Provable Concept Learning for Interpretable Predictions Using Variational Autoencoders

Anonymous Authors[1]

Abstract

In safety-critical applications, practitioners are reluctant to trust neural networks when no interpretable explanations are available. Many attempts to provide such explanations revolve around pixel-based attributions or use previously known concepts. In this paper we aim to provide explanations by provably identifying *high-level, previously unknown ground-truth concepts*. To this end, we propose a probabilistic modeling framework to derive (C)oncept (L)earning and (P)rediction (CLAP) – a VAE-based classifier that uses visually interpretable concepts as predictors for a simple classifier. Assuming a generative model for the ground-truth concepts, we prove that *CLAP* is able to identify them while attaining optimal classification accuracy. Our experiments on synthetic datasets verify that CLAP identifies distinct ground-truth concepts on synthetic datasets and yields promising results on the medical Chest X-Ray dataset.

1. Introduction

Suppose a hospital aims to deploy a model that classifies diseases $\mathbf{Y}$ from medical images $\mathbf{X}$ and informs the doctor about relevant predictive features. There may be multiple diseases such as lung atelectasis and lung infiltration and multiple *interpretable* ground-truth *features* (or *concepts*) $\mathbf{Z}_c$, such as lung or heart shape, that are relevant for predicting each disease. Ideally, in addition to identifying and utilizing these interpretable features, the model should perform prediction in an interpretable manner itself. The domain expert can then check whether the model is reasonable and also potentially make new scientific discoveries – i.e. discover new factors relevant for prediction.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Thus, in this paper, we seek an interpretable predictive model that uses the ground-truth features for prediction. But what makes a predictive model interpretable from a practical perspective? Even though the definite answer depends on the application domain, practitioners often agree on the following desiderata: first of all, the model should be *simple* – e.g. additive in the predictive features with a small number of relevant features. Simplicity allows us to interpret the relevance of each variable (Rudin, 2018), and ensure that the interpretation is robust to small changes to the input (Alvarez-Melis & Jaakkola, 2018a;b). Furthermore, the model ideally assigns *global and local* importance to the features used for prediction (Reyes et al., 2020; Stiglic et al., 2020); in the context of medical imaging for example, the former corresponds to the population-level importance, the latter to the patient-level one.

While there have been many works on interpretable predictions, none of them provide a prediction model that identifies and uses these previously unknown ground-truth features (see relate works for more discussion). This paper tries to go bottom-up, starting from a generative model to derive a procedure based on variational inference that satisfies all the desiderata. Our proposed framework i) mathematically formalizes concept learning and ii) provably identifies the ground-truth concepts and provides an accurate and simple prediction model using these discovered concepts.

More concretely, we view the recovery of the ground-truth concepts as a latent variable estimation problem. We start by assuming an explicit graphical model for the joint distribution of $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$. Here, the latent variables $\mathbf{Z}$ include all ground-truth latent features, as well as others irrelevant for prediction. Together, the latent variables $\mathbf{Z}$ generate the raw observation $\mathbf{X}$. The task of concept learning can then be mathematically thought of as obtaining identifiability and performing inference on the latent factors. Using a VAE-based architecture, we enable both visualization (and thus facilitate human interpretation) of the learned concepts, as well as prediction based on these.

In summary, we make the following contributions:

1. We present a framework to model ground-truth latent features $\mathbf{Z}_c$ (*Sec. 2*), and derive C(oncept) (L)earning

and (P)rediction (in short *CLAP*), an inherently interpretable prediction framework based on variational autoencoders (*Sec. 3*)

2. We prove that *CLAP* enables identification of the ground-truth concepts underlying the data and learns a simple optimal prediction model based on these. Importantly, our framework does not require knowing the number of latent features (*Sec. 4*)

3. We validate *CLAP* on various multi-task prediction scenarios on synthetic (MPI3D, Shapes3D and Small-Norbs) datasets that yield encouraging results on domain-specific application of the framework on real data (*Sec. 5*)

We believe that our theoretical framework is a useful step for formalizing interpretable predictions. In particular, in settings where it's reasonable to assume that the ground-truth features are themselves interpretable by a domain expert, *CLAP* provably provides an end-to-end interpretable prediction model. Even when the assumption does not hold, we can still guarantee that *CLAP* finds a simple and accurate prediction model using ground-truth features.

1.1. Related work

In this section, we compare existing interpretable prediction methods with *CLAP* in detail, with a concise summary provided in Table 1. Previous methods proposed in the context of explainable/interpretable AI can be broadly divided into two categories: (i) providing post-hoc explanations for black-box prediction models and (ii) designing models that explicitly incorporate transparency into the model design, where the explanation is learned during training.

Post-hoc methods The majority of work on interpretability so far has focused on (i), providing post-hoc explanations for a given prediction model. These include pixel attribution methods (Bach et al., 2015; Selvaraju et al., 2017; Simonyan et al., 2014), counterfactual explanations (Antoran et al., 2021; Chang et al., 2019), explanations based on pre-defined concepts (Kazhdan et al., 2020; Rezende et al., 2014; Yeh et al., 2020), and recently developed StyleGANs (Lang et al., 2021; Wu et al., 2021). Post-hoc methods have a number of shortcomings given our desired objectives: First, it is unclear whether post-hoc explanations indeed reflect the black-box model's true "reasoning" (Kumar et al., 2020; Rudin, 2018). Even if an expert deems the output of the explanation model as unreasonable, one is unable to determine whether the explanation method or the original model is at fault. Furthermore, by construction, post-hoc methods cannot come with statistical inference guarantees and ensure that the learned concepts align with the ground-truth features. Finally, post-hoc methods are typically used to explain complex classifiers; as a result, they are unable to provide meaningful global and local importance of features for prediction.

VAE-based methods for inherently interpretable prediction Our procedure *CLAP* is an inherently interpretable prediction model and similar in spirit to VAE-based prediction techniques. On a high level, existing procedures either are unable to identify the ground-truth latent features or require additional labels. Therefore, they are not applicable in the traditional supervised learning setting considered in this paper (where only $\mathbf{X}, \mathbf{Y}$ are available). Further, none of the existing methods provide simultaneous guarantees for learning the underlying concept and obtaining optimal predictions using these learned features. We provide more specific comparisons next.

First of all, unsupervised VAEs (Kingma et al., 2014) can easily be used for prediction tasks by training a classifier on the latent features. A massive literature proposes various structural adjustments to improve disentanglement (Burgess et al., 2018; Chen et al., 2018; Higgins et al., 2017; Kim & Mnih, 2019; Kumar et al., 2017). However, (Locatello et al., 2019) empirically and theoretically demonstrate that these methods generally do not successfully identify the ground-truth latent features. Recently proposed VAE methods address the issue of non-identifiability by assuming access to additional data and improve identifiability. However, they either require the label as direct input (Joy et al., 2021), or labels for auxiliary variables that contain information about the ground-truth latent factors (Khemakhem et al., 2020; Mita et al., 2021) or the ground-truth factors themselves (Locatello et al., 2019). None of these scenarios are applicable to the traditional supervised learning setting in our paper.

Other works With respect to model architecture, our method is similar to Self-Explaining Neural Networks (SENN) (Alvarez-Melis & Jaakkola, 2018b) which decomposes a complex prediction model into learning interpretable concepts (using an autoencoder) and a simple (linear) predictor. More broadly, methods based on contrastive learning or multi-view data (e.g. (Gresele et al., 2019; Hyvärinen et al., 2019; Locatello et al., 2020; Shu et al., 2020; von Kügelgen et al., 2021)) can identify underlying latent features, albeit with access to pairs of images that share similar sources. Furthermore, the focus of these methods is on representation learning rather than interpretable predictions.

2. Modeling interpretable and predictive concepts

We present a probabilistic graphical model that statistically relates the ground-truth latent features $\mathbf{Z}_c$ to the labels and observed variables; our proposed method later uses this model to learn the latent concepts as well as a simple clas-

Table 1: Comparison of *CLAP* with post-hoc explanation methods and other inherently interpretable techniques. Here, the symbol ✓⋆ highlights that for learning visually distinct features, existing predictive VAEs require strong knowledge of the latent variables or auxiliary variables (in addition to labels).

Properties	Post-hoc explanations			Inherently interpretable	
	pixel attribution+ counterfactuals	pre-defined concepts	StyleGANs	existing VAEs/ autoencoders	*CLAP*
Learning visually distinct features	×	×	✓	✓⋆	✓
Global importance of predictive features	×	✓	×	×	✓
Guarantees: concept learning+prediction	×	×	×	×	✓

sifier based on these features. We remark that, although the methodology in this paper is presented under a specific generative model, the framework is general and flexible to other modeling choices.

Let $\mathbf{X}$ be raw observations and $\mathbf{Y} \in \mathcal{Y}$ be the associated label vector taking a finite collection of values. In general, $\mathbf{X}$ is comprised of *style factors* $\mathbf{Z}_s$, that should not be relevant for prediction, and high-level *core factors* $\mathbf{Z}_c$ that are the desired ground-truth concepts. For example, in the context of medical imaging, $\mathbf{Y}$ are various disease labels such as the presence of lung atelectasis and lung infiltration. Core factors $\mathbf{Z}_c$ that one can see in the X-ray image $\mathbf{X}$, such as heart and lung shapes, are typically direct consequences of a patient contracting the disease. Style factors $\mathbf{Z}_s$ such as physiological characteristics of the subject or specialities of the scanner are also factors that appear in the image but are not related to the disease.

A natural model for settings such as the one above is to assume an *anti-causal* model as in Fig. 1(a), where $\mathbf{Z}_c$ is a child of $\mathbf{Y}$, and combines with $\mathbf{Z}_s$ to produce the raw observation $\mathbf{X}$. We assume $\mathbf{Z}_c$ to be independent conditionally on $\mathbf{Y}$, as in the X-ray example, they may often vary independently (across patients) given a disease label. We instead allow arbitrary dependencies within $\mathbf{Z}_s$ and $\mathbf{Y}$.

Aggregating style and core factors in the vector $\mathbf{Z} = (\mathbf{Z}_c, \mathbf{Z}_s)$, we impose the following structural equation model on the graph in Fig. 1(a):

$$\mathbf{X} = f^\star(\mathbf{Z}) + \epsilon \;\text{ where } \epsilon \perp\!\!\!\perp \mathbf{Z}, \mathbf{Y} \text{ and for all } y \in \mathcal{Y}:$$
$$\mathbf{Z}|\mathbf{Y}{=}y \sim \mathcal{N}\left(\begin{pmatrix}\mu_y^\star\\ \mu^\star\end{pmatrix}, \begin{pmatrix}D_y^\star & 0\\ 0 & G^\star\end{pmatrix}\right); D_y^\star \text{ diagonal}, \tag{1}$$

for some continuous one-to-one function $f^\star$, vectors $\mu_y^\star, \mu^\star$, and positive-definite matrices $D_y^\star, G^\star$. The model (1) encodes the conditional independence relationships in Fig. 1(a): the covariance of the distribution $\mathbf{Z}_c|\mathbf{Y}$ is diagonal; the mean and covariance corresponding to $\mathbf{Z}_s$ are not a function of y and the noise ϵ is independent of $\mathbf{Y}$ so that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}_c$ and $\mathbf{Z}_s \perp\!\!\!\perp \mathbf{Y}$.

3. *CLAP*: interpretable predictions using ground-truth concepts

Given data of $\mathbf{X}$ and $\mathbf{Y}$ arising from the graphical model in Fig. 1(a), our objective is to identify the ground-truth concepts and learn a simple classifier that uses these to accurately predict $\mathbf{Y}$. Additionally, to facilitate human interpretability, we aim to enable experts in the loop to visually interpret the learned concepts. For concreteness, we specialize our exposition to images, although our framework can in principle be used on other types of data.

Our proposed framework is based on variational autoencoders (VAEs) (Kingma et al., 2014; Rezende et al., 2014). VAEs offer a number of favorable properties for our objectives. First, they can be derived in a principled manner from the underlying data generating mechanism. Second, the encoder/decoder pair in VAEs provide an effective approach to visualize and thus interpret the learned latent features via latent traversals (see Sec. 3.4 for more details).

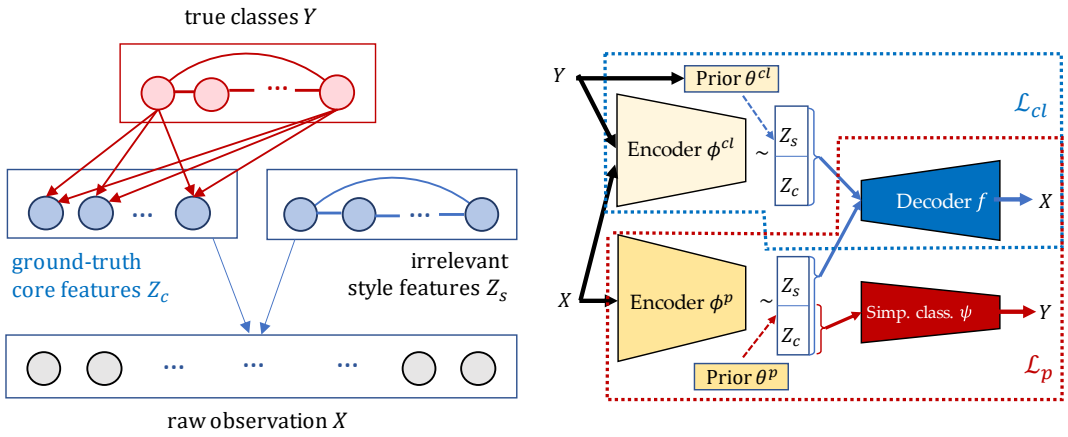
In that light, a natural first approach that might come to mind would be to train a VAE that uses the estimated latent features for prediction. In Sec. 3.1 we derive such a model, and show why, in its vanilla version, it can perform prediction but cannot identify the ground-truth core concepts. In Sec. 3.2, we overcome these challenges by introducing a novel VAE architecture *CLAP* shown in Fig. 1(b). Our proposed method combines the predictive VAE structure from earlier with a second VAE which helps with identifying the underlying ground-truth concepts.

