azimuth and elevation, and object colour. **3DShapes** (Burgess & Kim, 2018) contains $480,000$ images, with features: object shape and colour, floor colour, wall colour, and horizontal orientation. **MPI3D-Real** (Gondal et al., 2019) contains $103,680$ images of objects at the end of a robot arm, with features: object colour, size, shape, camera height, azimuth, and robot arm altitude. **Cars3D** (Reed et al., 2015) contains $16,185$ images with features: car-type, elevation, and azimuth. **CelebA** (Liu et al., 2015) contains over $200,000$ images of faces under a broad range of poses, facial expressions, and lighting conditions, totalling 40 different factors. All datasets under consideration consist of RGB images with dimensions 64×64. Among them, CelebA stands out as the most realistic and complex dataset. On the other hand, MPI3D-Real is considered the most realistic among factorized datasets, which we define as those compositionally generated using independent factors. To assess the generated images, we use the FID score (Heusel et al., 2017) to measure the distance between the distributions of generated and real images, and the Vendi score (Friedman & Dieng, 2022) to measure the diversity of sampled images. Both Vendi and FID use the Inception Network (Szegedy et al., 2017) to extract image features and compute the related similarity metrics. Since **DCI** (Eastwood & Williams, 2018) scores can produce unreliable results in certain cases, (Mahon et al., 2023; Cao et al., 2022; Do & Tran, 2020), we measure disentanglement using also single neuron classification SNC (Mahon et al., 2023). Further details on used datasets and metrics are given in Appendix C.

Baseline Methods. We compare α-TCVAE to four other VAE models: β-VAE (Higgins et al., 2016), β-TCVAE (Chen et al., 2018), β-VAE+HFS (Roth et al., 2023) and FactorVAE (Kim & Mnih, 2018), all of which are described in Section 2, as well as a vanilla VAE (Kingma & Welling, 2013). To assess diversity and visual fidelity beyond VAE-based models, we also compare to a generative adversarial network model, StyleGAN (Karras et al., 2019).

Generation Faithfulness and Diversity Analyses. We present image generation results from our model alongside baseline models, evaluating performance on the FID and Vendi metrics across datasets. For image generation using VAE-models, we adopt two strategies: (1) Sampling a noise vector from a multivariate standard normal and decoding it. (2) Encoding an actual image, then selecting a latent dimension. The value of this chosen dimension is adjusted by shifts of $+/-$ 1, 2, 4, 6, 8, or 10 standard deviations. Subsequently, we decode the adjusted representation. In Figures 2 and 3 the two sampling strategies are labeled as 'Sampled from Noise' and 'Sampled from Traversals' respectively. Figures 2 and 3 show that α-TCVAE consistently generates more diverse (higher Vendi) and more faithful (lower FID) images than baseline VAE models.

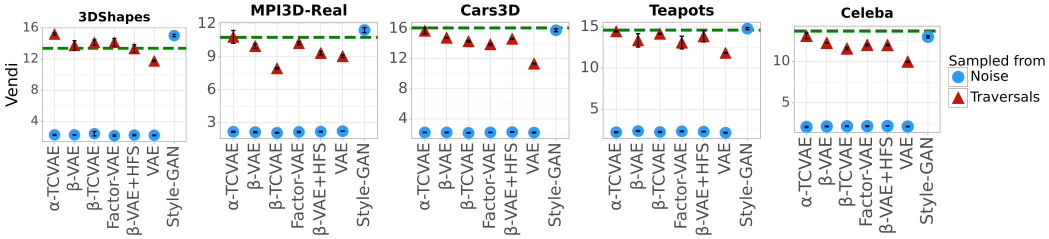


Figure 2: Diversity of generated images, as measured by Vendi score. Two different sampling strategies are considered: sampled from white noise and from traversals. The diversity of the images of our model, α-TCVAE, is consistently higher than baseline VAE models, and on par with StyleGAN. The green dashed line represents ground truth dataset diversity. Traversals produce significantly more diverse images than samples.

The Vendi score of α-TCVAE is comparable to that of Style-GAN, and its FID score is only slightly worse. Moreover, Style-GAN takes 15x the training time ($\sim$ 2hrs vs. $>$ 30hrs on a single Nvidia Titan XP) and learns only a generative model, whereas VAEs learn both a generative model and a representational model. Noticeably, all VAE-based models perform poorly in terms of both diversity and reconstruction quality when sampling from white noise, highlighting the benefit of a structured sampling strategy when using VAE-based models for generative tasks. Another finding is that traversal-generated images are superior to those obtained from the prior, i.e. sampling from a standard normal and decoding. This is in keeping with prior work showing that drawing latent samples from a distribution other than the standard normal, e.g. a GMM, often leads to higher quality generated images Chadebec et al. (2022), and it supports the claim that disentangled models allow more systematic exploration of the latent space leading to more diverse images. This claim is also supported by noting that all disentangled VAEs give higher diversity than the vanilla VAE.

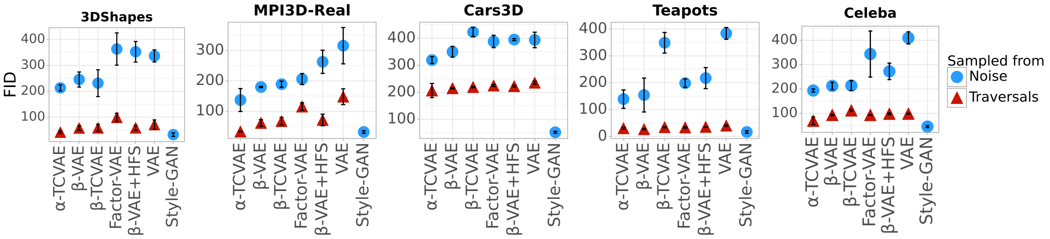


Figure 3: Faithfulness of generated images to the data distribution, as measured by FID score. Two different sampling strategies are considered: sampled from white noise and from traversals. The scores for the images of our model, α-TCVAE, are consistently better than baseline VAE models (lower FID is better), and only slightly worse than StyleGAN. Traversals produce significantly more faithful images than samples.

Disentanglement Analyses and Downstream Metrics Correlation Study In this section we examine the disentanglement capabilities of α-TCVAE and the related VAE baselines, and how it relates, statistically, with the diversity and quality of generated images, as measured in Section 4. Figures 4, 5 and 6 show that α-TCVAE consistently achieves comparable or better DCI, SNC and unfairness scores. The improvement of α-TCVAE over the baselines is most significant on the most realistic factorized dataset, namely MPI3D-Real. Interestingly, while there is a significant gap between the DCI scores of disentangled and entangled models across every factorized dataset, SNC shows that in terms of single neuron factorization, for both Cars3D and MPI3D-Real, α-TCVAE is the only model that significantly improves over the entangled VAE. This is perhaps due to the tendency of DCI to sometimes overestimate disentanglement Mahon et al. (2023); Cao et al. (2022). Furthermore, as illustrated in Figure 4, no model has been successful in learning disentangled representations from the CelebA dataset. To meaningfully encode CelebA images, we used high-dimensional latent representations (e.g., 48 dimensions). However, as highlighted by Do & Tran (2020), disentangling and measuring disentanglement in high-dimensional representations are notoriously challenging tasks. Indeed, while DCI and unfairness present unrealistic results, SNC gave all models a score of zero, and so we do not display the figures here. Figure 10 illustrates a significant correlation between the Vendi, unfairness, and DCI metrics. There is a compelling correlation between Vendi and DCI scores, underscoring that diversity and disentanglement are statistically related. This resonates with the understanding that disentangled latent spaces naturally exhibit superior generative diversity (Higgins et al., 2019). Additionally, Vendi and DCI both exhibit a negative correlation with unfairness. This observation is consistent with Locatello et al. (2019a)'s findings, implying that the fairness of downstream prediction tasks is deeply associated with the diversity and disentanglement of the representations being learned. Further correlations results are given in Appendix D, along with examples of latent traversals.

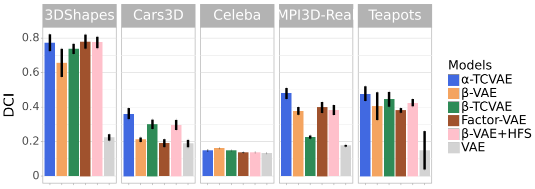


Figure 4: Comparison of DCI scores of our model with those of baseline models.

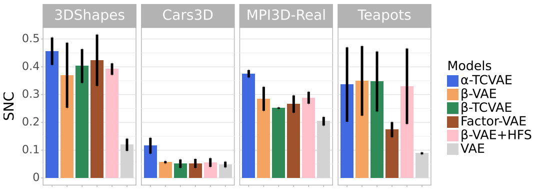


Figure 5: Comparison of SNC scores of our model with those of baseline models.

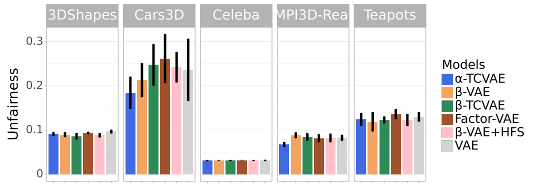


Figure 6: Comparison of unfairness scores of our model with those of baseline models.