3.1. Vanilla predictive VAE and its shortcomings

A natural first attempt at learning a predictive VAE procedure is to maximize the following ELBO of the log-evidence of $(\mathbf{X}, \mathbf{Y})$:

$$\log p(\mathbf{X}, \mathbf{Y}) \geq \mathbb{E}_{q_{\phi^p}(\mathbf{Z}|\mathbf{X})} \log \frac{p_f(\mathbf{X}|\mathbf{Z}) p_\psi(\mathbf{Y}|\mathbf{Z}_c) p_{\theta^p}(\mathbf{Z})}{q_{\phi^p}(\mathbf{Z}|\mathbf{X})}$$
$$=: \mathcal{L}_p(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y}). \tag{2}$$

The objective $\mathcal{L}_p$ corresponds to the VAE architecture in the red box in Fig. 1(b). Here, q is the approximate poste-



(a) Generative graphical model

(b) VAE architecture of *CLAP*

Figure 1: The graphical model in (a) describes how the desired high-level core latent features $\mathbf{Z}_c$ are related to the remaining variables $\mathbf{Y}, \mathbf{X}, \mathbf{Z}_s$. The VAE architecture in (b) is derived by lower-bounding the evidence values $p(\mathbf{X}, \mathbf{Y})$ and $p(\mathbf{X}|\mathbf{Y})$ and incorporating the generative assumptions from (a) (see main text). We utilize two separate encoders, correspondent to the $\mathcal{L}_{cl}$ and $\mathcal{L}_p$ terms of objective (3), and impose sharing of the decoder. The two encoders define two different sets of latents $\mathbf{Z} = (\mathbf{Z}_c, \mathbf{Z}_s)$, which are separately passed through f to get the relative reconstructions. The two resulting objectives $\mathcal{L}_p$ and $\mathcal{L}_{cl}$ are then summed in the full objective $\mathcal{L}_{CLAP}$. A simple classifier based on $\mathbf{Z}_c$ is trained as part of the model inside $\mathcal{L}_p$.

rior with encoder parameters ϕ^p, ψ parameterizes a simple classifier, f is the decoder's parameters, and θ^p the prior distribution's parameters. Specifically, from the data generating mechanism (1), the prior $p_{\theta^p}(\mathbf{Z})$ is a density of a Gaussian mixture distribution with $|\mathbf{Y}|$ (number of labels) components, where the covariance corresponding to the core features for each mixture component is diagonal. The ELBO (2) is derived in a standard fashion using Jensen's inequality $\log p(\mathbf{X}, \mathbf{Y}) \geq \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{Y})} \log \frac{p(\mathbf{X},\mathbf{Y}|\mathbf{Z})p(\mathbf{Z})}{q(\mathbf{Z}|\mathbf{X},\mathbf{Y})}$ and leveraging the assumed generative model (1).

The model learned by maximizing the objective $\mathcal{L}_p$ naturally yields a classifier $p_\psi(\mathbf{Y}|\mathbf{Z}_c)$ based on core features extracted from the encoder $q_{\phi^p}(\mathbf{Z}|\mathbf{X})$, which should approximate the ground-truth ones. Since the encoder does not rely on $\mathbf{Y}$ as an input, we can readily use it for end-to-end classification during test time. In fact, under a regularity condition, we show in Supp. Mat. Sec. A.2.1 that this architecture is optimal for prediction. However, it *does not* guarantee that the estimated core features $\hat{\mathbf{Z}}_c$ correspond to the ground-truth factors $\mathbf{Z}_c$. In fact, they can be arbitrary linear transformations of $\mathbf{Z}_c$ without sacrificing prediction performance (Locatello et al., 2019) (see ablation studies in Sec. 4), thus not satisfying our desired properties. In addition, as the dimensionality of the core features $\mathbf{Z}_c$ is typically unknown, a conservative choice for the number of latent features (over-parameterized setting) may wrongly include style features or redundant core features in the prediction model (see ablation study in Sec. 4). In the next section, we propose our framework *CLAP* that mitigates the aforementioned issues: it learns a prediction model using the ground-truth core concepts (even in the over-parameterized setting), without sacrificing classification accuracy.

3.2. CLAP to overcome shortcomings

To overcome the aforementioned challenges, we augment the objective $\mathcal{L}_p$ with two additional terms to arrive at our proposed objective function for *CLAP*:

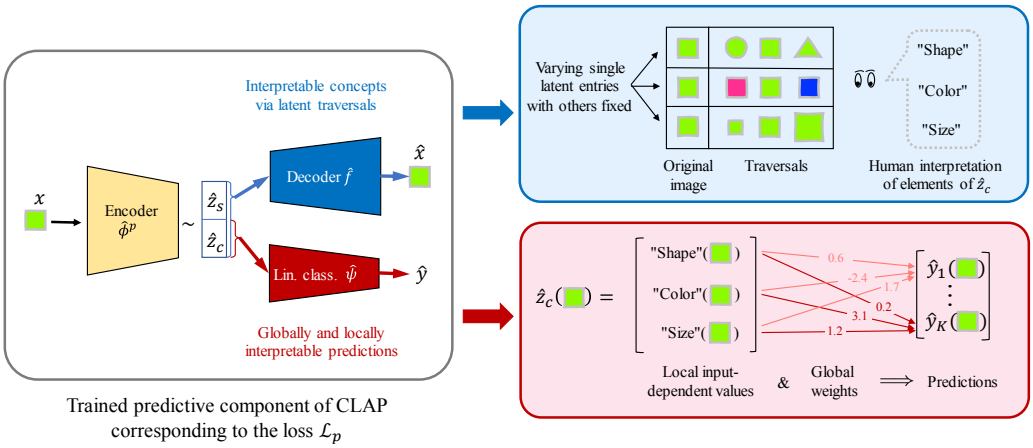
$$\mathcal{L}_{CLAP} := \mathcal{L}_p + \mathcal{L}_{cl} - \lambda_n \rho \,. \tag{3}$$

On a high level, the additional component $\mathcal{L}_{cl}$ ensures identifiability of the ground-truth concepts $\mathbf{Z}_c$ (*concept learning*) and the regularization term $\lambda_n \rho$ helps to identify a minimal number of ground-truth concepts in an over-parameterized latent space. In the following, we formalize each term.

Concept-learning component $\mathcal{L}_{cl}$ While the objective $\mathcal{L}_p$ is designed to maximize the full likelihood of image data $\mathbf{X}$ and target labels $\mathbf{Y}$, the term $\mathcal{L}_{cl}$ maximizes the likelihood of $\mathbf{X}$ conditioned on $\mathbf{Y}$. The fact that the labels act as additional input data in this likelihood objective, plays a central role in provably obtaining identifiability. Furthermore, the conditional independence of $\mathbf{Z}_c$ given $\mathbf{Y}$ can be more naturally captured when $\mathbf{Y}$ is considered as an input. Similarly to above, for any posterior q, we can lower-bound the conditional log-evidence as $\log p(\mathbf{X}|\mathbf{Y}) \geq \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{Y})} \log \frac{p(\mathbf{X}|\mathbf{Z},\mathbf{Y})p(\mathbf{Z}|\mathbf{Y})}{q(\mathbf{Z}|\mathbf{X},\mathbf{Y})}$, and incorporate the generative assumptions in (1) to obtain the final ELBO objective:

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{Y}) &\geq \mathbb{E}_{q_{\phi^{cl}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})} \log \frac{p_f(\mathbf{X}|\mathbf{Z})p_{\theta^{cl}}(\mathbf{Z}|\mathbf{Y})}{q_{\phi^{cl}}(\mathbf{Z}|\mathbf{X},\mathbf{Y})} \\ &:= \mathcal{L}_{cl}(\phi^{cl}, \theta^{cl}, f; \mathbf{X}, \mathbf{Y}) \,. \end{aligned} \tag{4}$$

The component of *CLAP* corresponding to $\mathcal{L}_{cl}$ is highlighted in blue in Fig. 1(b). Here, ϕ^{cl} are the parameters of the



Trained predictive component of CLAP corresponding to the loss $\mathcal{L}_p$

Figure 2: We present how the prediction model obtained by training *CLAP* can be used and interpreted at test time. Supplying a test images x to the component $\mathcal{L}_p$ of *CLAP*, we learn core features $\hat{\mathbf{Z}}_c$. These features are visualized using latent traversals and interpreted by a human, who assigns them to high-level concepts. Furthermore, the estimated linear classifier predicts a label and provides global (population wise) and local (instance wise) importance for the interpreted concepts.

encoder, and f those of the decoder. Appealing to the data generating mechanism (1), we can further factorize the prior in the form $p_{\theta^{cl}}(\mathbf{Z}|\mathbf{Y}) = p(\mathbf{Z}_c|\mathbf{Y})p(\mathbf{Z}_s)$. Here, $p(\mathbf{Z}_c|\mathbf{Y})$ is a Gaussian density function with diagonal covariance and different parameters for different $\mathbf{Y}$ while we model the prior $p(\mathbf{Z}_s)$ as a standard Gaussian distribution without loss of generality. We aggregate all these parameters in θ^{cl}.

In general, maximizing the ELBO or even the true log-evidence would not allow for of identification the true concepts. However, a simple heterogeneity assumption can alleviate this issue, formally stated in Supp. Mat. Sec. A.1.

Assumption 1 (Concept learning, informal). The functions $f, f^\star$ satisfy a regularity condition and the distribution of core features change 'enough' when conditioned on different realizations of $\mathbf{Y}$.

Lemma 1 (Maximizing $\mathcal{L}_{cl}$ identifies the ground-truth concepts). Suppose the data is generated according to the model in (1) with no noise, i.e. $\epsilon \equiv 0$ and Assumption 1 holds. Suppose $\mathcal{L}_{cl}$ is maximized in the infinite data limit with the correct number of latent features included in the model. Then, the posterior samples $\hat{\mathbf{Z}}_c$ obtained from the encoder $q_{\hat{\phi}^{cl}}$ are equal to the ground-truth features $\mathbf{Z}_c$ up to permutation and scaling.

We prove this lemma in Supp. Mat. Sec. A.2.2, and also extend to the noisy setting in Supp. Mat. Sec. A.4. Theoretical results for identifibiality were previously established in (Khemakhem et al., 2020). We note that our guarantees differ substantially and refer to Supp. Mat. Sec. B for more details. Despite the concept-learning capabilities, a model trained only on $\mathcal{L}_{cl}$ *cannot* be used for prediction since it requires the labels as input to the encoder $q_{\phi^{cl}}(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$.

Therefore, we combine the objectives $\mathcal{L}_p$ and $\mathcal{L}_{cl}$ by utilizing the same decoder f in (2) and (4), as represented in Fig. 1(b). This coupling via a shared decoder is crucial, as it forces the $\mathcal{L}_p$ architecture to also perform concept learning: During joint training, the two encoders of $\mathcal{L}_{cl}$ and $\mathcal{L}_p$ learn approximately the same latent space (we show in Theorem 1 that the latent spaces align in the infinite data limit). [1]Since $\mathcal{L}_{cl}$ provably identifies the ground-truth features in the latent space, it then follows that the estimated core features obtained by the encoder of $\mathcal{L}_p$ closely align with $\mathbf{Z}_c$. Thus, after training the combined objective $\mathcal{L}_p + \mathcal{L}_{cl}$, the trained VAE architecture corresponding to $\mathcal{L}_p$ provides an interpretable prediction model: an input image is mapped to accurate ground-truth core features, which are then used on top of a simple classifier to predict the target label $\mathbf{Y}$. We refer the reader to Sec. 3.4 for more discussion on how the trained *CLAP* is used at test time.

Sparsity penalty ρ to account for overparameterized latent space We add a regularization term $\lambda_n \rho(f, \psi)$ to impose simultaneous group sparsity on the prediction weights and decoder weights – this ensures that if an estimated core feature feature is predictive, it has non-negligible effect in the reconstruction of the image and vice versa. In particular, let k_c, k_s be the conservative choice on the dimensionality of the core and style features in our VAE model, respectively. Further, let $k = k_c + k_s$ be the total number of latent variables. We consider the following parameterization for the decoder $f = f' \circ B$, $B \in \mathbb{R}^{k \times k}$ and classifier $\psi = \psi' \circ C$, $C \in \mathbb{R}^{k_c \times k_c}$, where $|\mathbf{Y}|$ is the number of labels to be predicted and f', ψ' are one-to-one and continuous. Then, the sparsity inducing penalty $\rho(f, \psi)$ in the combined

[1]Informally speaking, the reason for this is that the latent features in each architecture reconstruct the image via the same decoder. Since the common decoder defines a generative model , the posteriors (i.e. the different encoders) need to be similar as well.

objective function (3) takes the form:

$$\rho(f,\psi) := \sum_{i=1}^{k_c} \mathbb{I}\left[\left\|\begin{pmatrix} B_{:,i} \\ C_{:,i}\end{pmatrix}\right\|_2 > 0\right] + \sum_{i=k_c+1}^{k} \mathbb{I}\left[\left\|B_{:,i}^T\right\|_2 > 0\right], \tag{5}$$

where the indicator function $\mathbb{I}[\cdot]$ counts the number of latent features effectively utilizes by the model. Note that the nonzero columns of C correspond to core features in the model with predictive power, and the nonzero columns of B correspond to core and style features that are used for reconstruction with the decoder f. For practical considerations, we consider the following convex surrogate in our experiments: $\rho(f,\psi) = \sum_{i=1}^{k_c} \left\|\begin{pmatrix} B_{:,i}^T & C_{:,i}^T\end{pmatrix}\right\|_2$.

3.3. Theoretical guarantees for *CLAP*

In Sec. 3.2, we described how after the training of *CLAP*, the component corresponding to $\mathcal{L}_p$ can be used as an interpretable prediction model. We next provide guarantees that this prediction model is optimal in terms of accuracy and is based on high-level features that align with the ground-truth concepts. In the sequel, we denote k_c, k_s to be the number of core and style features chosen in the VAE architecture and $k_c^\star, k_s^\star$ to be the dimensions of the corresponding true features of the generative model in Fig. 1(a). Further, we use $q_{\hat{\phi}^p}$, $q_{\hat{\phi}^{cl}}$ to denote the encoders obtained by maximizing the objective in (3) in the infinite data limit and let $\hat{\mathbf{Z}}$ be the posterior samples obtained from $q_{\hat{\phi}^p}$ for input $\mathbf{X}$?. Finally, we denote the trained classifier as $\hat{\psi} = \hat{\psi}' \circ \hat{C}$, and the core features $\hat{\mathbf{Z}}_c$ are specified as the vector of elements of $\hat{\mathbf{Z}}$ corresponding to nonzero columns of $\hat{C}$.

Our main theorem additionally requires an assumption about a simple classifier being optimal:

Assumption 2 (optimal classifier). The Bayes optimal classifier for predicting $\mathbf{Y}$ using $\mathbf{Z}_c$ belongs to the set of simple classifiers used in *CLAP*.

Theorem 1 (*CLAP* learns an optimal prediction model using interpretable ground-truth features). Consider the same setup as Lemma 1. Suppose $k_c \geq k_c^\star$, $k_s \geq k_s^\star$, and that Assumptions 1 and 2 hold. Then, the posterior samples $\hat{\mathbf{Z}}$ obtained from the encoder $q_{\hat{\phi}^p}$ are identical to the posterior samples obtained from the encoder $q_{\hat{\phi}^{cl}}$. Furthermore, the core features $\hat{\mathbf{Z}}_c$ are 1) optimally predictive: $\mathbf{Y}|\hat{\mathbf{Z}}_c \overset{\text{dist}}{=} \mathbf{Y}|\mathbf{X}$, and 2) aligned with the ground truth: $\hat{\mathbf{Z}}_c$ is equal to $\mathbf{Z}_c$ up to scaling and permutation.

The proof of Theorem 1 is presented in Supp. Mat. Sec. A.3. Our guarantees in Theorem 1 ensure that the prediction model obtained by *CLAP* is optimal. Furthermore, the core features $\hat{\mathbf{Z}}_c$ align with the ground-truth concepts. Finally, the number of predictive factors equals to the number of ground-truth concepts; that is, our model obtains the minimal set of predictive features.