Attribute Classification Task In this experiment, we train a multilayer perceptron (MLP)

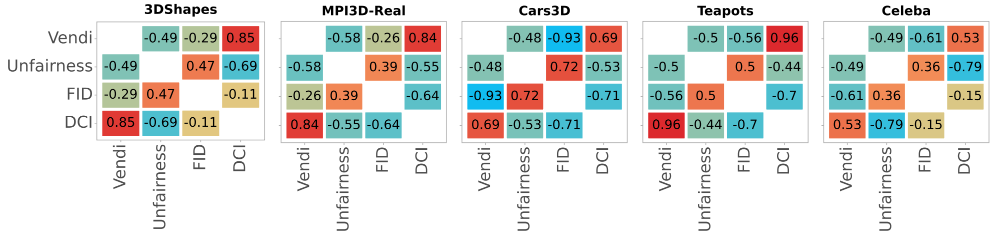


Figure 7: Correlations between diversity (Vendi score), generation faithfulness (FID score), unfairness and DCI. Correlations are computed using the results from all models across 5 different seeds.

to classify sample attributes using the models' encoded latent representations. Figure 8 reveals that α-TCVAE either matches or surpasses the baseline models in terms of attribute classification accuracy. The improvement is minor on 3DShapes and Teapots, but more significant on Cars3D and MIP3D-Real. Interestingly, the only dataset where all VAEs exhibit commendable performance is CelebA, where high-dimensional representations are used. This suggests that, for this particular downstream task, the dimensionality of the representation may be the main constraining factor. In fact, this downstream task inherently favours high-dimensional attributes, considering that a MLP is employed for the attribute classification.

Loconav Ant Maze Reinforcement Learning Task. In this experiment, a model-based RL agent has to learn its proprioceptive dynamical system while escaping from a maze. Recently, Hafner et al. (2022) introduced Director, a hierarchical model-based RL agent. Director employs a hierarchical strategy with a Goal VAE that learns and generates sub-goals, simplifying the planning task. The first hierarchy level represents the agent's internal states, while in the second one, the Goal VAE encodes the agent's state and infers sub-goals. As a result, the Goal VAE generates sub-goals to guide the agent through the environment. Given the enhanced generative diversity of α-TCVAE, we postulated that integrating our proposed TC bound could improve Director's exploration. In this experiment, we replaced the beta-VAE objective, used to train Director's Goal VAE, with our TC-bound, expecting a richer diversity in sub-goals, thus expediting environment exploration and enhancing overall learning behaviour. Figure 9 compares the performance of Director and Alpha-

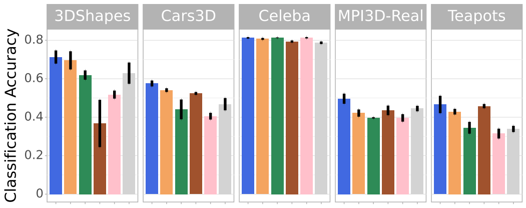
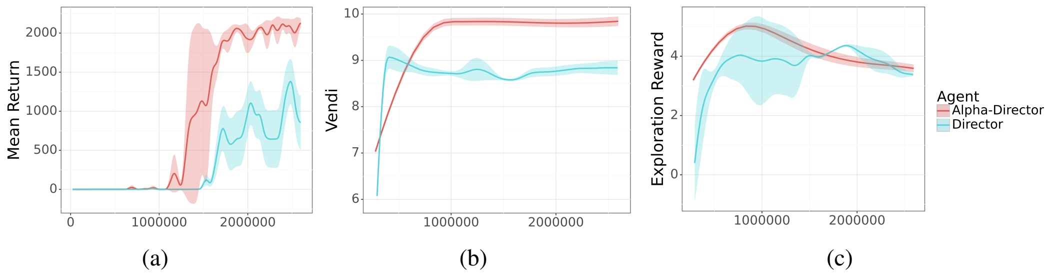


Figure 8: Comparison of α-TCVAE and baseline models on the Downstream Attribute Classification Task. Our proposed model either matches or surpasses the baseline models in terms of attribute classification accuracy



(a) (b) (c)

Figure 9: Performance of Director, a model-based hierarchical RL agent, and Alpha-Director on the Antmaze task. While director samples sub-goals using the original β-VAE, Alpha-Director samples sub-goals using the proposed α-TCVAE. Sampling using α-TCVAE gives more diverse goals (b), better exploration (c) and significantly higher mean return (a).

Director, which replaces β-VAE objective with the proposed TC-bound instead, the results are averaged across three seeds. Figure 9-(a) presents the mean return, which scores the performances of the agent on the given task (i.e., finding the exit of the maze while learning proprioceptive dynamics), showing that Alpha-Director significantly outperforms Director, learning faster and to a higher final high mean reward. Figure 9-(b) illustrates the Vendi score of sampled goals across batch and sequence length, showing that α-TCVAE generates sub-goals with a higher diversity score. As a result, Alpha-Director has a better exploration, as shown in Figure 9-(c), leading to faster learning. Collectively, these findings highlight that α-TCVAE enables the agent to sample a broader range of sub-goals, fostering efficient exploration and ultimately enhancing task performance.

5 Discussion and Future Work

Through comprehensive quantitative analyses, we answer the defined research questions while delineating the advantages and limitations of the proposed model relative to the evaluated baselines. Our findings resonate with the hypothesis posited by Higgins et al. (2019), emphasizing a strong correlation between disentanglement and generative diversity. Notably, disentangled representations consistently showcase enhanced visual fidelity and diversity compared to the entangled ones. This correlation persists across all datasets rendered using disentangled representations. Intriguingly, traversal analyses of α-TCVAE, illustrated in Figures 1 and 16 in Appendix C, reveal that it is able to discover novel generative factors, such as object positioning and vertical perspectives, which are absent from the training dataset. We hypothesize that the CEB term is responsible for this phenomenon. Most existing models optimize only the information bottleneck, and while this can result in factorized representations, it does not directly optimize latent variable informativeness. Our proposed bound also includes a CEB term, and so maximizes the average informativeness as well, which may push otherwise noisy variables to learn new generative factors. Future research will delve deeper into comprehending this phenomenon and exploring its potential applications.

In accordance with the literature, the main limitation of α-TCVAE is that, akin to other disentangled VAEs, it is difficult to scale efficiently. This scaling challenge permeates the entire disentanglement paradigm. In high-dimensional spaces, not only do disentangled VAE-based models struggle to produce disentangled representations, but also the metrics used to measure disentanglement tend not to be useful. (e.g., DCI and SNC(Eastwood & Williams, 2018; Mahon et al., 2023)). On the other hand, disentangled representations have a number of desirable properties, as already showcased in the literature (Higgins et al., 2022). In particular, their impact is undeniable in the Ant Maze RL experiment from Figure 9. Reinforcing this observation, our correlation study underscores the relationship between disentanglement and diversity, leading to the following question: can we leverage diversity as a surrogate for measuring disentanglement in complex and high-dimensional scenarios? We leave the answer to this question as a future work.

6 Conclusion

We introduce α-TCVAE, a VAE optimized through a convex lower bound on the joint total correlation (TC) between the latent representation and the input data. This proposed bound naturally reduces to a convex combination of the known variational information bottleneck (VIB) (Alemi et al., 2017) and the conditional entropy bottleneck (CEB) (Fischer & Alemi, 2020). Moreover, it generalizes the widely adopted β-VAE bound. By maximizing disentanglement and average informativeness of the latent variables, our approach enhances both representational and generative capabilities. A comprehensive quantitative evaluation indicates that α-TCVAE consistently produces superior representations. This is evident from its performance across key downstream metrics: disentanglement (i.e., DCI and SNC), generative diversity (i.e., Vendi score), visual fidelity (i.e., FID), and its demonstrated downstream usefulness. In particular, our α-TCVAE showcases significant improvements on the MPI3D-Real dataset, the most realistic factorized dataset in our evaluation, and in a downstream reinforcement learning task. This highlights the strength of maximizing the average informativeness of latent variables, offering a pathway to address the inherent challenges of disentangled VAE-based models.

7 Ethic statement and reproducibility

To the best of the authors' knowledge, this study does not involve any ethical issues. The authors aim to maximize the reproducibility of the study. The codes of this project will be released in the camera-ready version. In the methods section, notions align with existing literature.

8 Acknowledgements

We thank Prof. Yoshua Bengio for the useful feedback provided along the project. We thank Mila - Quebec AI institute, TUDelft, and Compute Canada for providing all the resources to make the project possible.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *International Conference on Learning Representations*, 2017.

Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511804090.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

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

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint ArXiv:1804.03599*, 2018.

Jinkun Cao, Ruiqian Nai, Qing Yang, Jialei Huang, and Yang Gao. An empirical study on disentanglement of negative-free contrastive learning. *Advances in Neural Information Processing Systems*, 35:1210–1222, 2022.

Clément Chadebec, Louis Vincent, and Stéphanie Allassonnière. Pythae: Unifying generative autoencoders in python-a benchmarking use case. *Advances in Neural Information Processing Systems*, 35:21575–21589, 2022.

Ricky TQ Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018.

Wenxiao Chen, Haowen Xu, Zeyan Li, Dan Pei, Jie Chen, Honglin Qiao, Yang Feng, and Zhaogang Wang. Unsupervised anomaly detection for intricate KPIs via adversarial training of VAE. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 1891–1899. IEEE, 2019.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2020.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *International Conference on Learning Representations*, 2020.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Ian Fischer and Alexander A Alemi. CEB improves model robustness. *Entropy*, 22(10):1081, 2020.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint ArXiv:2210.02410*, 2022.

Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1157–1166. PMLR, 2019.

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

Anirudh Goyal, Aniket Didolkar, Nan Rosemary Ke, Charles Blundell, Philippe Beaudoin, Nicolas Heess, Michael C. Mozer, and Yoshua Bengio. Neural production systems. *Advances in Neural Information Processing Systems*, 34:25673–25687, 2021.

Danijar Hafner, Kuang-Huei Lee, Ian Fischer, and Pieter Abbeel. Deep hierarchical planning from pixels. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26091–26104. Curran Associates, Inc., 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. In *Theoretical Physics for Deep Learning Workshop, ICML*, 2019.