3.4. Using *CLAP*'s output for interpretation

We now discuss how *CLAP*'s trained model can be used to produce an end-to-end interpretable prediction model pipeline, which we represent in Fig. 2.

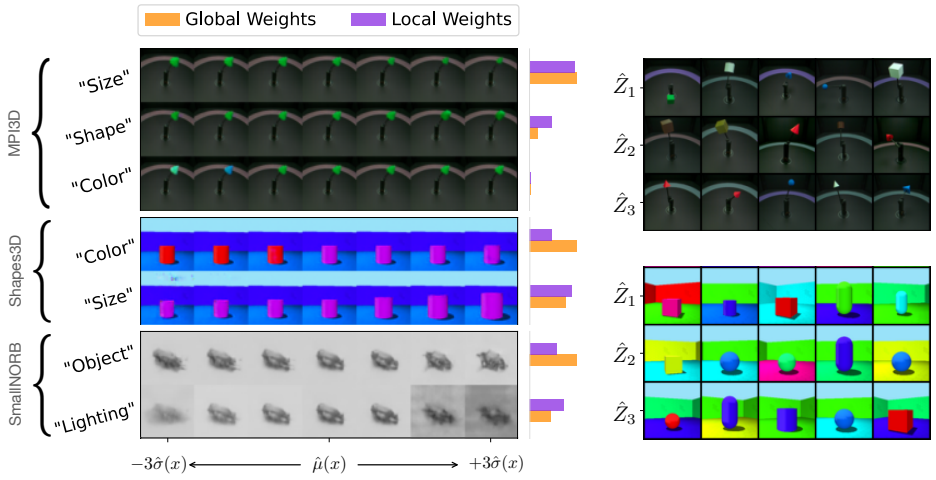
At inference time, the part of *CLAP*'s model corresponding to $\mathcal{L}_p$ is utilized, since it does not require a label as an input (Fig. 2 left). As we describe in detail next, the learned concepts are visualized using latent traversals; to conclude the pipeline, a human expert visually inspects these traversals and assigns a meaning to the relative latent variables.

Interpretations via latent traversals Generally, the visual explanations provided by the model need to be evaluated by a human expert (see Sec. 1). As is customary for VAE models, we provide such visualizations via latent traversals. Specifically, let x be an input image. The core concepts associated to x are obtained via the posterior mean $\hat{\mu}(x) := \mathbb{E}_{q_{\hat{\phi}^p}(\hat{\mathbf{Z}}_c|x)}[\hat{\mathbf{Z}}_c]$. The semantics of $\hat{\mathbf{Z}}_c$ are then discovered by performing latent traversals. In these, we change one component of $\hat{\mu}(x)$ at a time, while keeping the others fixed, and observe the reconstructions obtained through the decoder $\hat{f}$. Owing to the concept-learning capabilities of *CLAP*, the traversals on the core latent features will produce distinct changes in the reconstructed images corresponding to the different discovered ground-truth concepts, which will allow the human expert to assign them with a semantic meaning. This procedure is represented in the top-right of Fig. 2. There, for example, upon visual inspection, the first latent is assigned the meaning of *"Shape"* from the expert, the second *"Color"*, and so on.

Interpretable predictions using learned concepts We note here that in our experiments, we found a linear classifier to be well-performing across all datasets. For this reason, the following description assumes ψ to simply be the linear weights of the corresponding linear classifier $p_\psi(\mathbf{Y}|\hat{\mathbf{Z}}_c)$. For each concept, we provide both a global and local relevance for prediction, as depicted in the bottom right of Fig. 2. The global relevance represents the importance of a concept for prediction at a population level (i.e. across images) and is thus directly encoded in the entries of $\hat{\psi}$. The local relevance is instead image-specific, and is observed in the summands of the linear combination $\langle \hat{\mu}(x), \hat{\psi} \rangle$. These two measures allow the practitioner to transparently assess the decision process of the model, as they assign a prediction weight to human interpretable features.

4. Experimental results

We next present experiments on synthetic data to corroborate our theoretical results, and evaluate the ability of *CLAP* to learn an accurate prediction model using the ground-truth features. Since in most real-world datasets, ground-truth factors are unknown but necessary to verify whether *CLAP* can work in practice, we resort to three standard "disentan-



(a) *CLAP* traversals and interpretations (b) SENN prototypes

Figure 3: a) *CLAP* traversals on (in order) the MPI3D, Shapes3D and SmallNORB datasets, and b) SENN prototypes on (in order) the MPI3D and Shapes3D datasets.

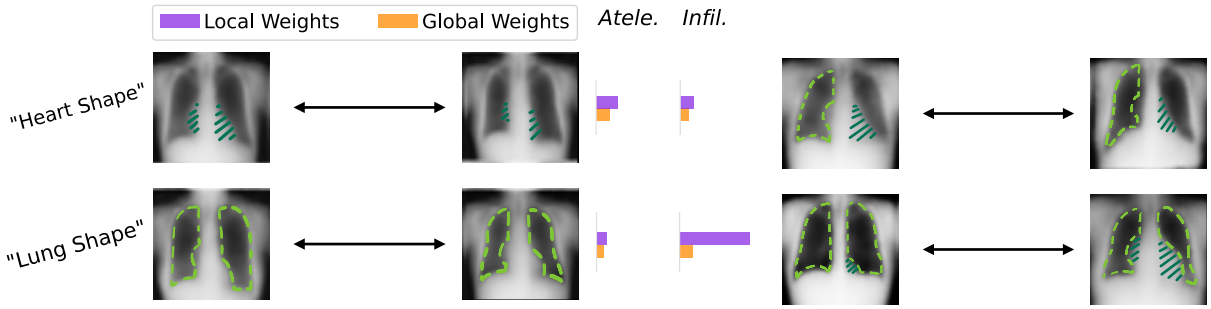
glement" datasets MPI3D (Gondal et al., 2019), Shapes3D (Burgess & Kim, 2018) and SmallNORB (LeCun et al., 2004). These datasets consist of collections of objects generated synthetically according to some ground-truth factors of variation. The images are a priori unlabeled; thus, we select some of the ground-truth factors, which represent the concepts $\mathbf{Z}_c$ to be discovered, and generate artificial binary labels $\mathbf{Y}$. The ground-truth factors $\mathbf{Z}_c$ are object shape, size and color for MPI3D, object color and size for Shapes3D and object type and lighting for SmallNORB (see Supp. Mat. Sec. D). For all the experiments and baselines in Sec. 4, details on training and architectures employed are deferred to Supp. Mat. Sec. C. In general, for all methods, we used neural network architectures comparable in complexity to those utilized in (Joy et al., 2021; Qiao et al., 2019).

As explained in Sec. 3.4, we proceed with the evaluation of *CLAP* by first generating latent traversals. The goal is to determine whether the discovered concepts have a one-to-one correspondence with the ground-truth $\mathbf{Z}_c$ that we used to generate the data. In Fig. 3(a), every row corresponds to the traversal for one latent feature. As can be observed, the estimated core features indeed represent the ground-truth ones; this means that the model identifies the ground-truth concepts underlying the data generating mechanism. Importantly, we remark that the concept names assigned to the single rows (e.g. *"Size", "Shape"*) are obtained by visual inspection; the model doesn't have direct access to them, but only to the images $\mathbf{X}$ and labels $\mathbf{Y}$.

Finally, the discovered $\mathbf{Z}_c$ are also fully predictive, as *CLAP* achieves classification accuracy above 0.99 on all the datasets. We include additional traversals in Supp. Mat. Sec. D.1; there, we also show that, due to the sparsity regularization penalty $\rho(f, \psi)$, the model accurately assigns negligible global and local weights (i.e. no predictive value) to the remaining latent features included in the model. This is in contrast to the concepts shown in Fig. 3(a) that have non-negligible global and local weights. In other words, in line with our theory, estimated core features that have prediction power align with the ground-truth concepts.

Comparison with baselines We compare the outputs of *CLAP* with those of SENN (Alvarez-Melis & Jaakkola, 2018b) and CCVAE (Joy et al., 2021), two prediction models in the existing literature that are closest to *CLAP*. To explain its predictions and visualize the learned concepts, SENN uses prototypes – a set of training images that "best represent" every latent variable. In Fig. 3(b), we depict the prototypes relative to some of these features. Similarly to *CLAP*, human inspection is needed to describe the concepts that such latents encode. However, the task here is substantially more difficult: for any of the latents, we can observe many different changes, e.g in the first row objects of different colors and shapes are observed, and from different camera angles. This indicates that not only SENN is not able to identify the ground-truth $\mathbf{Z}_c$, thus hindering interpretability, but also mixes them with non-predictive style features $\mathbf{Z}_s$. We also apply CCVAE on synthetic data and observe that its learned latent features do not align with the ground-truth ones; due to space constraints, we show these results in Supp. Mat. Sec. F.

Ablation studies In order to demonstrate the importance of each of our design choices, we also perform various ablation studies on the MPI3D dataset, presented in Supp. Mat. E. Firstly, we show that if the sparsity penalty $\lambda_n \rho(f, \psi)$ is removed from the learning objective, the resulting model utilizes separately some latent variables for visualization, and some others for prediction. On the other hand, with the



(a) *CLAP* traversals and interpretations (b) CCVAE traversals

Figure 4: Output of *CLAP* and traversals of CCVAE for the Chest X-ray dataset. In (a), we present the weights for both the atelectasis and lung infiltration disease predictions, as well as the human interpretations of the discovered concepts. For better visual comparison, we only show the images obtained at the extremes of the latent traversals. Moreover, we highlight the changes that occur during the traversals. We include magnified figures with full traversals in Supp. Mat. Sec. G, as well as a glossary on how to read the results.

use of $\lambda_n \rho(f, \psi)$, *CLAP* ensures correspondence between features utilized for prediction and visualization. Furthermore, we show latent traversals for a model trained only on $\mathcal{L}_p$. As explained in Sec. 3.1, the learned features are fully predictive, but do not correspond to the ground-truth one. In fact, it can be observed that various ground-truth features change jointly within one single traversal. Further, we empirically confirm that the concept-learning capabilities of *CLAP* rely on the labels $\mathbf{Y}$ being informative enough, as highlighted by the assumptions in Sec. 3.2. Practically, this means that multiple labels help with more accurate recovery of the ground-truth $\mathbf{Z}_c$; we show that the concept learning capabilities of *CLAP* indeed decrease on a dataset where only one label is available.

5. Future Outlook

So far, we have evaluated *CLAP* in synthetic scenarios where we know the ground-truth data generating mechanism and the core factors are easy to recognize for a layperson. For many scientific scenarios such as the example in the introduction, evaluating whether learned concepts correspond to the "ground-truth" can only be done by domain experts. Nevertheless, we provide the outputs of *CLAP* for some challenging real datasets to highlight some of its favorable properties compared to other competing methods.

In this section, we present results on the Chest X-ray dataset, and defer additional experiments on the PlantVillage dataset (Hughes et al., 2015) to Supp. Mat. Sec. H. The Chest X-ray dataset (Wang et al., 2017) consists of radiography images; each image has 14 associated binary disease labels. We emphasize that only the disease labels may be used to learn the underlying concepts and no additional supervision is available. As explained in Sec. 1.1, many inherently interpretable models cannot be applied successfully in this setting, since they generally assume further information on the ground-truth factors. Due to the negative results for SENN in Sec. 4, we only compare our method with CCVAE.

Both *CLAP* and CCVAE attain similar classification accuracies of 0.903 and 0.898, respectively. In Fig. 4, we compare the traversals obtained by both methods. First we observe that *CLAP* manages to learn concepts that are localized in the X-Ray image, corresponding to separate properties, such as *"Heart shape"* and *"Lung shape"*. Instead, in both the traversals presented for CCVAE, characteristics that can be associated to both the heart and lung shapes vary together. Thus, while CCVAE finds similar concepts for prediction, they do not appear as separate components of $\mathbf{Z}_c$. Hence it is harder for a human expert to uniquely label the learned concepts and, consequently, interpret the model's output.

Another desirable characteristic of *CLAP* is that the global and local weights reflect the importance of the concepts in predicting different diseases. For example, compared to atelectasis, the concept "lung shape" has higher weight (both global and local) in determining the presence of lung infiltration. Since lung infiltration is a condition related to dense substances in the lungs, the concept "lung shape" learned by *CLAP* is natural and indicative. Further, we remark that the discovered concepts manifest through very nuanced traversals. This is sensible, as it is to be expected that real life examples come with subtle and less pronounced features than synthetic and commonly used datasets. In conclusion, these experiments show the advancement and potential of *CLAP* compared to existing methods for providing real-life interpretable predictions.

There are a number of exciting future directions that can further improve *CLAP* for broader and more effective use in real-world scenarios. For example, the visualizations of the VAE are not optimally sharp compared to the status quo for GANs. Hence, it would be interesting to explore whether one can obtain provable concept learning when the VAE is replaced by a GAN structure. Further, in many scientific applications, the number of available images can be quite small. An interesting avenue for future research could be to develop solutions for the small data regime.

References

Alvarez-Melis, D. and Jaakkola, T. On the robustness of interpretability methods. In *arXiv preprint arXiv:1806.08049*, 2018a.

Alvarez-Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In *Neural Information Processing Systems*, 2018b.

Antoran, J., Bhatt, U., Adel, T., Weller, A., and Hernández-Lobato, J. M. Getting a CLUE: A method for explaining uncertainty estimates. In *International Conference in Learning Representations*, 2021.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

Burgess, C. and Kim, H. 3D Shapes Dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Burgess, C., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. In *arXiv preprint arXiv:1804.03599*, 2018.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. K. Explaining image classifiers by counterfactual generation. In *International Conference in Learning Representations*, 2019.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Neural Information Processing Systems*, 2018.

Gondal, M. W., Wüthrich, M., Miladinović, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Neural Information Processing Systems*, 2019.

Gresele, L., Rubenstein, P., Mehrjou, A., Locatello, F., and Schölkopf, B. The incomplete Rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Uncertainty in Artificial Intelligence*, 2019.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference in Learning Representations*, 2017.

Hughes, D., Salathé, M., et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. In *arXiv preprint arXiv:1511.08060*, 2015.

Hyvärinen, A., Sasaki, H., and Turner, R. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Joy, T., Schmon, S., Torr, P., Siddharth, N., and Rainforth, T. Capturing label characteristics in VAEs. In *International Conference in Learning Representations*, 2021.

Kazhdan, D., Dimanov, B., Jamnik, M., Liò, P., and Weller, A. Now you see me (CME): Concept-based model extraction. In *International Conference on Information and Knowledge Management*, 2020.

Khemakhem, I., Kingma, R., Monti, P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In *Artificial Intelligence and Statistics*, 2020.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, 2019.

Kingma, D., Rezende, D., Mohamed, S., and Welling, M. Semi-supervised learning with deep generative models. In *Neural Information Processing Systems*, 2014.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference in Learning Representations*, 2017.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C. E., and Friedler, S. A. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference in Machine Learning*, 2020.

Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W. T., Isola, P., Globerson, A., Irani, M., and Mosseri, I. Explaining in style: Training a GAN to explain a classifier in StyleSpace. In *International Conference in Computer Vision*, 2021.

LeCun, Y., Huang, F., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition*, 2004.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference in Machine Learning*, 2019.