Irina Higgins, Sébastien Racanière, and Danilo Rezende. Symmetry-based representations for artificial and biological general intelligence. *Frontiers in Computational Neuroscience*, 16:836498, 2022.

HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Multi-view representation learning via total correlation objective. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 848–856. IEEE, 2019.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint ArXiv:1312.6114*, 2013.

Andreas Kirsch, Clare Lyle, and Yarin Gal. Unpacking information bottlenecks: Surrogate objectives for deep learning, 2021. URL https://openreview.net/forum?id=5rc0K0ezhqI.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–51, 2018.

Yang Li, Cheng Yu, Guangzhi Sun, Weiqin Zu, Zheng Tian, Ying Wen, Wei Pan, Chao Zhang, Jun Wang, Yang Yang, et al. Cross-utterance conditioned VAE for speech generation. *arXiv preprint ArXiv:2309.04156*, 2023.

Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Mixnmatch: multifactor disentanglement and encoding for conditional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8039–8048, 2020.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint ArXiv:1701.07274*, 2017.

Zeming Lin, Tom Sercu, Yann LeCun, and Alexander Rives. Deep generative models create new and diverse protein structures. In *Machine Learning for Structural Biology Workshop, NeurIPS*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, 32, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019b.

Louis Mahon, Lei Shah, and Thomas Lukasiewicz. Correcting flaws in common disentanglement metrics. *arXiv preprint ArXiv:2304.02335*, 2023.

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1429–1437, 2019.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pp. 4402–4412. PMLR, 2019.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670, 2023.

Pol Moreno, Christopher KI Williams, Charlie Nash, and Pushmeet Kohli. Overcoming occlusion with inverse graphics. In *European Conference on Computer Vision*, pp. 170–185. Springer, 2016.

Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-VAE. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10775–10784, 2021.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint ArXiv:1511.06434*, 2016.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-VAE-2. *Advances in Neural Information Processing Systems*, 32, 2019.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2019.

Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multimodal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint ArXiv:1703.00810*, 2017.

Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu. Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6699–6703. IEEE, 2020.

Xiaohong Sun, Jinan Gu, and Hongying Sun. Research progress of zero-shot learning. *Applied Intelligence*, 51(6):3600–3614, 2021.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 2001.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Ziyu Wang, Yiyi Zhang, Yixiao Zhang, Junyan Jiang, Ruihan Yang, Junbo Zhao, and Gus Xia. Pianotree VAE: Structured representation learning for polyphonic music. *arXiv preprint ArXiv:2008.07118*, 2020.

Zachary Wu, Kadina E. Johnston, Frances H. Arnold, and Kevin K. Yang. Protein sequence design with deep generative models. *Current Opinion in Chemical Biology*, 65:18–27, 2021.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76, 2020.

A Total Correlation Lower Bound Derivation

In this section we are going to derive the TC lower bound defined in equation 6. Since it is defined as a convex combination of marginal log-likelihood, VIB, and CEB terms, we are going to split the derivation into two subsections. First, we will derive a first TC bound that introduces the VIB term. Then, we will derive another TC bound, which explicitly shows the CEB term. Finally, we will define the TC bound shown in equation 6 as a convex combination of the two bounds.

A.1 TC Bound and the Variational Information Bottleneck

Unfortunately, direct optimization of mutual information terms is intractable Alemi et al. (2017). Therefore, we first need to find a lower bound of equation 4. Following the approach used in Hwang et al. (2021), we can expand it as:

$$\begin{aligned} TC_\theta(\boldsymbol{z},\boldsymbol{x}) &= \sum_{k=1}^{K} I_\theta(\boldsymbol{z}_k,\boldsymbol{x}) - I_\theta(\boldsymbol{z},\boldsymbol{x}), \qquad (7)\\ &= \sum_{k=1}^{K}\left[\mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z}_k)}\left[\log\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})}\right]\right] - \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}\left[\log\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z})}{p_D(\boldsymbol{x})}\right],\\ &= \sum_{k=1}^{K}\left[\mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z}_k)}\left[\log\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})}\frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}\right]\right] - \mathbb{E}_{q_\theta(\boldsymbol{z},\boldsymbol{x})}\left[\log\frac{q_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\theta(\boldsymbol{z})}\frac{r(\boldsymbol{z})}{r(\boldsymbol{z})}\right]. \end{aligned}$$

Let's expand these two terms:

$$\begin{aligned} &\mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z}_k)}\left[\log\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})}\frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}\right] = \int\int q_\theta(\boldsymbol{z}_k,\boldsymbol{x})\log\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})}\frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}d\boldsymbol{z}_k d\boldsymbol{x}, \qquad (8)\\ &= \int\int q_\theta(\boldsymbol{z}_k|\boldsymbol{x})p_D(\boldsymbol{x})\left(\log\left(\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}\right) + \log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k) - \log p_D(\boldsymbol{x})\right)d\boldsymbol{z}_k d\boldsymbol{x},\\ &= H(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z}_k)}[D_{KL}(q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)\|p_\phi(\boldsymbol{x}|\boldsymbol{z}_k))] + \mathbb{E}_{q_\theta(\boldsymbol{z}_k,\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)]. \end{aligned}$$

$$\begin{aligned} &\mathbb{E}_{q_\theta(\boldsymbol{z},\boldsymbol{x})}\left[\log\frac{q_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\theta(\boldsymbol{z})}\frac{r(\boldsymbol{z})}{r(\boldsymbol{z})}\right], \qquad (9)\\ &= \int q_\theta(\boldsymbol{x},\boldsymbol{z})\log\left(\frac{q_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\theta(\boldsymbol{z})}\frac{r(\boldsymbol{z})}{r(\boldsymbol{z})}\right)d\boldsymbol{z}d\boldsymbol{x},\\ &= \int q_\theta(\boldsymbol{z}|\boldsymbol{x})p_D(\boldsymbol{x})\left(\left(\log\frac{q_\theta(\boldsymbol{z}|\boldsymbol{x})}{r(\boldsymbol{z})}\right) + \log\left(\frac{r(\boldsymbol{z})}{q_\theta(\boldsymbol{z})}\right)\right)d\boldsymbol{z}d\boldsymbol{x},\\ &= \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))] - \mathbb{E}_{q_\theta(\boldsymbol{x}|\boldsymbol{z})}[D_{KL}(q_\theta(\boldsymbol{z})\|r(\boldsymbol{z}))]. \end{aligned}$$

As a result, we can write:

$$
\begin{aligned}
TC_{\theta}(\boldsymbol{z},\boldsymbol{x}) = \sum_{k=1}^{K} \big[H(\boldsymbol{x}) &+ \mathbb{E}_{q_\theta(\boldsymbol{z}_k)}[D_{KL}(q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)\|p_\phi(\boldsymbol{x}|\boldsymbol{z}_k))] + \mathbb{E}_{q_\theta(\boldsymbol{z}_k,\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)]\big]\,, \\
&- \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))] + \mathbb{E}_{q_\theta(\boldsymbol{x}|\boldsymbol{z})}[[D_{KL}(q_\theta(\boldsymbol{z})\|r(\boldsymbol{z}))], \qquad (10)\\
&\geq \sum_{k=1}^{K} \big[H(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z}_k,\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)\big] - \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))], \\
&= \sum_{k=1}^{K} \left[H(\boldsymbol{x}) + \int \left(\int q_\theta(\boldsymbol{z},\boldsymbol{x}) d\boldsymbol{z}_{\neq k}\right) \log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k) d\boldsymbol{z}_k d\boldsymbol{x}\right] - \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))], \\
&= \sum_{k=1}^{K} \big[H(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z},\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)\big] - \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))], \\
&= KH(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z},\boldsymbol{x})}[\log \prod_{k=1}^{K} p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)] - \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))], \\
&= KH(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z},\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}) + \log p_D(\boldsymbol{x})^{K-1}] - \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))], \\
&= \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})] - \underbrace{\mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))]}_{\text{VIB}} =: \mathcal{L}(\boldsymbol{z},\boldsymbol{x}).
\end{aligned}
$$

Maximizing $\mathcal{L}(\boldsymbol{z},\boldsymbol{x})$ not only maximizes the original objective $TC(\boldsymbol{z},\boldsymbol{x})$, but at the same time minimize the gap produced by upper bounding equation 10. As a result,

$$
\sum_{k=1}^{K} \big[\mathbb{E}_{q_\theta(\boldsymbol{z}_k)}[D_{KL}(q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)\|p_\phi(\boldsymbol{x}|\boldsymbol{z}_k))]\big] + \mathbb{E}_{q_\theta(\boldsymbol{x}|\boldsymbol{z})}[[D_{KL}(q_\theta(\boldsymbol{z})\|r(\boldsymbol{z}))], \qquad (11)
$$

will be minimized, leading to: $r(\boldsymbol{z}) \approx q_\theta(\boldsymbol{z})$ and $p_\phi(\boldsymbol{x}|\boldsymbol{z}_k) \approx q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)$.

Moreover, since $H(\boldsymbol{x})$ and $\log p_D(\boldsymbol{x})^{K-1}$ do not depend on θ, we can drop them from $\mathcal{L}(\boldsymbol{z},\boldsymbol{x})$. Finally, to avoid using a heavy notation, we will denote the VIB term as $D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))$, leading to the first TC bound which introduces the VIB term:

$$
TC_\theta(\boldsymbol{z},\boldsymbol{x}) \geq \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})] - \underbrace{D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|r(\boldsymbol{z}))}_{\text{VIB}}. \qquad (12)
$$

A.2 TC BOUND AND THE CONDITIONAL VARIATIONAL INFORMATION BOTTLENECK

Expanding Eq. equation 2, we can reformulate $TC(\boldsymbol{z},\boldsymbol{x})$ as follow:

$$
\begin{aligned}
TC_\phi(\boldsymbol{z},\boldsymbol{x}) &= \sum_{k=1}^{K} I_\phi(\boldsymbol{z}_k,\boldsymbol{x}) - I_\phi(\boldsymbol{z},\boldsymbol{x}), \qquad (13)\\
&= \sum_{k=1}^{K} \left[\frac{K-1}{K} I_\phi(\boldsymbol{z}_k,\boldsymbol{x}) + \frac{1}{K} I_\phi(\boldsymbol{z}_k,\boldsymbol{x}) - \frac{1}{K} I_\phi(\boldsymbol{z},\boldsymbol{x})\right], \\
&= \sum_{k=1}^{K} \left[\frac{K-1}{K} I_\phi(\boldsymbol{z}_k,\boldsymbol{x}) + \frac{1}{K}\left(I_\phi(\boldsymbol{z}_k,\boldsymbol{x}) - I_\phi(\boldsymbol{z},\boldsymbol{x})\right)\right]. \qquad (14)
\end{aligned}
$$

Interestingly, can write the last term of Eq. equation 14 as:

$$\begin{aligned}
I_\phi(\boldsymbol{z}_k, \boldsymbol{x}) - I_\phi(\boldsymbol{z}, \boldsymbol{x}) &= \mathbb{E}_{p_\phi(\boldsymbol{x}, \boldsymbol{z}_k)}\left[\log \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})}\right] - \mathbb{E}_{p_\phi(\boldsymbol{x}, \boldsymbol{z})}\left[\log \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z})}{p_D(\boldsymbol{x})}\right], \qquad (15)\\
&= \int \left(\int p_\phi(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}_{\neq \boldsymbol{z}_k}\right) \log \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})} d\boldsymbol{z}_k d\boldsymbol{x},\\
&- \int p_\phi(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) \log \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z})}{p_D(\boldsymbol{x})} d\boldsymbol{z} d\boldsymbol{x},\\
&= \int p_\phi(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) \log \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p(\boldsymbol{x}|\boldsymbol{z})} d\boldsymbol{z} d\boldsymbol{x},\\
&= -\int p_\phi(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}|\boldsymbol{z})}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)} d\boldsymbol{z} d\boldsymbol{x},\\
&= -\int p_\phi(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}|\boldsymbol{z}_{\neq k}, \boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)} d\boldsymbol{z} d\boldsymbol{x},\\
&= -I_\phi(\boldsymbol{z}_{\neq k}, \boldsymbol{x}|\boldsymbol{z}_k).
\end{aligned}$$

We can now write equation 5:

$$TC_\theta(\boldsymbol{z}, \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\left[(K-1) I_\theta(\boldsymbol{z}_k, \boldsymbol{x}) - I_\theta(\boldsymbol{z}_{\neq k}, \boldsymbol{x}|\boldsymbol{z}_k)\right].$$

Interestingly, the second IB term in Eq. (8) can now be expressed as multiple conditional MIs between the observation and $K-1$ other latent variables given the k-th latent representation variable, penalizing the extra information of the observation not inferable from the given latent representation variable. Moreover, we can further expand the TC as:

$$\begin{aligned}
TC_\theta(\boldsymbol{z}, \boldsymbol{x}) &= \frac{1}{K}\sum_{k=1}^{K}\left[(K-1) I_\theta(\boldsymbol{z}_k, \boldsymbol{x}) - I_\theta(\boldsymbol{z}_{\neq k}, \boldsymbol{x}|\boldsymbol{z}_k)\right], \qquad (16)\\
&= \frac{1}{K}\sum_{k=1}^{K}\Bigg[(K-1)\left[\mathbb{E}_{q_\theta(\boldsymbol{z}_k, \boldsymbol{x})}\left[\log \frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})} \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}\right]\right]\\
&- \mathbb{E}_{q_\theta(\boldsymbol{x}, \boldsymbol{z})}\left[\log \frac{q_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\theta(\boldsymbol{z}_k|\boldsymbol{x})} \frac{r_p(\boldsymbol{z}|\boldsymbol{x})}{r_p(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_\theta(\boldsymbol{x}, \boldsymbol{z})}\left[\log q_\theta(\boldsymbol{z}_{\neq k})\right]\Bigg],\\
&= \frac{1}{K}\sum_{k=1}^{K}\Bigg[(K-1)\left[\mathbb{E}_{q_\theta(\boldsymbol{z}_k, \boldsymbol{x})}\left[\log \frac{q_\theta(\boldsymbol{x}|\boldsymbol{z}_k)}{p_D(\boldsymbol{x})} \frac{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}{p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)}\right]\right]\\
&- \mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x}))] - \mathbb{E}_{q_\theta(\boldsymbol{x}, \boldsymbol{z})}\left[\log \frac{r_p(\boldsymbol{z}_k|\boldsymbol{x}) r_p(\boldsymbol{z}_{\neq k}|\boldsymbol{x})}{q_\theta(\boldsymbol{z}_k|\boldsymbol{x}) q_\theta(\boldsymbol{z}_{\neq k})}\right]\Bigg],\\
&= \frac{K-1}{K}\sum_{k=1}^{K}\left[H(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z}_k, \boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)]\right] - \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x}))]\right]\\
&+ \frac{K-1}{K}\sum_{k=1}^{K}\left[\mathbb{E}_{q_\theta(\boldsymbol{z}_k)}[D_{KL}(q_\theta(\boldsymbol{x}|\boldsymbol{z}_k) \| p_\phi(\boldsymbol{x}|\boldsymbol{z}_k))]\right] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{q_\theta(\boldsymbol{z}_{\neq k}, \boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}_k|\boldsymbol{x}) \| r_p(\boldsymbol{z}_k|\boldsymbol{x})] \qquad (17)\\
&+ \int D_{KL}(q_\theta(\boldsymbol{z}_{\neq k}) \| r_p(\boldsymbol{z}_{\neq k}|\boldsymbol{x})) d\boldsymbol{x},\\
&\geq \frac{K-1}{K}\sum_{k=1}^{K}\left[H(\boldsymbol{x}) + \mathbb{E}_{q_\theta(\boldsymbol{z}_k|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)]\right] - \underbrace{\frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}_{p_D(\boldsymbol{x})}[D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x}))]\right]}_{\text{CEB}}. \qquad (18)
\end{aligned}$$

Maximizing Eq. 18 not only maximizes the original objective $TC(\boldsymbol{z}, \boldsymbol{x})$ but at the same time minimizes the gap produced by upper bounding Eq. equation 17, leading to: $r_p(\boldsymbol{z}_k|\boldsymbol{x}) \approx q_\theta(\boldsymbol{z}_k|\boldsymbol{x})$, $q_\theta(\boldsymbol{z}_{\neq k}) \approx r_p(\boldsymbol{z}_{\neq k}|\boldsymbol{x})$ and $q_\theta(\boldsymbol{x}|\boldsymbol{z}_k) \approx p_\phi(\boldsymbol{x}|\boldsymbol{z}_k)$. Moreover, since $H(\boldsymbol{x})$ does not depend on θ, we can drop it from Eq. equation 18. Finally, to avoid using a heavy notation, we will denote the CEB term as $D_{KL}(q_\theta(\boldsymbol{z}_k|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x}))$, leading to the second TC bound which introduces the CEB term:

$$TC_\theta(\boldsymbol{z}, \boldsymbol{x}) \geq \frac{K-1}{K} \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}|\boldsymbol{z})] - \underbrace{D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x}))}_{\text{CEB}} . \tag{19}$$

A.3 Final TC bound

In order to obtain the final expression of the derived TC bound, we can compute a convex combination of the two bounds defined in Eq. equation 12 and equation 19.

$$TC(\boldsymbol{z}, \boldsymbol{x}) = (1-\alpha) \left(\sum_{k=1}^{K} I_\theta(\boldsymbol{z}_k, \boldsymbol{x}) - I_\theta(\boldsymbol{z}, \boldsymbol{x}) \right) \tag{20}$$

$$+ \alpha \left(\sum_{k=1}^{K} \left[\frac{K-1}{K} I_\theta(\boldsymbol{z}_k, \boldsymbol{x}) + \frac{1}{K} I_\theta(\boldsymbol{z}_k, \boldsymbol{x}) - \frac{1}{K} I_\theta(\boldsymbol{z}, \boldsymbol{x}) \right] \right), \tag{21}$$

$$= \frac{K(1-\alpha) + \alpha(K-1)}{K} \sum_{k=1}^{K} I_\theta(\boldsymbol{z}_k, \boldsymbol{x}) - \frac{\alpha}{K} \sum_{k=1}^{K} (I_\theta(\boldsymbol{z}, \boldsymbol{x}) - I_\theta(\boldsymbol{z}_k, \boldsymbol{x})) - (1-\alpha) I_\theta(\boldsymbol{z}, \boldsymbol{x}),$$

$$\geq \frac{K-\alpha}{K} \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})} [\log p_\phi(\boldsymbol{x}|\boldsymbol{z})] - \alpha D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x})) - (1-\alpha) D_{KL}(p_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r(\boldsymbol{z})),$$

$$= \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})} [\log p_\phi(\boldsymbol{x}|\boldsymbol{z})] - \underbrace{\frac{K\alpha}{K-\alpha} D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r_p(\boldsymbol{z}|\boldsymbol{x}))}_{\text{CEB}} - \underbrace{\frac{(1-\alpha)}{(1-\frac{\alpha}{K})} D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| r(\boldsymbol{z}))}_{\text{VIB}} .$$

where α is a hyperparameter that balances the effects of VIB and CEB terms. Table A.3 illustrates the lower bounds defined for β-VAE Higgins et al. (2016), FactorVAE Kim & Mnih (2018), HFS Roth et al. (2023) and β-TCVAE Chen et al. (2018) comparing them to the derived TC bound. We can see that the three bounds present a similar structure, presenting a marginal log-likelihood term and either one or two KL regularizers that impose some kind of information bottleneck.