Locatello, F., Poole, B., Raetsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *International Conference in Machine Learning*, 2020.

Mita, G., Filippone, M., and Michiardi, P. An identifiable double VAE for disentangled representations. In *International Conference on Machine Learning*, 2021.

Qiao, J., Li, Z., Xu, B., Cai, R., and Zhang, K. Disentanglement challenge: From regularization to reconstruction. *arXiv preprint arXiv:1912.00155*, 2019.

Reyes, M., Meier, R., Pereira, S., Silva, C., Dahlweid, F., von Tengg-Kobligk, H., Summers, R., and Wiest, R. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology:Artificial intelligence*, 23:e190043, 2020.

Rezende, D., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference in Machine Learning*, 2014.

Rudin, C. Please stop explaining black box models for high stakes decisions. In *arXiv preprint arXiv:1811.10154*, 2018.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *arXiv preprint arxiv:1312.6034*, 2014.

Stiglic, G., Kocbek, P., Fijavko, N., Zitnik, M., Verbert, K., and Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10, 2020.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Scholkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. In *Neural Information Processing Systems*, 2021.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition*, 2017.

Wu, Z., Lischinski, D., and Shechtman, E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Conference on Computer Vision and Pattern Recognition*, 2021.

Yeh, C., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. In *Neural Information Processing Systems*, 2020.

A. Proof of Theoretical results

For simplicity, we first prove guarantees for *CLAP* in the setting where the number of latent variables is specified correctly. Subsequently, we will extend the analysis to the setting where the number of latent variables is miss-specified (i.e. chosen conservatively).

Throughout, we use the following notation. Let $\mathbb{P}_{\mathbf{X}}$ be the distribution of $\mathbf{X}$, $\mathbb{P}_{\mathbf{Y}}$ be the probability distribution of, $\mathbb{P}_{\mathbf{X},\mathbf{Y}}$ be the joint probability distribution of and $(\mathbf{X}, \mathbf{Y})$, and $\mathbb{P}_{\mathbf{X}|\mathbf{Y}}$ be the probability distribution of $\mathbf{X}|\mathbf{Y}$, all with respect to the data generating model. Associated with the probability distributions $\mathbb{P}_{\mathbf{X}}$ and $\mathbb{P}_{\mathbf{X}|\mathbf{Y}}$ are the density functions with we denote by $p^\star(\mathbf{X})$ and $p^\star(\mathbf{X}|\mathbf{Y})$

A.1. Formal description of our assumptions

Let f' be the function from the decomposition $f = f' \cdot B$ introduced in relation to the sparsity regularization term (B is the identity matrix when the number of latents is correctly specified). Assumption 1 then is formally described as follows:

Assumption 1 (Concept learning, formal).

Assumption 1.1. The functions f', $f^\star$ are one-to-one and continuous

Assumption 1.2. There exists $(y, \tilde{y}) \in \mathcal{Y}$ s.t. $D_y^\star (D_{\tilde{y}}^\star)^{-1}$ has distinct diagonal entries not equal to one.

The first component of Assumptions 1 is rather mild and ensures that the functions mapping from the latent space to the input space are injective. The second component of Assumption 1 states that variations in the label Y should impact the variance of all the core latent features (hence the exclusion of value "one") and in a distinct manner. This type of assumption is similar in spirit to requiring "heterogeneous interventions" in causal structural learning. Specifically, in the context of our anti-causal graphical model 1(a), the labels $y \in \mathcal{Y}$ can be viewed as an "environment" variable where an environment dictates the distribution of the core latent features. A change in an environment can then be viewed as interventions on the core latent features. Thus, in this perspective, the second component of Assumption 1 requires the impact of the interventions to be on all of the core features and to be sufficiently heterogeneous. Furthermore, we note that the second component of Assumption 1 requires that changes to the label lead to simultaneous changes to all of the core features. This assumption can be relaxed so that not all of the core features must vary at once with a change in the label, as long as each feature varies for some change to the label. Mathematically, the assumption can be relaxed to the following: for every $i, j \in [k_c], i \neq j$, there exists $y, \tilde{y} \in \mathcal{Y}$ such that $[D_y^\star D_{\tilde{y}}^{\star -1}]_{i,i} \neq [D_y^\star D_{\tilde{y}}^{\star -1}]_{j,j}$ and $[D_y^\star D_{\tilde{y}}^{\star -1}]_{i,i} \neq 1$, $[D_y^\star D_{\tilde{y}}^{\star -1}]_{j,j} \neq 1$.

We now state Assumption 2 formally. Specifically, letting $\mathcal{S}$ be the set of simple classifiers in the training, we require:

Assumption 2 (optimal prediction, formal). The Bayes optimal classifier for predicting $\mathbf{Y}$ using $\mathbf{Z}_c$ belongs to the set of simple classifiers $\mathcal{S}$ used in *CLAP*, i.e.

$$\mathbb{E}[\log p^\star(\mathbf{Y}|\mathbf{Z}_c)] = \max_{\psi \in \mathcal{S}} \mathbb{E}[\log p_\psi(\mathbf{Y}|\mathbf{Z}_c)].$$

A.2. Analysis with known number of latent features

A.2.1. Maximizing $\mathcal{L}_p$ achieves optimal prediction

As described in Section 3.1, maximizing the objective $\mathcal{L}_p$ achieves optimal prediction. We formalize this below.

Lemma 2 (Maximizing $\mathcal{L}_p$ achieves optimal prediction). Suppose the data is generated according to the model in (1) with no noise, i.e. $\epsilon \equiv 0$ and Assumptions 1.1 and 2 hold. Suppose $\mathcal{L}_[$ is maximized in the infinite data limit with the correct number of latent features included in the model. Then, the posterior samples $\hat{\mathbf{Z}}_c$ obtained from the encoder $q_{\hat{\phi}^p}$ are optimallly prediction: $\mathbf{Y}|\hat{\mathbf{Z}}_c \overset{\text{dist}}{=} \mathbf{Y}|\mathbf{X}$.

Proof of Lemma 2. We analyze the following estimator in the infinite data limit:

$$\underset{\phi^p, \theta^p, f, \psi}{\operatorname{argmax}} \mathbb{E}_{\mathbf{X},\mathbf{Y} \sim \mathcal{P}_{\mathbf{X},\mathbf{Y}}} [\mathcal{L}_{\mathrm{p}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})]. \tag{6}$$

The optimization program (6) can be equivalently expressed as:

$$\begin{aligned}\underset{f,\psi,\phi^p,\theta^p}{\operatorname{argmax}} \quad & \underbrace{\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right]}_{\text{reconstruction loss}} \\ & + \underbrace{\mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}}_c)]\right]}_{\text{classification term}}.\end{aligned} \tag{7}$$

Here, $\tilde{\mathbf{Z}}$ is an approximation for the true latent variables $\mathbf{Z}$ with $p_f(x|\tilde{\mathbf{Z}}) = \delta_{f(\tilde{\mathbf{Z}})}$. Consider maximizing the reconstruction loss in (7). In this setting, the VAE model searches for an approximation $\hat{f}(\tilde{\mathbf{Z}}) \overset{\text{dist}}{\approx} \mathbf{X}$ where the parameters of the VAE model (e.g. posterior ϕ^p, prior θ^p, f) are optimized to yield the best approximation of $\mathbf{X}$. In other words, VAE training approximates the following optimization:

$$\underset{f,\theta^p}{\operatorname{argmax}} \ \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}[\log p_{f,\theta^p}(\mathbf{X})], \tag{8}$$

where the likelihood $p_{f,\theta^p}(\mathbf{X})$ is defined with respect to the distribution $\tilde{\mathbf{X}} \overset{\text{dist}}{=} f(\tilde{\mathbf{Z}})$ with $\tilde{\mathbf{Z}}$ being a Gaussian mixture distribution with parameters θ^p. Optimality for (8) is achieved if $\tilde{\mathbf{X}} \overset{\text{dist}}{=} \mathbf{X}$, i.e. $p_{f,\theta^p}(\mathbf{X}) = p^\star(\mathbf{X})$. In particular, by definition this is achieved for $f = f^\star$ since $\mathbf{X} \overset{\text{dist}}{=} f^\star(\mathbf{Z})$. Similarly, since $\mathbf{X} \overset{\text{dist}}{=} \left[f^\star \circ \begin{pmatrix} PD & 0 \\ 0 & N \end{pmatrix}\right] \begin{pmatrix} PD & 0 \\ 0 & N \end{pmatrix}^{-1} \mathbf{Z}$ and the family of Gaussian mixture distributions is invariant to linear transformations, we can further conclude that $f = f^\star \circ \begin{pmatrix} PD & 0 \\ 0 & N \end{pmatrix}$ for any permutation matrix P and diagonal matrix D is also an optimum of (8).

Since it is hard to maximize over p_{f,θ^p} directly, the VAE training approach uses a surrogate for the density p via the ELBO approximation. Specifically, recall that the ELBO is a lower bound for the log-likelihood:

$$\log\left(p^\star(\mathbf{X})\right) \geq \mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right). \tag{9}$$

Equality holds if the approximate posterior matches the true posterior, that is in the noiseless case, for any $x \in \mathcal{X}$:

$$q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X} = x) = p_f(\tilde{\mathbf{Z}}|\tilde{\mathbf{X}} = x) = \delta_{f^{-1}(x)}. \tag{10}$$

Here, we have appealed to one-to-one property of the function f by Assumption 1.1 and that the number of latents is specified correctly. Hence, with the choice of the posterior in (10), and setting $\tilde{\mathbf{X}} \overset{\text{dist}}{=} \mathbf{X}$, the maximization over f, θ^p of (8) and the reconstruction loss in (7) are equivalent.

We finally need to verify that ϕ^p in (10) leads to an optimal prediction in the classification term in (7). This follows by noting that $\mathbf{Y}|f^{-1}(\mathbf{X}) \overset{\text{dist}}{=} \mathbf{Y}|\mathbf{X}$ according to the graphical model in Figure 1(a), from the optimum (10) for q_{ϕ^p}, and Assumption 2. □

A.2.2. Proof of Lemma 1: maximizing $\mathcal{L}_{cl}$ identifies the true concepts

We analyze the following estimator in the infinite data limit

$$\begin{aligned}&\underset{\phi^{cl},\theta^{cl},f}{\operatorname{argmax}} \sum_{y\in\mathcal{Y}} \mathbb{P}_{\mathbf{Y}}(\mathbf{Y} = y)\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}|\mathbf{Y}=y}}[\mathcal{L}_{\texttt{c1}}(\phi^{cl}, \theta^{cl}, f; \mathbf{X}, \mathbf{Y})] \\ = &\underset{\phi^{cl},\theta^{cl},f}{\operatorname{argmax}} \sum_{y\in\mathcal{Y}} \mathbb{P}_{\mathbf{Y}}(\mathbf{Y} = y)\mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}|\mathbf{Y}}}\Big[\mathbb{E}_{q_{\phi^{cl}}}(\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y} = y)[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] \\ &\qquad - \mathrm{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y} = y), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y} = y)\right)\Big],\end{aligned} \tag{11}$$

where equality follows from the definition of $\mathcal{L}_{\text{c1}}$. Here, $\tilde{\mathbf{Z}}$ is an approximation for the underlying latent variables $\mathbf{Z}$ with $\tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right)$, for some vector μ_y and (diagonal) matrix $D_y \in \mathbb{R}^{k_c \times k_c}$ and a general matrix $G \in \mathbb{R}^{k_s \times k_s}$, altogether accumulated in the parameter θ^{cl}. Finally, we have that $p(\mathbf{X}|\tilde{\mathbf{Z}}) = \delta_{f(\tilde{Z})}$.

The proof of Lemma 1 relies on the following lemmas, which we state below and prove later.

Lemma 3. Let $a, b \in \mathbb{N}_+$. Suppose $Q\tilde{D} = DQ$ for orthogonal matrix $Q \in \mathbb{R}^{(a+b)\times(a+b)}$ and diagonal matrices $D, \tilde{D}$ where in the first a coordinates, D has unequal diagonal entries with no entry equal to one. Suppose that the last b entries of $D, \tilde{D}$ are equal to 1. Then, Q takes the following form: $Q = \begin{pmatrix} P & 0 \\ 0 & \tilde{Q} \end{pmatrix}$ where $P \in \mathbb{R}^{a\times a}$ is a permutation matrix and $\tilde{Q} \in \mathbb{R}^{b\times b}$ is an orthogonal matrix.

Lemma 4. The following two statements are equivalent:

1. the parameters $(\hat{\phi}^{cl}, \hat{f}, \{\hat{\mu}_y, \hat{D}_y, \hat{G}\}_{y\in\mathcal{Y}})$ are optimizers of (11).
2. for all $y \in \mathcal{Y}$, $\mathbf{X} \overset{\text{dist}}{=} \hat{f}(\tilde{\mathbf{Z}})$; $\tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\hat{\mu}_y, \begin{pmatrix} \hat{D}_y & 0 \\ 0 & \hat{G} \end{pmatrix}\right)$; $\hat{\phi} =$ parameters of $\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y} = y$.

With Lemmas 3 and 4 at hand, we are ready to prove Lemma 1.

Proof of Lemma 1. Lemma 4 states that in the noiseless case $\epsilon = 0$, we have $\mathbf{X} \overset{\text{dist}}{=} f^\star(\mathbf{Z}) \overset{\text{dist}}{=} \hat{f}(\tilde{\mathbf{Z}})$ for $\tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\hat{\mu}_y, \begin{pmatrix} \hat{D}_y & 0 \\ 0 & \hat{G} \end{pmatrix}\right)$. We now show that the set of all possible solutions for $\hat{f}$, denoted by $\mathcal{H}$, is restricted to maps of the form $f^\star \circ \begin{pmatrix} PD & 0 \\ 0 & G \end{pmatrix}$ for permutation and diagonal matrices of dimension $k_c \times k_c$ and general $k_s \times k_s$ matrix G.