Model	Lower Bound
β-VAE	$\mathbb{E}_{q_\theta(\boldsymbol{z}\mid\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}\mid\boldsymbol{z})] - \beta D_{KL}(q_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Vert p(\boldsymbol{z}))$
FactorVAE	$\mathbb{E}_{q_\theta(\boldsymbol{z}\mid\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}\mid\boldsymbol{z})] - \beta D_{KL}(q_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Vert p(\boldsymbol{z})) - \gamma D_{KL}(q_\theta(\boldsymbol{z})\Vert \prod_{k=1}^{K} q_\phi(\boldsymbol{z}_k))$
β-TCVAE	$\mathbb{E}_{q(z\mid n)p(n)}[\log p(n\mid z) - \alpha I_q(z;n) - \beta D_{KL}(q(z)\Vert \prod_j q(\boldsymbol{z_j})) - \gamma \sum_j D_{KL}(q(z_j)\Vert p(z_j))$
HFS	$\mathbb{E}_{q_\theta(\boldsymbol{z}\mid\boldsymbol{x})}[\log p_\phi(\boldsymbol{x}\mid\boldsymbol{z})] - \gamma[\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \max_{z \in Z_{:,1} \times Z_{:,2} \times \cdots \times Z_{:,K}} \min_{z' \in Z_{:,(i,j)}} d(z, z')]$
α-TCVAE	$\mathbb{E}_{q_\theta(\boldsymbol{z}\mid\boldsymbol{x})}\left[\log p_\phi(\boldsymbol{x}\mid\boldsymbol{z})\right] - \frac{K\alpha}{K-\alpha} D_{KL}(q_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Vert r_p(\boldsymbol{z}\mid\boldsymbol{x})) - \frac{(1-\alpha)}{(1-\frac{\alpha}{K})} D_{KL}(q_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Vert r(\boldsymbol{z}))$

Table 1: This table compares the lower bound objective functions of β-VAE, β-TCVAE, FactorVAE and HFS-VAE. The lower bound objective function of β-VAE is composed of the expected log-likelihood of the data given the latent variables and the KL divergence between the approximate posterior and the prior of the latent variables (i.e., VIB term). The FactorVAE model further adds a KL divergence term between the approximate posterior and the factorized prior of the latent variables, which approximates the total correlation of the latent variables, and HFS-VAE further adds a Monte-Carlo approximation of Hausdorff distance. α-TCVAE, on the other hand, uses a convex combination of VIB term and KL divergence between the approximate posterior and the prior of the latent variables conditioned on the k-th latent variable (i.e., CEB term). K represents the dimensionality of the latent variables, while β, γ and α are hyperparameters of the models.

B Architectures and Hyperparameters Details

The hyperparameters used for the different experiments are shown in Table 2.

Table 2: Comparison of the different hyperparameters used across the Datasets

Dataset	β	γ	α	latent dim K	Training Epochs
Teapots	2	10	0.25	10	50
3DShapes	3	10	0.25	10	50
Cars3D	4	10	0.25	10	50
MPI3D-Real	5	10	0.25	10	50
Celeba	5	10	0.25	48	50

All encoder, decoder and discriminator architectures are taken from Roth et al. (2023).

C Further Details on Datasets and Metrics

C.1 Datasets

We test on five datasets. **Teapots** (Moreno et al., 2016) contains $200,000$ images of size 64×64. Each image features a rendered, camera-centered teapot with 5 uniformly distributed generative factors of variation: azimuth and elevation (sampled between 0 and 2π), along with three RGB colour channels (each sampled between 0 and 1). **3DShapes** (Burgess & Kim, 2018) consists of $480,000$ images of size 64×64. Every image displays a rendered, camera-centered object with 6 uniformly distributed generative factors of variation: shape (sampled from [cylinder, tube, sphere, cube]), object colour, object hue, floor colour, wall colour, and horizontal orientation, all determined using linearly spaced values. **MPI3D-Real** (Gondal et al., 2019) comprises $103,680$ images of size 64×64. Each image captures objects at a robot arm's end, characterized by 6 factors: object colour, size, shape, camera height, azimuth, and robot arm altitude. **Cars3D** (Reed et al., 2015) is made up of $16,185$ images of size 64×64. Each image portrays a rendered, camera-centered car, categorized by 3 factors: car-type, elevation, and azimuth. **CelebA** (Liu et al., 2015) encompasses over $200,000$ images of size 64×64. Every image presents a celebrity, highlighted by a broad range of poses, facial expressions, and lighting conditions, which sum up to 40 different factors. Every model is trained using a subset containing the 80% of the selected dataset images in a fully unsupervised way. The models are evaluated on the remaining images using the following downstream scores. While

CelebA is the most complex dataset, MPI3D-Real is the most realistic among the ones usually used in the disentanglement community.

C.2 METRICS

When using the **FID** score to assess image quality, we compare the distribution of generated images to that of the real images. Specifically, FID (Heusel et al., 2017) measures the distance between two distributions of images, and we apply it to measure the distance between the generated images and the real ones. A lower distance is better, indicating that the generated images belong to the distribution of ground truth images.

The **Vendi** score (Friedman & Dieng, 2022), which we use to measure the diversity of the generated images, is computed with respect to a similarity measure. Specifically, it is calculated as the exponential of the entropy of the eigenvalues of the similarity matrix, i.e. the matrix whose (i, j)th entry is the similarity between the ith and jth data points. It can be interpreted as the effective number of distinct elements in the set.

To assess the quality of encoded latent representations, we use DCI, SNC/NK (Mahon et al., 2023) and the unfairness measure of Locatello et al. (2019a).

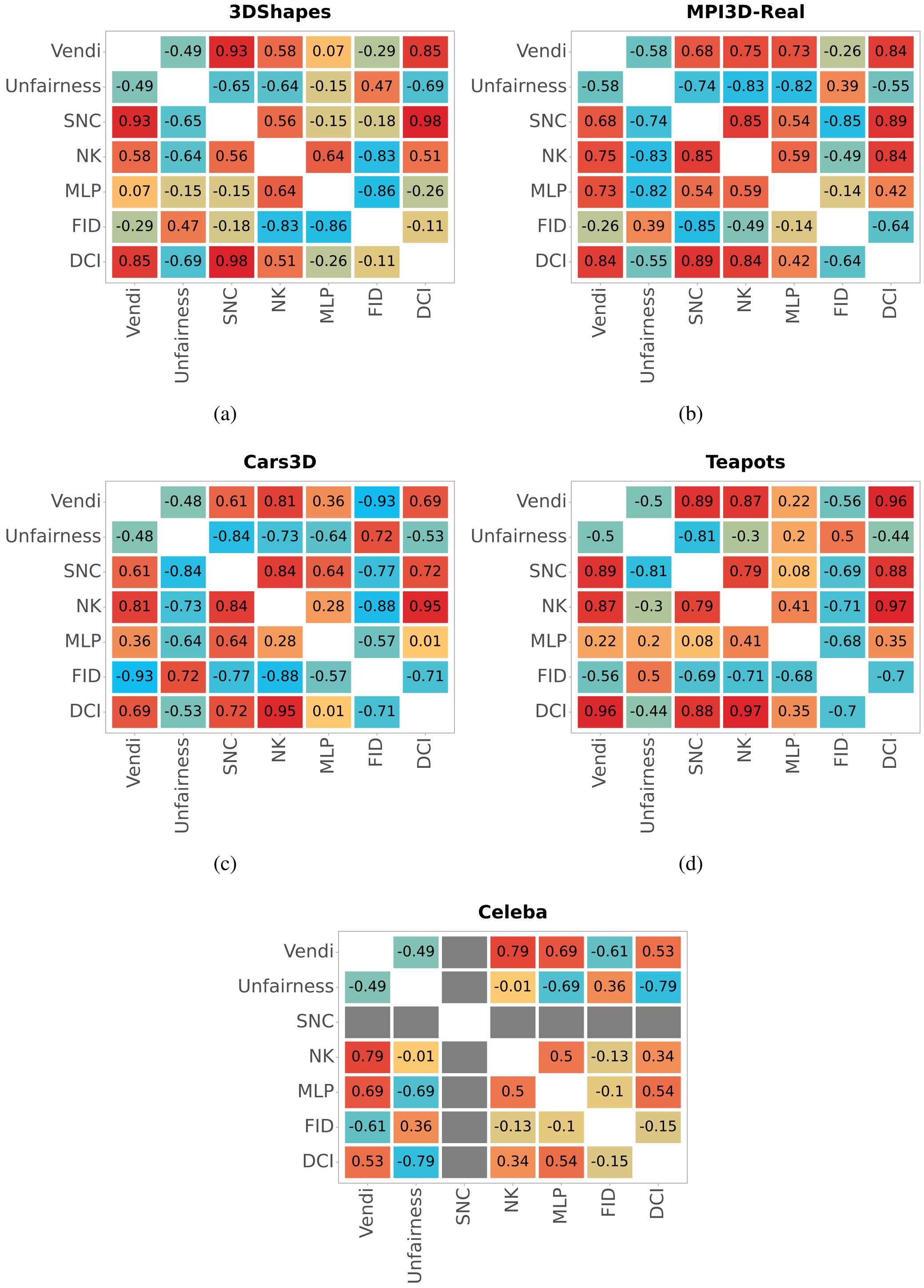
DCI, the first disentanglement metric we compute, first trains a regressor to predict the generative factors from the latent representation, and from this regressor extracts a matrix of feature importances, where the (i, j)th entry is the import of the ith latent dimension to predicting the jth generative factor. It then takes (a normalized version of) the entropy of rows and columns to compute 'disentanglement' and 'completeness', respectively. The accuracy of the regressor is taken as the 'informativeness' score. The average of these three scores, across all factors and neurons, is the final DCI score.