Remember in the noiseless case $\mathbf{X} := f^\star(\mathbf{Z})$. We then have the following equality:

$$\begin{aligned}\mathcal{H} &= \left\{\text{continuous, one-to-one } f \;\middle|\; f^{-1}(\mathbf{X})|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right); D_y \text{ diagonal for all } y \in \mathcal{Y}\right\} \\ &\overset{(a)}{=} \left\{f^\star \circ g \text{ for a continuous, one-to-one } g \;\middle|\; g^{-1}(\mathbf{Z})|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right); D_y \text{ diagonal for all } y \in \mathcal{Y}\right\},\end{aligned}$$

where for every $y \in \mathcal{Y}$, $D_y \in \mathbb{R}^{k_c\times k_c}$ and $G \in \mathbb{R}^{k_s\times k_s}$. Here the relation $\overset{(a)}{=}$ follows from $f^\star$ being one-to-one and continuous from Assumption 1.1 as well as $\mathbf{X} \overset{\text{dist}}{=} f^\star(\mathbf{Z})$. We further have

$$\begin{aligned}\mathcal{H} &\overset{(b)}{=} \left\{f^\star \circ M \text{ for invertible matrix } M \;\middle|\; M^{-1}(\mathbf{Z})|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right); D_y \text{ diagonal for all } y \in \mathcal{Y}\right\} \\ &\overset{(c)}{=} \left\{f^\star \circ M \;\middle|\; M = \begin{pmatrix} D_y^\star & 0 \\ 0 & G^\star \end{pmatrix}^{1/2} Q_y \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}^{-1/2}; D_y \text{ diagonal}, Q_y \text{ orthogonal for all } y \in \mathcal{Y}\right\}.\end{aligned} \tag{12}$$

The relation $\overset{(b)}{=}$ follows from the fact that the set of one-to-one continuous operators that preserve Gaussianity are linear; the relation $\overset{(c)}{=}$ follows from $\mathbf{Z}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y^\star, \begin{pmatrix} D_y^\star & 0 \\ 0 & G_s^\star \end{pmatrix}\right)$ and the following calculations:

$$\begin{aligned} & M^{-1}\begin{pmatrix} D_y^\star & 0 \\ 0 & G^\star \end{pmatrix} M^{-T} = \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix} \\ \Leftrightarrow\ & \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}^{-1/2} M^{-1}\begin{pmatrix} D_y^\star & 0 \\ 0 & G^\star \end{pmatrix} M^{-T}\begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}^{-1/2} = \mathrm{Id} \\ \Leftrightarrow\ & \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}^{-1/2} M^{-1}\begin{pmatrix} D_y^\star & 0 \\ 0 & G^\star \end{pmatrix}^{1/2} \text{ is an orthogonal matrix.}\end{aligned} \tag{13}$$

Consider the pair $y, \tilde{y}$ satisfying Assumption 1.2. Then, since M doesn't depend on y, we have that:

$$\begin{pmatrix} D_y^\star & 0 \\ 0 & G^\star \end{pmatrix}^{1/2} Q_y \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}^{-1/2} = \begin{pmatrix} D_{\tilde{y}}^\star & 0 \\ 0 & G^\star \end{pmatrix}^{1/2} Q_{\tilde{y}} \begin{pmatrix} D_{\tilde{y}} & 0 \\ 0 & G \end{pmatrix}^{-1/2} . \tag{14}$$

Define the quantities:

$$A^\star := \begin{pmatrix} (D_y^\star)^{-1} D_{\tilde{y}}^\star & 0 \\ 0 & \mathrm{Id} \end{pmatrix} \quad ; \quad A := \begin{pmatrix} (D_y)^{-1} D_{\tilde{y}} & 0 \\ 0 & \mathrm{Id} \end{pmatrix} .$$

Then the relation (14) reduces to the following condition:

$$(A^\star)^{-1/2} Q_y A^{1/2} \text{ is an orthogonal matrix.} \tag{15}$$

The relation (15) leads to the conclusion:

$$Q_y A = A^\star Q_y \, . \tag{16}$$

Since by Assumption 1.2, the first k_c diagonal elements of $A^\star$ are distinct, $A, A^\star$ satisfy the assumptions of Lemma 3. This implies that $Q_y = \begin{pmatrix} P_y & 0 \\ 0 & \bar{Q}_y \end{pmatrix}$ for some permutation matrices P_y and orthogonal matrices $\bar{Q}_y$ and hence

$$\mathcal{H} = \left\{ f^\star \circ M \;\middle|\; M = \begin{pmatrix} PD & 0 \\ 0 & G \end{pmatrix} ; D \in \mathbb{R}^{k_c \times k_c} \text{ diagonal}, P \in \mathbb{R}^{k_c \times k_c} \text{ permutation matrix}, G \in \mathbb{R}^{k_s \times k_s} \right\},$$

which concludes the proof. □

Proof of Lemma 3. For convenience, we first decompose the matrices $Q = \begin{pmatrix} Q_1 & Q_{12} \\ Q_{21} & Q_2 \end{pmatrix}$ where $Q_1 \in \mathbb{R}^{k_c \times k_c}$ and similarly for $D, \tilde{D}$. We first show that $Q_{12} = 0$. The relation $Q\tilde{D} = DQ$ implies that for any pair of indices (i, j) either $[Q]_{ij} = 0$ or $[Q]_{ij} \neq 0 \;\&\; [D]_{ii} = [\tilde{D}]_{jj}$. Consider $i \in \{1, 2, \ldots, k_c\}$ and $j \in \{k_c + 1, k + 2, \ldots, k_c + k_s\}$. By the above relation and the assumption that the first k_c entries of D are not equal to one and the last k_s entries of $\tilde{D}$ are equal to one, we conclude that $[Q]_{ij} = 0$. Thus, we have established that $Q_{12} = 0$. Further, because $QQ^\top = I$, we have that $Q_{21} = 0$.

It remains to show that Q_1 is diagonal. First note that since Q is orthogonal, the matrix Q_1 must be orthogonal. The equality $Q\tilde{D} = DQ$ now implies that $Q_1 D_1 = \tilde{D}_1 Q_1$. In particular, $Q_1 D_1 Q_1^T = \tilde{D}_1$, that is $Q_1 D_1 Q_1^T$ is an eigen-decomposition of a diagonal matrix with distinct eigenvalues. By the uniqueness of eigen-decompositions, Q_1 must therefore be a permutation matrix. □

Proof of Lemma 4. Consider maximizing the reconstruction loss in (11). In this setting, the VAE model searches for an approximation $\hat{f}(\tilde{\mathbf{Z}}) \overset{\text{dist}}{\approx} \mathbf{X}$ where the parameters of the VAE model (e.g. posterior ϕ^{cl}, prior θ^{cl}, f) are optimized to yield the best approximation of $\mathbf{X}$. In other words, VAE training approximates the following optimization:

$$\underset{f, \theta^{cl}}{\operatorname{argmax}} \sum_{y \in \mathcal{Y}} \mathbb{P}_{\mathbf{Y}}(\mathbf{Y} = y) \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{\mathbf{X}|\mathbf{Y}=y}} [\log p_{f, \theta^{cl}}(\mathbf{X}|\mathbf{Y} = y)], \tag{17}$$

where the likelihood $p_{f,\theta^{cl}}(\mathbf{X})$ is defined with respect to the distribution $\tilde{\mathbf{X}} \overset{\text{dist}}{=} f(\tilde{\mathbf{Z}})$ with $\tilde{\mathbf{Z}}|\mathbf{Y} = y$ being a Gaussian distribution with parameters θ^{cl}. Optimality for (17) is achieved if $\tilde{\mathbf{X}} \overset{\text{dist}}{=} \mathbf{X}$, i.e. $p_{f,\theta^{cl}}(\mathbf{X}|\mathbf{Y} = y) = p^\star(\mathbf{X}|\mathbf{Y} = y)$. Since it is hard to maximize over $p_{f,\theta^{cl}}$ directly, the VAE training approach uses a surrogate for the density p via the ELBO approximation. Specifically, recall that the ELBO is a lower bound for the log-likelihood:

$$\log \left(p^\star(\mathbf{X}|\mathbf{Y} = y) \right) \geq \mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X})} [\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL} \left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y} = y), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y} = y) \right).$$

Equality holds if the approximate posterior matches the true posterior, that is in the noiseless case, for any $x \in \mathcal{X}$:

$$q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X} = x, \mathbf{Y} = y) = p_f(\tilde{\mathbf{Z}}|\tilde{\mathbf{X}} = x, \mathbf{Y} = y) = \delta_{f^{-1}(x)}. \tag{18}$$

where we have appealed to f being one-to-one from Assumption 1.1.. Hence, with the choice of the posterior in (18), and setting $\tilde{\mathbf{X}} \overset{\text{dist}}{=} \mathbf{X}$, the maximization over f, θ^{cl} of (17) and the reconstruction loss in (11) are equivalent. □

A.3. Proof of Theorem 1: miss-specified number of latent variables

In this section, we analyze the regularized estimator:

$$\max_{\phi^p,\theta^p,\phi^{cl},\theta^{cl},f,\psi} \hat{\mathbb{E}}_{\mathbf{X},\mathbf{Y}}[\mathcal{L}_p(\phi^p,\theta^p,f,\psi;\mathbf{X},\mathbf{Y})] + \hat{\mathbb{E}}_{\mathbf{X},\mathbf{Y}}[\mathcal{L}_{cl}(\phi^{cl},\theta^{cl},f;\mathbf{X},\mathbf{Y})] - \lambda_n \rho(f,\psi). \tag{19}$$

For our analysis, we define a few quantities. Let $f^\star_{\text{extend}}$ be one-to-one extension of $f^\star$ to a domain $\mathbb{R}^{\tilde{k}}$ where for any $v \in \mathbb{R}^{\tilde{k}}$:

$$f^\star_{\text{extend}} \circ B^\star(v) = f^\star \begin{pmatrix} v_{1:k_c} \\ v_{\tilde{k}_c+1:\tilde{k}_c+k_s} \end{pmatrix} \quad \text{where } B^\star = \begin{pmatrix} \mathrm{Id}_{k_c} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{Id}_{k_s} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{\tilde{k}\times\tilde{k}}.$$

Here, $\mathrm{Id}_{(\cdot)}$ denotes an identify matrix with its size specified in the subscript. Finally, we let $\mathbf{Z}_{\text{exend}} \in \mathbb{R}^{\tilde{k}}$ be a random variable that is identical to $\mathbf{Z}$ (true core and style features) in certain coordinates, and a standard Gaussian in other coordinates. Specifically:

$$\begin{pmatrix} \mathrm{Id}_{k_c} & 0 & 0 & 0 \\ 0 & 0 & \mathrm{Id}_{k_s} & 0 \end{pmatrix} \mathbf{Z}_{\text{extend}} = \mathbf{Z} \;\; ; \quad \begin{pmatrix} 0 & \mathrm{Id}_{\tilde{k}_c-k_c} & 0 & 0 \\ 0 & 0 & 0 & \mathrm{Id}_{\tilde{k}_s-k_s} \end{pmatrix} \mathbf{Z}_{\text{extend}} \text{ is standard Gaussian.}$$

In the infinite data limit with the regularization parameter λ_n tending to zero with larger sample size, the optimal parameters of (19) are solutions to :

$$\begin{aligned} \underset{f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl}}{\text{argmin}} \quad & \rho(f,\psi) \\ \text{subject-to} \quad & f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl} \in \underset{f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl}}{\text{argmax}} \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathcal{L}_{\mathrm{p}}(\phi^p,\theta^p,f,\psi;\mathbf{X},\mathbf{Y})] \\ & \qquad + \mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathcal{L}_{\mathrm{c1}}(\phi^p,\theta^p,f,\psi;\mathbf{X},\mathbf{Y})]. \end{aligned} \tag{20}$$

The proof of the corollary requires a few lemmas which we provide next and prove later.

Lemma 5. We have the following equivalence for a set of parameters $f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl}$:

$$\begin{aligned} f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl} \in & \underset{f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl}}{\text{argmax}} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\mathrm{p}}(\phi^p,\theta^p,f,\psi;\mathbf{X},\mathbf{Y})] + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\mathrm{c1}}(\phi^p,\theta^p,f,\psi;\mathbf{X},\mathbf{Y})] \\ & \Leftrightarrow \\ f,\phi^p,\theta^p,\phi^{cl},\theta^{cl} \in & \underset{\phi^p,\theta^p,\phi^{cl},\theta^{cl},f}{\text{argmax}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right] \\ & \qquad + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right] \\ \psi,\phi^p \in & \underset{\psi,\phi^p}{\text{argmax}}\, \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}}_c)]\right] \end{aligned}$$

Lemma 6. We have the following implication for the parameters $\phi^p, \theta^p, \phi^{cl}, \theta^{cl}, f$:

$$\phi^p, \theta^p, \phi^{cl}, \theta^{cl}, f \in \underset{\phi^p, \theta^p, \phi^{cl}, \theta^{cl}, f}{\operatorname{argmax}} \quad \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\mathbf{X}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right]$$
$$+ \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right]$$

$$\psi, \phi^p \in \underset{\psi, \phi^p}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}}_c)]\right]$$

$$\Rightarrow \quad \text{for every } y \in \mathcal{Y} \quad f = f' \circ B, \mathbf{X} \overset{\text{dist}}{=} f(\tilde{\mathbf{Z}}) \text{ where } \tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right) \text{ and}$$

$$D_y \in \mathbb{R}^{\tilde{k}_c \times \tilde{k}_c}, G \in \mathbb{R}^{\tilde{k}_s \times \tilde{k}_s} \quad : \quad q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}) = p(\tilde{\mathbf{Z}}|\mathbf{X}), q_{\phi^{\mathrm{cl}}}(\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y} = y) = p(\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y} = y).$$

Lemma 7. Consider the following optimization problem:

$$\begin{aligned}(f'_{\text{opt}}, B_{\text{opt}}) = \underset{f' \text{one-to-one}, B}{\operatorname{argmin}} \quad & \sum_{i=1}^{\tilde{k}} \mathbb{I}(\|B_{:,i}\|_2 > 0) \\ \text{subject-to} \quad & \text{there exists a random vector } \tilde{\mathbf{Z}} \in \mathbb{R}^{\tilde{k}} \text{ with } \mathbf{X} \overset{\text{dist}}{=} f' \circ B(\tilde{\mathbf{Z}}) \\ & \tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right) \text{for all } y \in \mathcal{Y}\end{aligned}$$

Then, the following statements hold:

1. Any feasible B has k (nonzero) linearly independent columns.
2. Any optimal B_{opt} has exactly k columns that are nonzero and linearly independent.
3. Letting $B_{\text{opt.reduced}}$ be those nonzero columns, then, $f'_{\text{opt}} \circ B_{\text{opt.reduced}} = f^\star_{\text{extend}} \circ \begin{pmatrix} PD & 0 & 0 & 0 \\ 0 & 0 & H & 0 \end{pmatrix}^T$ for some diagonal matrix $D \in \mathbb{R}^{k_c \times k_c}$, permutation matrix $P \in \mathbb{R}^{k_c \times k_c}$ and non-singular matrix $H \in \mathbb{R}^{k_s \times k_s}$.
4. Letting $\tilde{\mathbf{Z}}_{\text{reduced}}$ be the latent features corresponding to nonzero columns in B, $\tilde{\mathbf{Z}}_{\text{reduced}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D^{-1}P^{-1}D_y^\star P^{-T}D^{-1} & 0 \\ 0 & H^{-1}G^\star H^{-T} \end{pmatrix}\right)$ for all $y \in \mathcal{Y}$.

Proof of Theorem 1. Throughout the proof, we take $f = f' \circ B$ and $\psi = \psi' \circ C$. We consider the following reformulation of the minimal value of the optimization (20):

$$\begin{aligned} l^{(1)} := \min \quad & t \\ \text{subject-to} \quad & \rho(f, \psi) \le t \\ & f, \psi, \phi^p, \theta^p, \phi^{cl}, \theta^{cl} \in \underset{f, \psi, \phi^p, \theta^p, \phi^{cl}, \theta^{cl}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\mathrm{p}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})] \\ & \qquad + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\mathrm{c1}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})], \end{aligned}$$

where we denote the set of feasible parameters $(t, f, \psi, \phi^p, \theta^p, \phi^{cl}, \theta^{cl})$ of the optimization problem above by $\mathcal{S}^{(1)}$. Our objective is to show that any optimal encoder produces core features that are permutation and scaling of the true core features.

To that end, we consider the following optimization problem:

$$
\begin{aligned}
l^{(2)} := \min \quad & t \\
\text{subject-to} \quad & \rho(f,\psi) \le t \\
& f, \phi^p, \theta^p, \phi^{cl}, \theta^{cl} \in \underset{f,\phi^p,\theta^p,\phi^{cl},\theta^{cl}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right] \\
& \qquad\qquad + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right] \\
& \psi, \phi^p \in \underset{\psi,\phi^p}{\operatorname{argmax}} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}})]\right],
\end{aligned}
$$

where the set of feasible parameters $(t, f, \psi, \phi^p, \theta^p, \phi^{cl}, \theta^{cl})$ are denoted by $\mathcal{S}^{(2)}$. By Lemma 5, we have that $\mathcal{S}^{(2)} = \mathcal{S}^{(1)}$ so that $l^{(2)} = l^{(1)}$. We then relax the constraint set as follows:

$$
\begin{aligned}
l^{(3)} := \min \quad & t \\
\text{subject-to} \quad & \sum_{i=1}^{\tilde{k}} \mathbb{I}(\|B_{:,i}\|_2 > 0) \le t \\
& f, \phi^p, \theta^p, \phi^{cl}, \theta^{cl} \in \underset{f,\phi^p,\theta^p,\phi^{cl},\theta^{cl}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right] \\
& \qquad\qquad + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \mathrm{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right] \\
& \psi, \phi^p \in \underset{\psi,\phi^p}{\operatorname{argmax}} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}})]\right],
\end{aligned}
$$

where the set of feasible parameters $(t, f, \psi, \phi^p, \theta^p, \phi^{cl}, \theta^{cl})$ are denoted by $\mathcal{S}^{(3)}$. Evidently, $\mathcal{S}^{(3)} \supseteq \mathcal{S}^{(2)} \supseteq \mathcal{S}^{(1)}$ and thus $l^{(1)} \ge l^{(2)} \ge l^{(3)}$. Let $\mathcal{S}^{(3)}_{\text{opt}}$ be the optimal set of parameters in $\mathcal{S}^{(3)}$. Appealing to Lemma 6 and Lemma 7, the set $\mathcal{S}^{(3)}_{\text{opt}}$ given by:

$$
\begin{aligned}
\mathcal{S}^{(3)}_{\text{opt}} = \Bigg\{ & (t, f, \psi, \phi^p, \theta^p, \phi^{cl}, \theta^{cl}) \mid t = k, \text{there exists a random vector } \tilde{\mathbf{Z}} \in \mathbb{R}^{\tilde{k}} \text{ and matrices } P, D, H \text{ s.t.} \\
& \mathbf{X} \overset{\text{dist}}{=} f' \circ B\tilde{\mathbf{Z}} \;\;,\;\; f' \circ B_{\text{reduced}} = f^\star_{\text{extend}} \begin{pmatrix} PD & 0 & 0 & 0 \\ 0 & 0 & H & 0 \end{pmatrix}^T \text{ where } B_{\text{reduced}} \in \mathbb{R}^{\tilde{k}\times k} \\
& \theta^{cl} \text{ parameters of the Gaussian random vector } \tilde{\mathbf{Z}}|\mathbf{Y} = y \text{ where:} \\
& \quad \tilde{\mathbf{Z}}_{\text{reduced}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D^{-1}P^{-1}D_y^\star P^{-T}D^{-1} & 0 \\ 0 & H^{-1}G^\star H^{-T} \end{pmatrix}\right) \\
& \theta^p \text{ parameters of the distribution of } \tilde{\mathbf{Z}} \;\;,\;\; q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}) = q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}) = p(\tilde{\mathbf{Z}}|\mathbf{X}) \\
& \psi \in \operatorname{argmax} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}}_c)]\right] \Bigg\}.
\end{aligned}
\tag{21}
$$

Here, P is a $k_c \times k_c$ permutation matrix, D is a $k_c \times k_c$ diagonal matrix, and H is a $k_s \times k_s$ non-singular matrix. Furthermore, B_{reduced} is the k nonzero columns of B and $\tilde{\mathbf{Z}}_{\text{reduced}}$ are the components of $\tilde{\mathbf{Z}}$ corresponding to the nonzero columns of B. Take any optimal set of parameters in (21). Noting that $\mathbf{X} \overset{\text{dist}}{=} f^\star_{\text{extend}} B^\star \mathbf{Z}_{\text{extend}}$, it is straightforward to check that the first k_c components $\tilde{\mathbf{Z}}_{\text{reduced}}|\mathbf{X}$ are a permutation and linear scaling of $\mathbf{Z}_c$, and thus the posterior samples are optimally predictive. As such, one possibility for an optimal predictor is $\psi = \psi' \circ C$ where C has the same nonzero columns as B in (21). Notice that the resulting set of parameters is feasible in the set $\mathcal{S}^{(1)}$ and yield the objective value $t = k$. In other words, we have shown that $l^{(1)} = l^{(2)} = l^{(3)} = k$.

Now let $\mathcal{S}^{(1)}_{\text{opt}}$ be the optimal set of parameters associated with $l^{(1)}$. By Lemmas 5, 6 and 7, any optimal B should have at least k nonzero linearly independent columns. Thus, to attain the lower bound $l^{(1)} = k$, the optimal B should indeed only have k nonzero columns. This observation implies that the constraint $\sum_{i=1}^{\tilde{k}} \mathbb{I}(\|B_{:,i}\|_2 > 0) \leq t$ can be added to feasibility set $\mathcal{S}^{(1)}$ without changing the optimal value. Thus, we have concluded that $\mathcal{S}^{(1)}_{\text{opt}} \subseteq \mathcal{S}^{(4)}_{\text{opt}}$. Since any parameters in (21) lead to the posterior samples $\begin{pmatrix}\text{Id}_{k_c} & 0\end{pmatrix} \tilde{\mathbf{Z}}_{\text{reduced}}|\mathbf{X}$ that are a permutation and linear scaling of the samples of $\mathbf{Z}_c$, we have the desired result. □

We now prove the Lemmas 5,6, and 7 that were used in the proof of the corollary.

Proof of Lemma 5. The direction ← follows in a straightforward manner. For the direction →, we introduce some notation. Define:

$$h(f, \phi^p, \theta^p, \phi^{cl}, \theta^{cl}) := \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \text{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right]$$
$$+ \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \text{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right]$$
$$g(\phi^p, \psi) := \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_\psi(\mathbf{Y}|\tilde{\mathbf{Z}})]\right]$$

Note that $\mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\texttt{p}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})] + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\texttt{c1}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})] = h(f, \phi^p, \theta^p, \phi^{cl}, \theta^{cl}) + g(\phi^p, \psi)$. Thus, we have the following inequality:

$$\begin{aligned}&\max_{f,\psi,\phi^p,\theta^p,\phi^{cl},\theta^{cl}} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\texttt{p}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})] + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}[\mathcal{L}_{\texttt{c1}}(\phi^p, \theta^p, f, \psi; \mathbf{X}, \mathbf{Y})] \\ &\leq \max_{\phi^p,\theta^p,\phi^{cl},\theta^{cl},f} h(f, \phi^p, \theta^p, \phi^{cl}, \theta^{cl}) + \max_{\psi,\phi^p} g(\phi^p, \psi).\end{aligned} \tag{22}$$

Consider the parameters in (21). Notice that they optimize each term in the right hand side of the inequality above and are feasible in the optimization problem in left hand side of the inequality. Thus, the inequality in (22) is actually an equality. Let t_{opt} be the optimal value of either side of the equality.

Suppose for a proof of contradiction that the direction → is not valid. In other words, consider a set of maximizers $(f_{\text{opt}}, \psi_{\text{opt}}, \phi^p_{\text{opt}}, \theta^p_{\text{opt}}, \phi^{cl}_{\text{opt}}, \theta^{cl}_{\text{opt}})$ for the left hand side of the equation above that are not maximal in either term in the right hand side. Then:

$$\begin{aligned}t_{\text{opt}} &= h(f_{\text{opt}}, \phi^p_{\text{opt}}, \theta^p_{\text{opt}}, \phi^{cl}_{\text{opt}}, \theta^{cl}_{\text{opt}}) + g(\phi^p_{\text{opt}}, \psi_{\text{opt}}) \\ &< \max_{\phi^p,\theta^p,\phi^{cl},\theta^{cl},f} h(f, \phi^p, \theta^p, \phi^{cl}, \theta^{cl}) + \max_{\psi,\phi^p} g(\phi^p, \psi).\end{aligned}$$

This however contradicts the fact that the inequality (22) is an equality. □

Proof of Lemma 6. We have the following inequality:

$$\begin{aligned}&\operatorname*{argmax}_{f,\phi^p,\theta^p,\phi^{cl},\theta^{cl}} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \text{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right] \\ &\quad + \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \text{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right] \\ &\leq \underbrace{\operatorname*{argmax}_{\phi^p,\theta^p,f} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}\left[\mathbb{E}_{q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \text{KL}\left(q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}), p_{\theta^p}(\tilde{\mathbf{Z}})\right)\right]}_{\text{Term 1}} \\ &\quad + \underbrace{\operatorname*{argmax}_{\phi^{cl},\theta^{cl},f} \mathbb{E}_{\mathbf{X},\mathbf{Y}\sim\mathbb{P}_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y})}[\log p_f(\mathbf{X}|\tilde{\mathbf{Z}})] - \text{KL}\left(q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X},\mathbf{Y}), p_{\theta^{cl}}(\tilde{\mathbf{Z}}|\mathbf{Y})\right)\right]}_{\text{Term 2}}.\end{aligned}$$

Notice that the different terms in the equation above have the common parameters f. Thus, the inequality in the equation above is an equality if there exists an f that is optimal for each of the terms in the relation above. Furthermore, from proof of Lemma 2, Term 1 is optimized when $\mathbf{X} \overset{\text{dist}}{=} f(\tilde{\mathbf{Z}})$ for $\tilde{\mathbf{Z}}$ being a mixture Gaussian and $q_{\phi^p}(\tilde{\mathbf{Z}}|\mathbf{X}) = p(\tilde{\mathbf{Z}}|\mathbf{X})$. From Lemma 4, Term 2 is maximized when $\mathbf{X} \overset{\text{dist}}{=} f(\tilde{\mathbf{Z}})$ for $\tilde{\mathbf{Z}}|\mathbf{Y} = y$ being a Gaussian distribution with appropriate covariance matrix. Consider the parameters in (21); they satisfy the properties above for the same decoder.

It remains to check that there exists a set of parameters in the right hand implication of Lemma 6 that are optimally predictive. This follows from taking the parameters (21) and noting that that the posterior samples $\tilde{\mathbf{Z}}|\mathbf{X}$ are optimally predictive. □

Proof of Lemma 7. By definition, any feasible $f = f' \circ B$ satisfies:

$$\mathbf{X} \overset{\text{dist}}{=} f' \circ B(\tilde{\mathbf{Z}}) \text{ and } \tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right) \text{ where } D_y \in \mathbb{R}^{\tilde{k}_c \times \tilde{k}_c} \text{ and } G \in \mathbb{R}^{\tilde{k}_s \times \tilde{k}_s}.$$

Proof of 1. Applying $(f^\star_{\text{extend}})^{-1}$ to both sides of the relation $\mathbf{X} \overset{\text{dist}}{=} f' \circ B\tilde{\mathbf{Z}}$ and noting that f' can be expressed as $f' = f^\star_{\text{extend}} \circ g$ for a continuous and one-to-one function g, we have:

$$B^\star \mathbf{Z}_{\text{extend}} \overset{\text{dist}}{=} g \circ B\tilde{\mathbf{Z}}.$$

Since $\mathbf{Z}|\mathbf{Y} = y$ and $\tilde{\mathbf{Z}}|\mathbf{Y} = y$ are Gaussian and g is a continuous and one-to-one function, g must be a linear map; we denote g by the matrix $N \in \mathbb{R}^{\tilde{k} \times \tilde{k}}$. It is straightforward then to argue that B must have $k = k_c + k_s$ total number of linearly independent columns; these linearly independent columns are nonzero by definition. We have proven the first item in Lemma 7.

Proof of 2. Gathering all the facts so far, we have that for any feasible $f = f' \circ B$:

$$f' = f^\star_{\text{extend}} \circ N \quad ; \quad B^\star \mathbf{Z}_{\text{extend}} \overset{\text{dist}}{=} NB\tilde{\mathbf{Z}} \; ; \; \text{rank}(B) = k_c + k_s. \tag{23}$$

Furthermore, by the objective of (20), we have that an optimal B must have exactly $k = k_c + k_s$ total number of linearly independent nonzero columns (i.e. k of the $\tilde{k}$ latent features have some visualization power). We have thus concluded the second item in the lemma.

Proof of 3. Let $B_{\text{reduced}} \in \mathbb{R}^{\tilde{k} \times k}$ be the non-zero columns of B, so that $B\tilde{\mathbf{Z}} = B_{\text{reduced}}\tilde{\mathbf{Z}}_{\text{reduced}}$ where $\tilde{\mathbf{Z}}_{\text{reduced}} \in \mathbb{R}^k$ and for every $y \in \mathcal{Y}$:

$$\tilde{\mathbf{Z}}_{\text{reduced}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right).$$

Here, D_y is a diagonal matrix. Without loss of generality, we assume that every diagonal entry in D_y has some variation across $y \in \mathcal{Y}$; otherwise, we can concatenate the components that do not vary to the general matrix G. The dimension of D_y and G corresponds to the number of core and style features that are selected by the nonzero columns of B. Note that so far, we have only established that the dimensions of D_y and G sum up to k. In what follows we show that the numbers of estimated core and style features equal to k_c, k_s.

Let M be the $k \times k$ matrix $M := \begin{pmatrix} \text{Id}_{k_c} & 0 & 0 & 0 \\ 0 & 0 & \text{Id}_{k_s} & 0 \end{pmatrix} NB_{\text{reduced}}$. By relation (23), $\mathbf{Z} \overset{\text{dist}}{=} M\tilde{\mathbf{Z}}_{\text{reduced}}$. Since the distribution of $\mathbf{Z}$ is non-degenerate, M is a non-singular matrix. Therefore, for every $y \in \mathcal{Y}$ we have

$$M^{-1}\mathbf{Z}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right).$$

Following a similar analysis as in the proof of Lemma 1, we conclude that for every $y \in \mathcal{Y}$, there exists an orthogonal matrix Q_y such that

$$M = \begin{pmatrix} D^\star_y & 0 \\ 0 & G^\star \end{pmatrix}^{1/2} Q_y \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}^{-1/2}. \tag{24}$$

Choosing $y, \tilde{y} \in \mathcal{Y}$ that satisfy Assumption 1.2, relation (24) implies that

$$A^{\star -1/2} Q_y A^{1/2} \text{ is an orthogonal matrix.} \tag{25}$$

Here, $A^\star := \begin{pmatrix} A_1^\star & 0 \\ 0 & A_2^\star \end{pmatrix}$ where $A_1^\star = D_y^\star [D_{\tilde{y}}^\star]^{-1}$ and $A_2^\star = \text{Id}$ and $A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$, where $A_1 = D_y D_{\tilde{y}}^{-1}$ and $A_2 = \text{Id}$. Relation (25) implies that:

$$Q_y A = A^\star Q_y.$$

Notice that the singular values of $Q_y A$ are equal to the singular values of A, since product by orthogonal matrices preserves the singular values. Similarly, singular values of $A^\star Q_y$ are equal to the singular values of $A^\star$. Thus, $A^\star$ and A have the same singular values. In other words, since $A^\star$ has k_s singular values equal to one and k_c singular values not equal to one (by Assumption 1.2), A_1 must have exactly k_c diagonal elements not equal to one. This allows us to conclude that the dimensions of D_y are greater than or equal to k_c.