SNC/NK, the second disentanglement metric we compute, works by first aligning neurons to latent factors using the Kuhn-Munkres algorithm to enforce uniqueness. Then each aligned neuron is used as a classifier for the corresponding factor, by binning its values. A higher accuracy of this single-neuron classifier (SNC) is better, indicating that the factor is well-represented by a single unique neuron. Neuron knockout (NK) is calculated as the difference between an MLP classifier that predicts the generative factor from all neurons, and one that predicts using all neurons but the one that factor was aligned to. A high NK is also better, indicating that no neurons, other than the one it was aligned to, contain information about the given factor. SNC/NK measures a slightly different and stronger notion of disentanglement than DCI, as it explicitly assumes an inductive bias that enforces each factor to be represented by a single latent variable.

MIG, is a disentanglement metric that quantifies the degree of separation between the latent variables and the generative factors in a dataset. It calculates the mutual information between each latent variable and each generative factor, identifying the variable that shares the most information with each factor. The gap, or difference, in mutual information between the top two variables for each generative factor is then computed. A larger gap indicates that one latent variable is significantly more informative about a generative factor than the others, signifying a higher degree of disentanglement. This metric is particularly useful in scenarios where a clear and distinct representation of generative factors is desired in the latent space. MIG thus complements DCI and SNC/NK by providing a measure of how well-separated the representations of different generative factors are within the model's latent space.

D EXTENDED RESULTS

Here, we present further results, in addition to those from Section 4. Figure 10 extended Fig. 7, reporting the correlations also with SNC, NK and the attribute classification accuracy as shown in Figure 8. Unsurprisingly, there is a strong correlations between the three metrics designed to measure disentanglement: DCI, SNC and NK. This, to some extent, verifies the reliability of these different disentanglement metrics. SNC and NK also correlate strongly with Vendi, as DCI does. This further supports the finding in our paper of a relationship between disentanglement and diversity.



3DShapes

	Vendi	Unfairness	SNC	NK	MLP	FID	DCI
Vendi		-0.49	0.93	0.58	0.07	-0.29	0.85
Unfairness	-0.49		-0.65	-0.64	-0.15	0.47	-0.69
SNC	0.93	-0.65		0.56	-0.15	-0.18	0.98
NK	0.58	-0.64	0.56		0.64	-0.83	0.51
MLP	0.07	-0.15	-0.15	0.64		-0.86	-0.26
FID	-0.29	0.47	-0.18	-0.83	-0.86		-0.11
DCI	0.85	-0.69	0.98	0.51	-0.26	-0.11	

(a)

MPI3D-Real

	Vendi	Unfairness	SNC	NK	MLP	FID	DCI
Vendi		-0.58	0.68	0.75	0.73	-0.26	0.84
Unfairness	-0.58		-0.74	-0.83	-0.82	0.39	-0.55
SNC	0.68	-0.74		0.85	0.54	-0.85	0.89
NK	0.75	-0.83	0.85		0.59	-0.49	0.84
MLP	0.73	-0.82	0.54	0.59		-0.14	0.42
FID	-0.26	0.39	-0.85	-0.49	-0.14		-0.64
DCI	0.84	-0.55	0.89	0.84	0.42	-0.64	

(b)

Cars3D

	Vendi	Unfairness	SNC	NK	MLP	FID	DCI
Vendi		-0.48	0.61	0.81	0.36	-0.93	0.69
Unfairness	-0.48		-0.84	-0.73	-0.64	0.72	-0.53
SNC	0.61	-0.84		0.84	0.64	-0.77	0.72
NK	0.81	-0.73	0.84		0.28	-0.88	0.95
MLP	0.36	-0.64	0.64	0.28		-0.57	0.01
FID	-0.93	0.72	-0.77	-0.88	-0.57		-0.71
DCI	0.69	-0.53	0.72	0.95	0.01	-0.71	

(c)

Teapots

	Vendi	Unfairness	SNC	NK	MLP	FID	DCI
Vendi		-0.5	0.89	0.87	0.22	-0.56	0.96
Unfairness	-0.5		-0.81	-0.3	0.2	0.5	-0.44
SNC	0.89	-0.81		0.79	0.08	-0.69	0.88
NK	0.87	-0.3	0.79		0.41	-0.71	0.97
MLP	0.22	0.2	0.08	0.41		-0.68	0.35
FID	-0.56	0.5	-0.69	-0.71	-0.68		-0.7
DCI	0.96	-0.44	0.88	0.97	0.35	-0.7	

(d)

Celeba

	Vendi	Unfairness	SNC	NK	MLP	FID	DCI
Vendi		-0.49	■	0.79	0.69	-0.61	0.53
Unfairness	-0.49		■	-0.01	-0.69	0.36	-0.79
SNC	■	■		■	■	■	■
NK	0.79	-0.01	■		0.5	-0.13	0.34
MLP	0.69	-0.69	■	0.5		-0.1	0.54
FID	-0.61	0.36	■	-0.13	-0.1		-0.15
DCI	0.53	-0.79	■	0.34	0.54	-0.15	

(e)

Figure 10: Correlations between all metrics we measure, both for the generated images and the representations.

Figure 11 shows the results for neuron knockout (NK), the second metric introduced by Mahon et al. (2023) alongside SNC, which is shown in Figure 5. Similar to SNC, the NK score for α-TCVAE

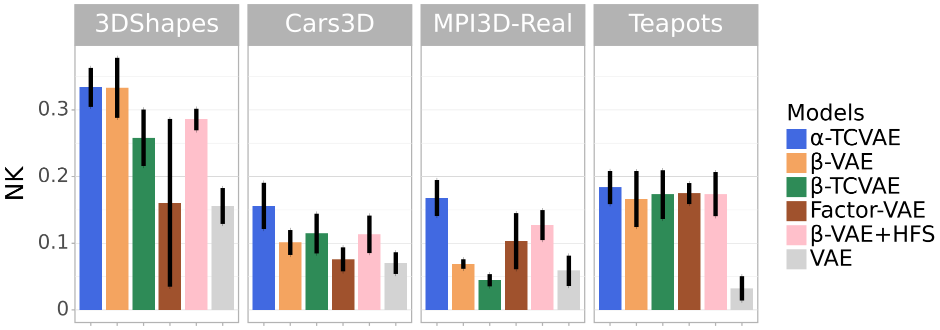


Figure 11: Comparison of the neuron-knockout score of α-TCVAE with that of baseline models. As with other metrics presented in the main paper, the improvement of α-TCVAE is minor on 3DShapes and Teapots, but more substantial on Cars3D and MPI3D-Real.

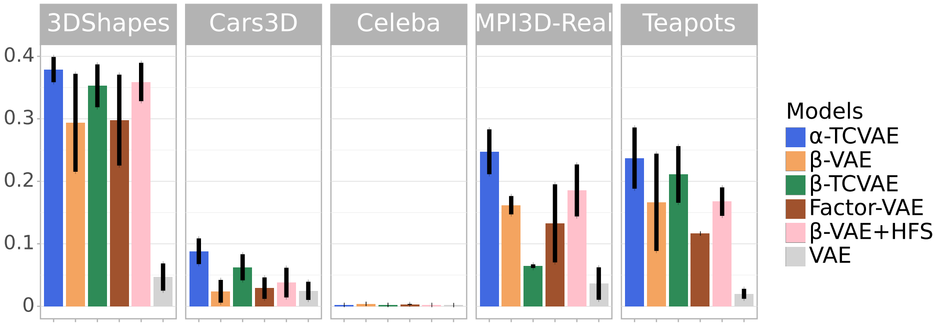


Figure 12: Comparison of the MIG score of α-TCVAE with that of baseline models. As with other metrics presented in the main paper, the improvement of α-TCVAE is minor on 3DShapes and Teapots, but more substantial on Cars3D and MPI3D-Real.