From the analysis above, we can partition the matrix A as follows: $\begin{pmatrix} \tilde{A}_1 & 0 \\ 0 & \text{Id} \end{pmatrix}$ where $\tilde{A}_1$ is a diagonal matrix of dimension $k_c \times k_c$ with all distinct entries. Appealing to Lemma 3, we then conclude that $Q_y = \begin{pmatrix} P & 0 \\ 0 & H \end{pmatrix}$ for a permutation matrix $P \in \mathbb{R}^{k_c \times k_c}$, diagonal matrix $D \in \mathbb{R}^{k_c \times k_c}$ and a non-singular matrix $H \in \mathbb{R}^{k_s \times k_s}$. Combining this with the expression of M (24) and the fact that the dimensions of D_y are greater than or equal to k_c, we conclude that M takes the form:

$$M = \begin{pmatrix} PD & 0 \\ 0 & H \end{pmatrix}.$$

Finally, combining the relation above with (23) and the fact that $\begin{pmatrix} 0 & \text{Id}_{\tilde{k}_c - k_c} & 0 & 0 \\ 0 & 0 & 0 & \text{Id}_{\tilde{k}_s - k_s} \end{pmatrix} NB = 0$ yields the third item of the lemma.

Proof of 4. The final component of the lemma also follows from the relation $\mathbf{Z} \overset{\text{dist}}{=} M \tilde{\mathbf{Z}}_{\text{reduced}}$. □

A.4. Analysis of $\mathcal{L}_{cl}$ in the noisy setting

In this section, our objective is to show that the density $q_{\phi^{cl}}(\tilde{\mathbf{Z}}|\mathbf{X}, \mathbf{Y})$ matches the density $p^\star(PD\mathbf{Z}_c|\mathbf{X}, \mathbf{Y})$ for some $k_c \times k_c$ permutation matrix and $k_c \times k_c$ diagonal matrix. Throughout the following discussion, $p^\star(\cdot)$ represents the ground truth density corresponding to a specified random variable.

In addition to Assumptions 1, analysis of the noisy setting requires the following mild assumption:

Assumption 3 : the Fourier transform of the density of ϵ is non-negative everywhere.

It is straightforward to extend the characterization in Lemma 4 to the noisy case and conclude that $\mathbf{X} \overset{\text{dist}}{=} f^\star(\mathbf{Z}) + \epsilon \overset{\text{dist}}{=} f(\tilde{\mathbf{Z}}) + \epsilon$ where $\tilde{\mathbf{Z}}|\mathbf{Y} = y \sim \mathcal{N}\left(\mu_y, \begin{pmatrix} D_y & 0 \\ 0 & G \end{pmatrix}\right)$. Here, D_y is a $k_c \times k_c$ diagonal matrix and G is a $k_s \times k_s$ matrix. Since $\tilde{\mathbf{Z}} \perp\!\!\!\perp \epsilon$, $\mathbf{Z} \perp\!\!\!\perp \epsilon$, we have that:

$$\mathcal{F}[p(f^\star(\mathbf{Z}) + \epsilon)] = \mathcal{F}[p(\epsilon)]\mathcal{F}[p(f^\star(\mathbf{Z}))]$$
$$\mathcal{F}[p(f(\tilde{\mathbf{Z}}) + \epsilon)] = \mathcal{F}[p(\epsilon)]\mathcal{F}[p(f(\tilde{\mathbf{Z}}))],$$

where $\mathcal{F}[\cdot]$ represents the Fourier transform and $p(\cdot)$ represents the density function with respect to a random variable. Since $p(f^\star(\mathbf{Z}) + \epsilon) = p(f(\tilde{\mathbf{Z}}) + \epsilon)$, appealing to Assumption 2, we have that $\mathcal{F}[p(f^\star(\mathbf{Z}))] = \mathcal{F}[p(f(\tilde{\mathbf{Z}}))]$, or equivalently, $f(\tilde{\mathbf{Z}}) \overset{\text{dist}}{=} f^\star(\mathbf{Z})$. We then have from the chain of equalities in (A.2.2) that $f = f^\star \circ \begin{pmatrix} PD & 0 \\ 0 & N \end{pmatrix}^{-1}$. Combining this with the fact that $f(\tilde{\mathbf{Z}}) \overset{\text{dist}}{=} f^\star(\mathbf{Z})$, we conclude that $\tilde{\mathbf{Z}} \overset{\text{dist}}{=} \begin{pmatrix} PD & 0 \\ 0 & N \end{pmatrix} \mathbf{Z}$. Notice that:

$$p(\tilde{z}_c|x, y) = \int p(\tilde{z}_c, \tilde{z}_s|x, y) \partial z_s \overset{(a)}{=} \frac{\int p_f(x|\tilde{z}_c, \tilde{z}_s) p(\tilde{z}_c|y) p(\tilde{z}_s) \partial \tilde{z}_s}{p^\star(x|y)} \overset{(b)}{=} \frac{\int p_f\left(x|\tilde{z}_c; N z_s\right) p(\tilde{z}_c|y) p^\star(z_s) \partial z_s}{p^\star(x|y)}, \tag{26}$$

The equality $\overset{(a)}{=}$ follows from $\mathbf{Z}_c \perp\!\!\!\perp \mathbf{Z}_s$ and the equality $\overset{(b)}{=}$ follows from the change of variables $Nz_s \leftarrow \tilde{z}_s$.

Now we examine the posterior density of $PD\mathbf{Z}_c|\mathbf{X}, \mathbf{Y}$. Appealing to the same line of reasoning as (26), we have that:

$$p^\star(z_c|x,y) = \int p^\star(z_c, z_s|x,y)\partial z_s \overset{(a)}{=} \frac{\int p^\star(x|z_c,z_s)p^\star(z_c|y)p^\star(z_s)\partial z_s}{p^\star(x|y)} \overset{(b)}{=} \frac{\int p^\star(x|D^{-1}P^{-1}z_c; z_s)p^\star(z_c|y)p^\star(z_s)\partial z_s}{p^\star(x|y)}$$
$$\overset{(c)}{=} \frac{\int p_f\left(x|z_c; Nz_s\right)p^\star(z_c|y)p^\star(z_s)\partial z_s}{p^\star(x|y)} \overset{(d)}{=} \frac{\int p_f\left(x|z_c; Nz_s\right)p(\tilde{z}_c|y)p^\star(z_s)\partial z_s}{p^\star(x|y)}, \tag{27}$$

where $p^\star(x|z_c, z_s)$ is the density of the distribution $\mathbf{X}|PD\mathbf{Z}_c, \mathbf{Z}_s$ and $p^\star(z_c|y)$ is the density of the distribution $PD\mathbf{Z}_c|\mathbf{Y}$. Here, the equality $\overset{(a)}{=}$ follows from $PD\mathbf{Z}_c \perp\!\!\!\perp \mathbf{Z}_s$, $\overset{(b)}{=}$ follows from the density $\mathbf{X}|(PD\mathbf{Z}_c, \mathbf{Z}_s) = p^\star(\mathbf{X}|(D^{-1}P^{-1}\mathbf{Z}_c; \mathbf{Z}_s))$, and $\overset{(c)}{=}$ follows from the relationship between f and $f^\star$. Finally, $\overset{(d)}{=}$ follows from the equality $p^\star(z_c|y) = p(\tilde{z}_c|y)$ since $PD\mathbf{Z}_c \overset{\text{dist}}{=} \tilde{\mathbf{Z}}_c$ (due to the relation between $\tilde{\mathbf{Z}}, \mathbf{Z}$ and that they are both Gaussian random variables after conditioning on a label). Comparing (26) and (27), we have the desired result.

B. Comparisons with (Khemakhem et al., 2020)

Previously, (Khemakhem et al., 2020) proved that supervision enables identifiability of the latent features. The only similarity of our results with this work can be found in Lemma 1, although our guarantees distinguish core and style features and allow for more relaxed assumptions. In particular, while the target label can take two distinct values for Assumption 2 to be satisfied, the assumption in (Khemakhem et al., 2020) – when specialized to the Gaussian prior – requires that the target label takes at least $2k$ distinct values where k is the number of the latent features. On all the remaining theoretical and methodological aspects, our setting differs substantially from (Khemakhem et al., 2020). Importantly, while our method can be employed for interpretable predictions, their proposed I-VAE is simply not applicable in our setting, as it requires labels as inputs and thus cannot perform prediction. Furthermore, our methodology and theoretical guarantees (see Theorem 1) covers the case of overparameterized latent spaces, i.e. the case in which the model allows for mode latent features than the ground truth ones. This is another practically relevant novelty, as in general the number of ground truth features is unknown.

C. Implementation details

C.1. Datasets and pre-processing

For all the datasets, pixels are transformed to have values between 0 and 1 and the image size of 64x64 is kept. We randomly fix train and test set with sizes respectively of 60% and 40% for MPI3D, 90% and 10% for shapes3D, SmallNORB and Plantvillage datasets. For the ChestXRay dataset, we use the pre-defined train test splits.

C.2. Hyperparameter selection

Hyperparameter selection been performed via visual inspection of the traversals *on the training set*. This is a correct validation of the algorithm, since no test data has been utilized for model selection. Furthermore, it reflects the procedure that we propose in the paper, where model selection is carried out with a human expert. All the traversals and results we report are then obtained on the test set as usual.

C.3. Hyper-parameters and training configuration

Most hyper-parameters and training configurations are kept fixed across datasets. An overview of hyper-parameter settings is shown in Table 2. We set a fixed dimension of $(\mathbf{Z}_c, \mathbf{Z}_s) = (10, 20)$. Notice that, similar to the real-life setting where the exact dimension of the true underlying latent features is unknown we allow the latent dimension of core and style latent features to be higher than theoretically needed for all the synthetic datasets.
We set the following values of (prediction term weight, group sparsity regularization) for the experiments: (50, 0.05) for MPI3D; (10, 0.0001) for Shapes3D; (50, 0.01) for SmallNORB; (200, 0.01) for PlantVillage; (200, 0.05) for ChestXRay.

Table 2: Hyperparameter settings for all datasets.

Hyperparameter	Value
Core latent space dimension	10
Style latent space dimension	20
Batch size	132
Optimizer	ADAM
Learning rate	5e-4
Decoder type	Bernoulli
Prediction loss	binary cross entropy
Training steps	$150,000$

C.4. *CLAP* model architecture

The architecture of *CLAP* consists of five main modules that share weights where appropriate. The five modules are: the encoder, decoder and predictor in the prediction VAE, and the encoder and decoder in the concept-learning VAE.

Predictor The predictor in the prediction VAE is a simple linear predictor mapping from core latent features to the labels where each label has a separate linear predictor.

Decoders The decoders in both the prediction VAE and the concept-learning VAE share weights with an architecture as depicted on the right in Table 3.

Encoders The encoder in the prediction VAE consists of a backbone with architecture as depicted on the left in Table 3 and two parallel fully connected layers. The output of the backbone is fed into the two separate fully connected layers where one is learning the posterior distribution of core latent features and one is learning the posterior distribution of style latent features.

The encoder in the concept-learning VAE shares all weights that are associated with style latent features with the encoder of the prediction VAE. In particular, the weights of the backbone and fully connected layer that output the posterior distribution of style latent features are shared across both. No weights are shared for that part of the concept-learning VAE associated with core latent features. Thus, a separate backbone and fully connected layer are used to learn the posterior distribution of core latent features in the concept-learning VAE. The label y is incorporated in the concept-learning VAE by feeding it jointly with the output of the backbone to the fully connected layer that outputs the posterior distribution of the core latent features.

C.5. SENN and CCVAE model architecture

For SENN, we employ the same architecture as for *CLAP*. In particular, we utilize the encoder backbone presented in Table 3, and map the encoding obtained from the backbone to the core and style features via two parallel linear layers. The decoder utilized is the same. Furthermore, SENN employs an additional mapping from the input $\mathbf{X}$ to the prediction weights utilized on top of the core features. This mapping is given by a neural network with structure `Conv(32, 4, 2, 0), MaxPool(2, 2), ReLU, Conv(32, 4, 2, 0), MaxPool(2, 2), ReLU`. To map to the prediction weights, we flatten and then utilize two linear layers with output dimension 288 and 128 and activations ReLU and Tanh respectively.

For CCVAE, we employ the publicly available architecture from the authors at https://github.com/thwjoy/ccvae. We note that the encoder-decoder pair is roughly equivalent to that utilized for *CLAP*.

Table 3: Encoder backbone and decoder architecture. **Abbreviations:** c denotes the channel size of the input image, $\tilde{k}_c$ the dimension of core latent factors, and $\tilde{k}_s$ the dimension of style latent factors. **Layer parameters:** For fully connected layers the first parameter denotes input dimension, and the second parameter output dimension. For LeakyReLU the parameter denotes its negative slope. For dropout layers the parameter denotes the probability that a whole channel is dropped out (2D dropout). For convolutional and transposed convolutional layers the parameters can be interpreted as follows: output channel size, kernel size, stride, padding.

Encoder backbone		**Decoder**	
Input size:	64 x 64 x c	Input size:	$30 = \tilde{k}_c + \tilde{k}_s$
Output size:	256	Output size:	64 x 64 x c
Layer	**Parameters**	**Layer**	**Parameters**
Conv	64, 3, 2, 1	FC	30, 512
LeakyReLU	0.01	ReLU	
Dropout	0.1	FC	512, 1024
Conv	64, 3, 2, 1	ConvTranspose	64, 3, 2, 0
LeakyReLU	0.01	ReLU	
Dropout	0.1	ConvTranspose	64, 3, 2, 1
Conv	64, 3, 2, 1	ReLU	
LeakyReLU	0.01	ConvTranspose	64, 3, 2, 1
Dropout	0.1	ReLU	
Conv	64, 3, 2, 1	ConvTranspose	c, 4, 2, 2
LeakyReLU	0.01		
Dropout	0.1		
Flatten			
FC	1024, 256		

D. Details on synthetic datasets

D.1. MPI3D traversals and details

The MPI3D dataset is an artificial dataset of images where the ground truth factors of variation are object color, shape and size, as well height of the camera, background color, horizontal and vertical axes of the camera. In Figure 5, we present some example images from the dataset.