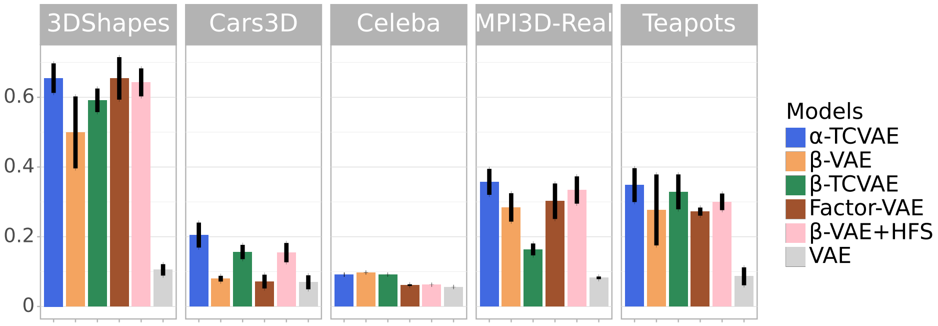


Figure 13: Comparison of the DCI-C completeness score of α-TCVAE with that of baseline models. As with other metrics presented in the main paper, the performance of α-TCVAE is comparable on 3DShapes, CelebA and Teapots, and better on Cars3D and MPI3D-Real.

is higher than that for baseline VAE models and, while the errorbars often overlap, the superiority of α-TCVAE is consistent across all five datasets and is most substantial on MPI3D-Real. Figure 12 shows the results for mutual information gap (MIG), which follows the same trend of NK, SNC and DCI scores. Figures 13, 14 and 15 present the results of Completeness, Disentanglement and

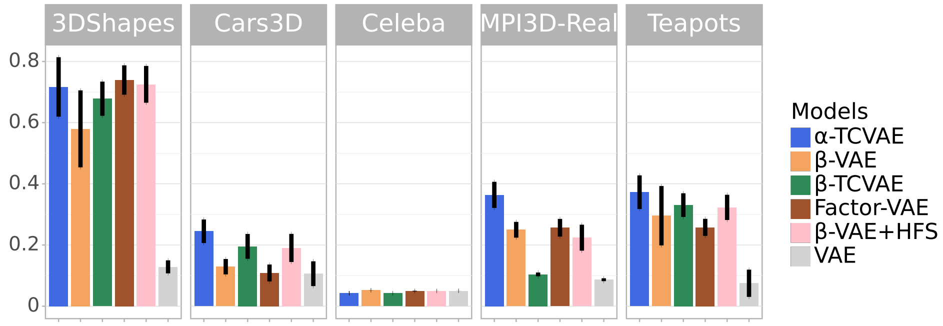


Figure 14: Comparison of the DCI-D disentanglement score of α-TCVAE with that of baseline models. As with other metrics presented in the main paper, the performance of α-TCVAE is comparable on 3DShapes, CelebA and Teapots, and better on Cars3D and MPI3D-Real.

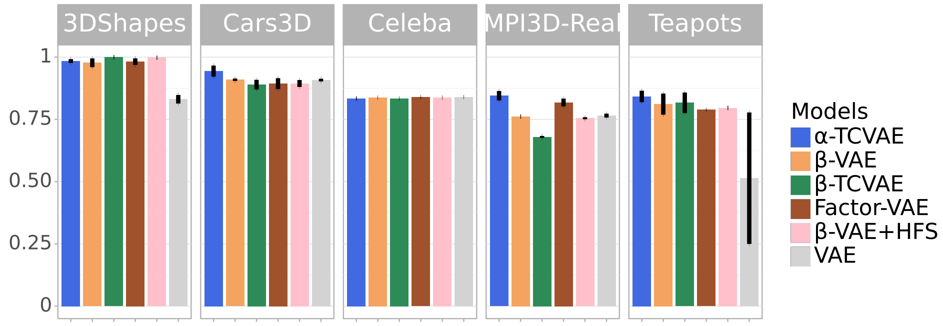
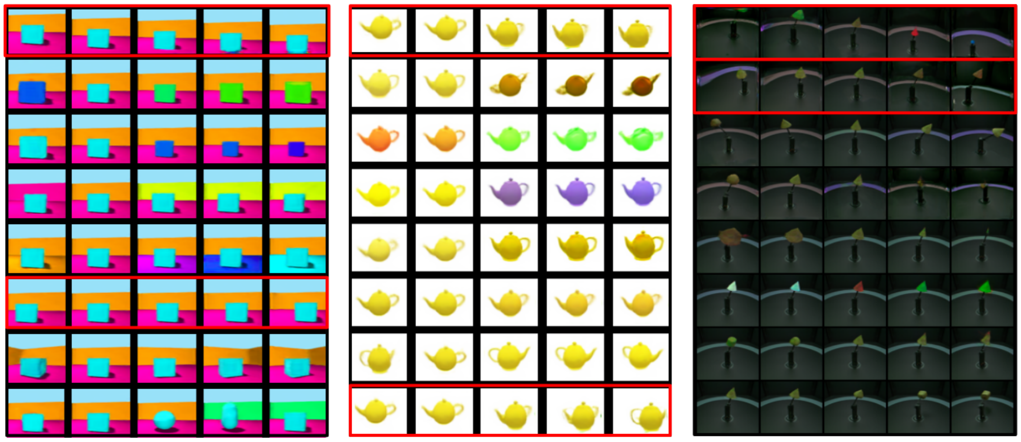


Figure 15: Comparison of the DCI-I informativeness score of α-TCVAE with that of baseline models. As with other metrics presented in the main paper, the performance of α-TCVAE is comparable on 3DShapes, CelebA and Teapots, and better on Cars3D and MPI3D-Real.

Informativeness metrics (DCI-C, DCI-D and DCI-I, respectively). The final DCI scores shown in fig. 4 is computed as geomtric mean of the three scores.

E Discovering novel factor of variations

Figure 16 presents α-TCVAE traversals across 3DShapes, Teapots and MPI3D-Real datasets. The red boxes indicate the discovered novel generative factors, that are not present within the train dataset, namely object position and vertical camera perspective. While we do not have a comprehensive explanation of why such an intriguing phenomenon is shown, we believe that the intuition behind can be explained considering the effects of VIB and CEB terms in the defined bound. Indeed, while VIB pushes individual latent variables to represent different generative factors, CEB pushes them to be informative. As a result, the otherwise noisy dimensions, are pushed to be informative (i.e., CEB) and to represent a distinct generative factor (i.e., VIB), resulting in the discovery of novel generative factors.



(a) 3DShapes Traversals (b) Teapots Traversals (c) MPI3D-Real Traversals

Figure 16: α-TCVAE generated latent traversals the 3DShapes, Teapots and MPI3D-Real datasets. The generated latent traversals reveal that α-TCVAE can learn and represent generative factors that are not present in the ground-truth dataset, namely vertical perspective and object position. The discovered generative factors are indicated with a red box.

F Relationship between CEB and Diversity

F.1 Fisher's Definition of Conditional Entropy Bottleneck

Fisher's approach to the Conditional Entropy Bottleneck Fischer & Alemi (2020) is an extension of the Information Bottleneck (IB) principle Alemi et al. (2017), aimed at finding an optimally compressed representation of a variable X that remains highly informative about another variable Y, under the influence of a conditioning variable Z. The CEB objective, according to Fisher, is formalized as a trade-off between two competing conditional mutual information terms:

$$\min_{p(z|x)} [I(X;Z|C) - \beta I(Y;Z|C)]$$

Here, $I(X;Z|C)$ quantifies the amount of information that the representation Z shares with X, conditioned on C. Simultaneously, $I(Y;Z|C)$ measures how much information Z retains about Y, also under the condition of C. The parameter β serves as a crucial tuning parameter, balancing these two aspects.

F.2 Adapting CEB to VAEs without Conditioning Variables

In the realm of Variational Autoencoders, where the training strategy is to reconstruct the input data X using a latent representation Z without any external conditioning C, the CEB framework undergoes a significant simplification. Given that $X = Y$ in a typical VAE setup, the CEB objective reduces to a form where the focus shifts to optimizing the mutual information between X and its latent representation Z:

$$\min_{p(z|x)} [(1 - \beta) I(X; Z)]$$

This objective can be further broken down as $(1 - \beta)(H(X) - H(X|Z))$, where $H(X)$ represents the entropy of the input data, and $H(X|Z)$ is the conditional entropy of the input given its latent representation. This formulation underscores the trade-off between compressing the input data in the latent space and retaining essential information for accurate reconstruction.

F.3 Incorporating Diversity into the CEB Objective

Following Friedman & Dieng (2022), Diversity can be quantitatively expressed as the exponential of the entropy of the latent space distribution $q(Z|X)$:

$$\text{Diversity} = \exp(H(Z|X))$$

To understand how the CEB framework relates to this notion of diversity, we utilize the entropy chain rule $\mathrm{H}(Y|X) = \mathrm{H}(X, Y) - \mathrm{H}(X)$, which allows to decompose $H(X|Z)$ in terms of the joint entropy $H(X, Z)$ and the conditional entropy $H(Z)$. Consequently, the CEB objective evolves into a more comprehensive form that explicitly accounts for the diversity of the latent space:

$$\min_{q(z|x)} [(1 - \beta)(H(X) - H(X, Z) + H(Z))]$$

$$\min_{q(z|x)} [(1 - \beta)(-H(Z|X) + H(Z))]$$

The latter one, makes clear the connection between the CEB term and Diversity as defined in Friedman & Dieng (2022). Indeed, we can see that when minimizing the CEB term the Diversity term is maximized.

G Disentanglement and Variational Information Bottleneck

Disentanglement in VAEs, following Higgings' β-VAE framework, seeks to learn representations where individual latent variables capture distinct, independent factors of variation in the data. This is achieved by modifying the traditional VAE objective to apply a stronger constraint on the latent space information bottleneck, controlled by a hyperparameter β. The β-VAE, introduced by Higgins et al. (2016), represents a seminal approach to disentanglement, promoting the learning of factorized and interpretable latent representations.

On a related front, the Variational Information Bottleneck (VIB) method, formulated by Alemi et al. (2017), extends the Information Bottleneck principle to deep learning. The VIB approach seeks to find an optimal trade-off between the compression of input data and the preservation of relevant information for prediction tasks. By employing a variational approximation, VIB efficiently learns compressed representations that are predictive of desired outcomes. Interestingly, Alemi formulates a VIB objective that is equivalent to Higgins' β-VAE one. Such result makes evident how imposing a higher information bottleneck leads to higher disentanglement.

H Sensitivity Analysis of α

In this section we present a sensitivity analysis of how α affects Vendi and FID results across the considered datasets. To be consistent and analyze how Alpha influences disentanglement scores, we also report a sensitivity analysis of the DCI metric and a correlation study showing how alpha is statistically correlated with FID, Vendi, and DCI metrics.

H.1 Diversity and Visual Fidelity Sensitivity Analysis with respect to α

To analyse how α influences the presented results, we performed an evaluation of FID, Vendi and DCI using $\alpha \in [0.00, 0.25, 0.50, 0.75, 1.00]$, where for $\alpha = 0.00$ we obtain β-VAE model, while for $\alpha = 0.25$ we get the results presented in the main paper. Figures 17 and 18 show that, when $\alpha \in [0.25, 0.50]$ α-TCVAE presents the highest diversity scores, while keeping a FID score comperable to β-VAE.

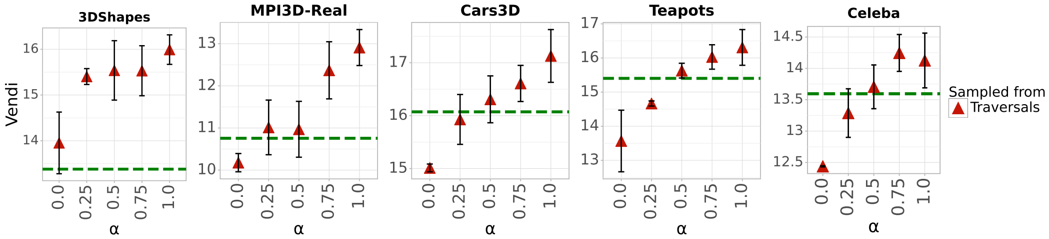


Figure 17: Sensitivity Analysis of the Diversity of generated images with respect to α. Only one sampling strategy is considered: sampled from traversals. The green dashed line represents ground truth dataset diversity. It can be seen that the higher alpha the higher will be the Vendi Score.

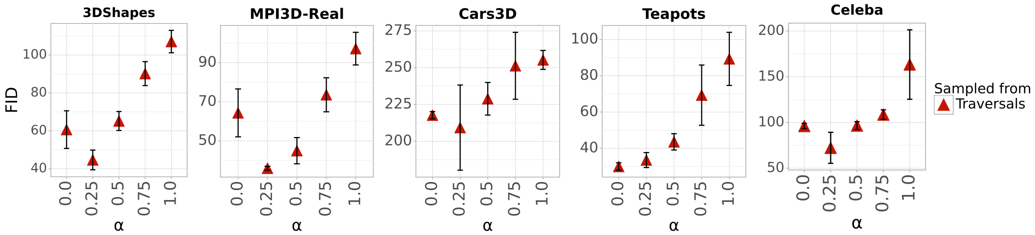


Figure 18: Sensitivity Analysis of the Faithfulness of generated images to the data distribution, as measured by FID score, with respect to α. Only one sampling strategy is considered: sampled from traversals. It can be seen that for $\alpha \in [0.25, 0.50]$ the model presents higher visual fidelity.

Interestingly, the two sensitivity analyses show two main trends:

- Diversity increases when using higher values of Alpha.
- FID score improves when using smaller values of Alpha.

Indeed, when using higher values of α , we increase the contribution of the CEB term in equation 6, which enhances diversity at the cost of visual fidelity. As a result, the higher the value of α, the more diverse the generated batch of images, and the lower will be the generation quality. However, it can be noticed that when using values of α between 0.25 and 0.50, we get a set of generated images that are more diverse and still have a better or comparable visual fidelity than β-VAE (i.e., α=0).

H.2 Disentanglement Sensitivity Analysis with respect to α

Here, we present a sensitivity analysis of the DCI metric. Figure 19 shows that the interval [0.25-0.50] presents higher values of disentanglement, following Diversity and Visual Fidelity analyses that show the best results in the same range. Such a trend can be explained by considering that α weights the contributions of VIB and CEB terms. While the CEB term enhances diversity, the VIB term encourages disentanglement. As a result, we can see that DCI scores decrease when α gets closer to 1. Interestingly, when α is in [0.25,0.50], the combination of CEB and VIB terms produces a better bound for the Total Correlation objective than when using , which results in higher DCI scores.

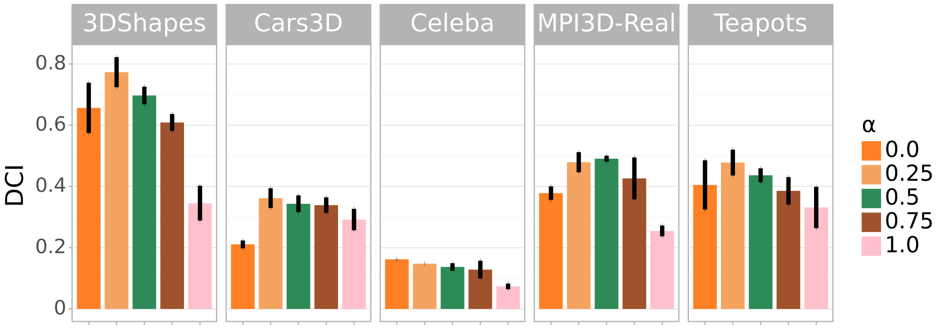


Figure 19: DCI scores sensitivity analysis with respect to α. On average when $\alpha \in [0.25, 0.50]$ α-TCVAE presents the best DCI scores.

H.3 Correlation Study: How is α correlated with Vendi, FID, and DCI metrics?

Here, we present correlation matrices for all the considered datasets. We computed them using the models trained for the alpha sensitivity analyses. The correlation matrices in fig. 20 confirm the trends observed in the other sensitivity analyses (i.e., Vendi, FID, and DCI). Indeed, α has a strong positive correlation with both FID and Vendi, showing that when α increases, diversity increases and FID deteriorates. On the other hand, α has a strong negative correlation with DCI for all datasets besides the Cars 3D dataset, showing that, on average, the higher the value of α, the lower the disentanglement.

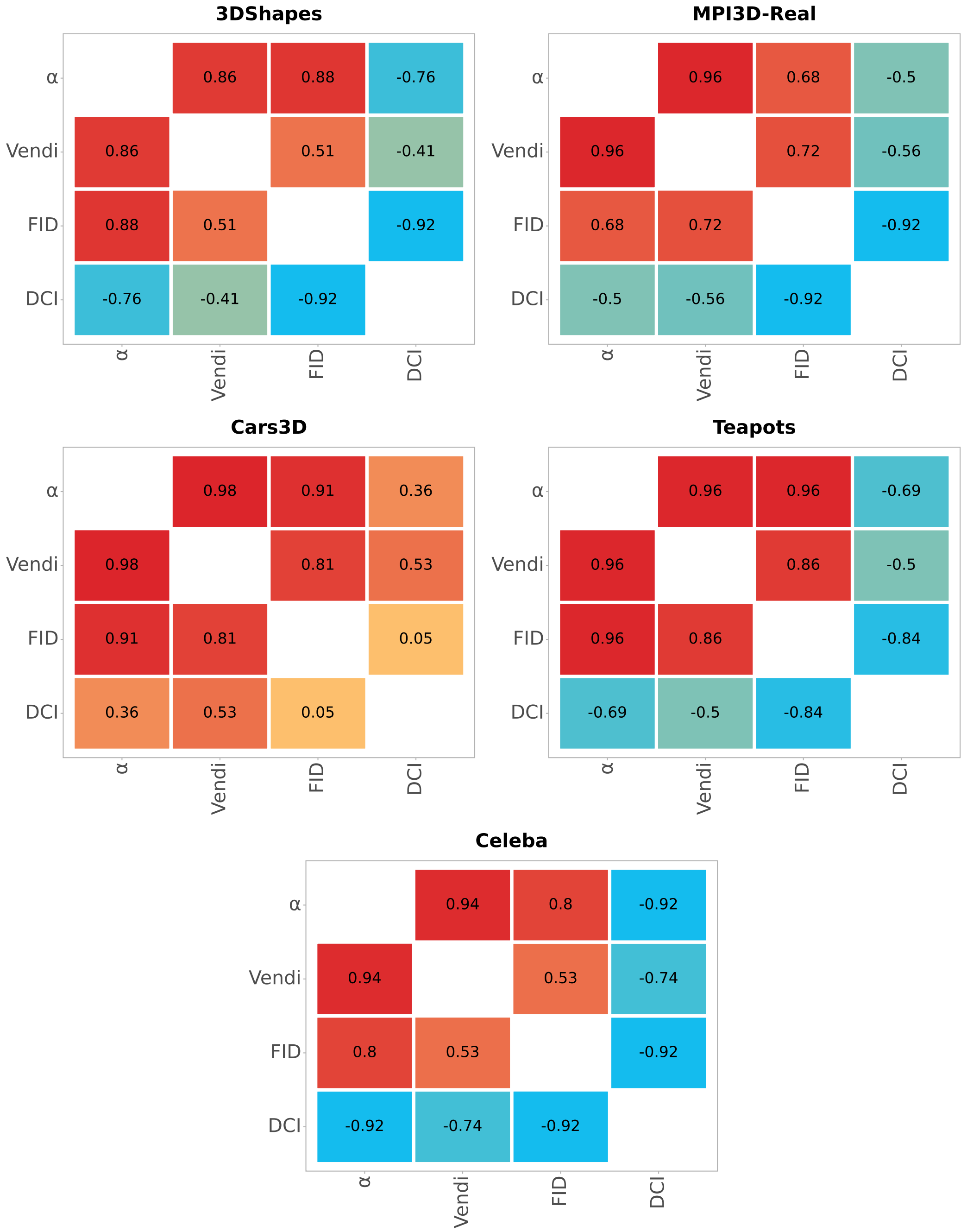


Figure 20: Correlations between α, diversity (Vendi score), generation faithfulness (FID score), and disentanglement (DCI). Correlations are computed using the results from all models across 5 different seeds.