We create synthetic labels according to the following rules: for the first label, $y = 1$ if color in $\{$white, green, brown, olive$\}$ and shape in $\{$cone, cube, cylinder, sphere$\}$ and size in $\{$small$\}$, and $y = 0$ otherwise; for the second label, $y = 1$ if color in $\{$green, red, blue$\}$ and size in $\{$large$\}$, and $y = 0$ otherwise; for the third label, $y = 1$ if shape in $\{$cone, pyramid$\}$, and $y = 0$ otherwise; for the final label, $y = 1$ if shape in $\{$cylinder, hexagonal, pyramid$\}$, and $y = 0$ otherwise.

We present additional MPI3D traversals of *CLAP* for $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ (second row) in Figure 6.

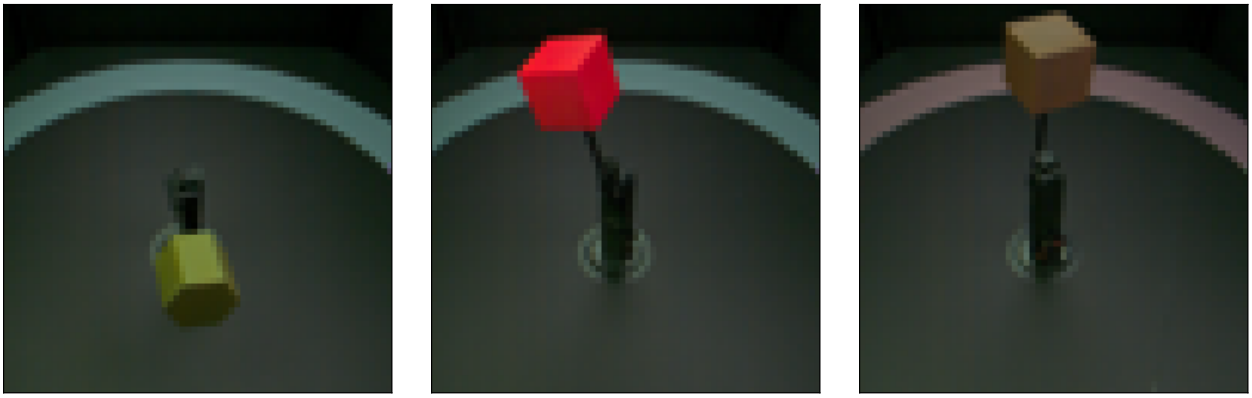


Figure 5: Some example images from the MPI3D dataset.

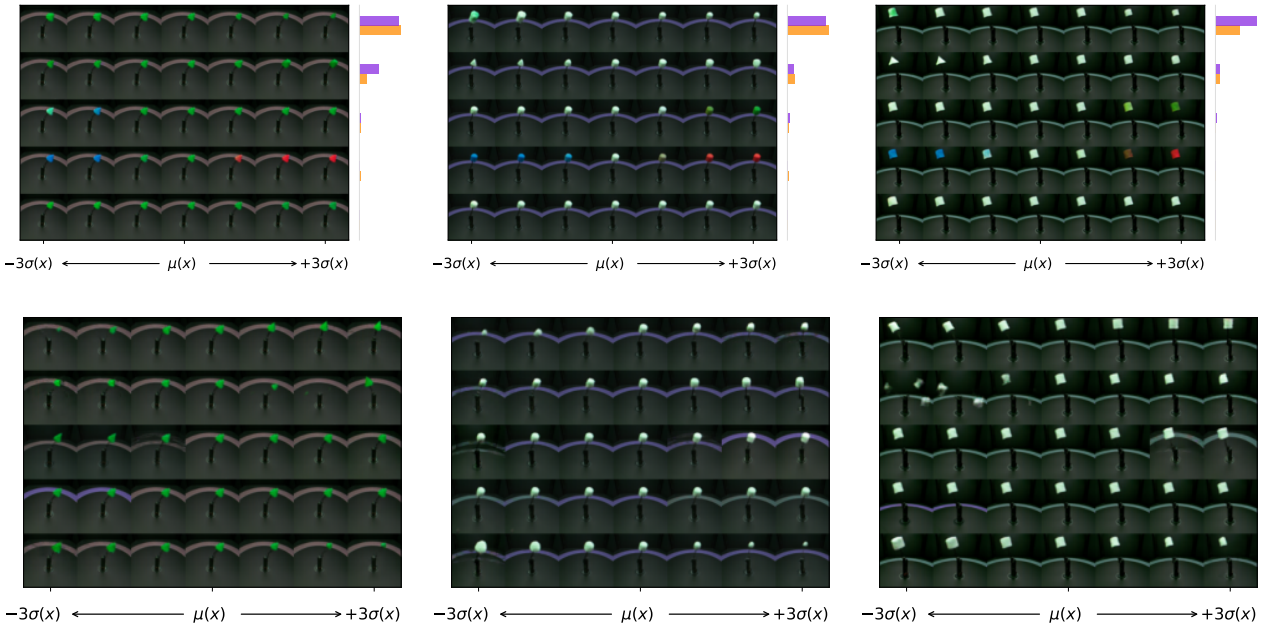


Figure 6: *CLAP* traversals of $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ on MPI3D dataset.

D.2. Shapes3D

The Shapes3D dataset is composed of synthetic images of different 3D objects. The ground truth factors of variations are floor hue, background wall hue, orientation, and the object's hue, scale, and shape. In Figure 7, we present some example images from the dataset.

We rescale all the factors of variation, which are already discrete, to take integer values starting from 0. In particular, hue and scale of the object have values in $[0, 9]$ and are used to create the synthetic labels. We create synthetic labels according to the following rules: for the first label, $y = 1$ if scale ≤ 5 and hue ≥ 3, $y = 0$ otherwise; for the second label, $y = 1$ if scale ≥ 3 and hue ≥ 3, $y = 0$ otherwise; for the second label, $y = 1$ if scale ≤ 4 and hue ≥ 2, $y = 0$ otherwise; for the final label, $y = 1$ if scale ≥ 5, $y = 0$ otherwise.

We present *CLAP* traversals on the Shapes3D dataset for $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ (second row) in Figure 8.

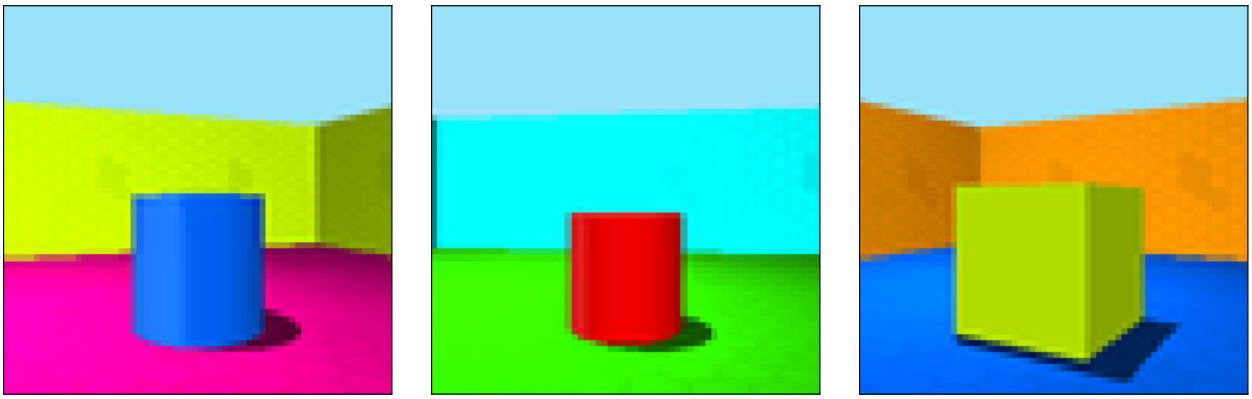


Figure 7: Some example images from the Shapes3D dataset.

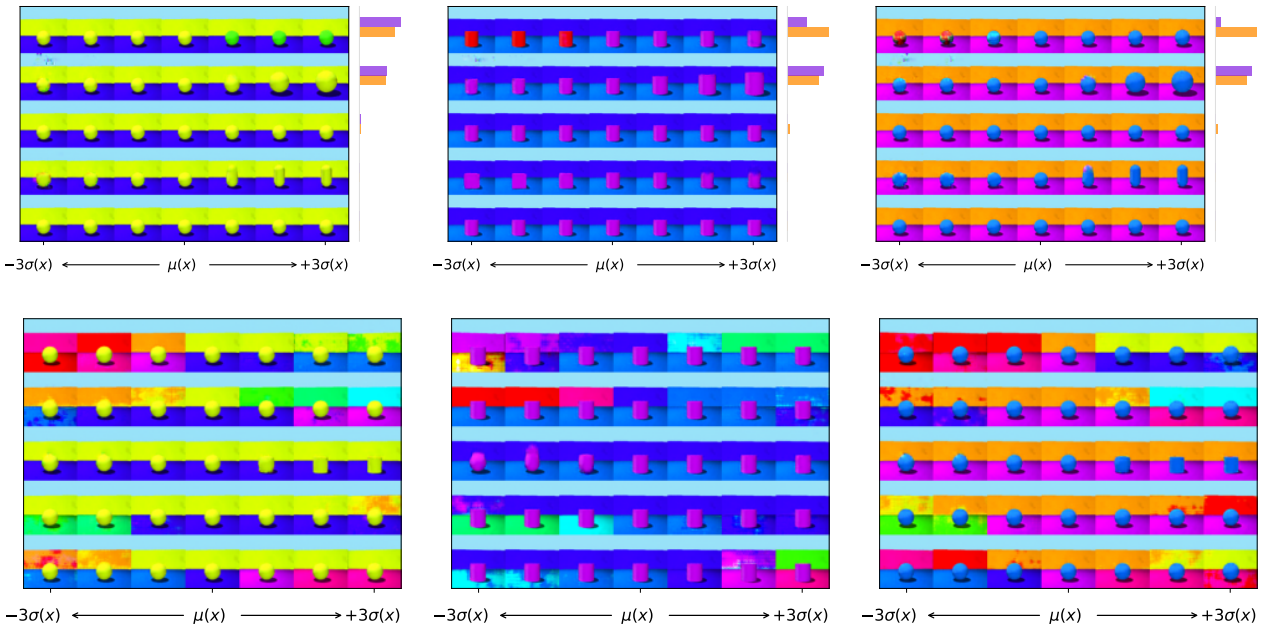


Figure 8: *CLAP* traversals of $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ on the Shapes3D dataset.

D.3. SmallNORB

The SmallNORB dataset is a dataset of black and white images. In Figure 9, we present some example images from the dataset.

The ground truth factors of variation are the object (9 classes), elevation of the camera (0 to 8), azimuth (even values from 0 to 34) and lightning condition (0 to 5). We create synthetic labels according to the following rules: for the first label, $y = 1$ if object type ≥ 5 and lightning ≥ 3, and $y = 0$ otherwise; for the second label, $y = 1$ if object type ≥ 5 and lightning < 3, and $y = 0$ otherwise; for the third label, $y = 1$ if object type < 5 and lightning ≥ 3, and $y = 0$ otherwise; for the final label, $y = 1$ if lightning < 3, and $y = 0$ otherwise.

We present *CLAP* traversals on the SmallNORB dataset for $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ (second row) in Figure 10.

Figure 9: Some example images from the Shapes3D dataset.

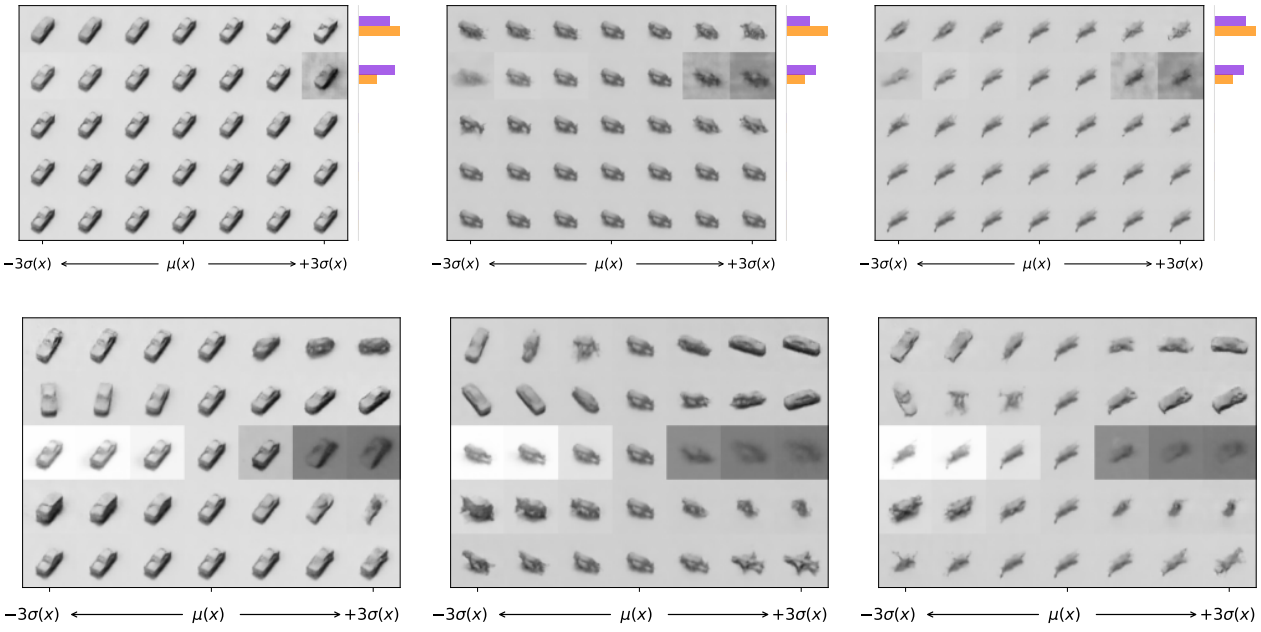


Figure 10: *CLAP* traversals of $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ on the SmallNORB dataset.

E. Ablation studies

E.1. No Group Sparsity

In Figure 11 we present traversals for *CLAP* trained without group sparsity, i.e. $\rho_n(f, \psi) = 0$.

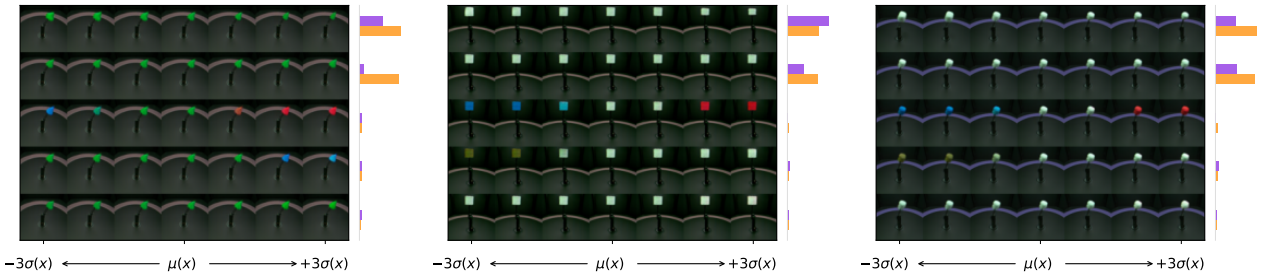


Figure 11: No-group-sparsity *CLAP* traversals of $\mathbf{Z}_c$ on the MPI3D dataset.

E.2. Prediction-Only Model

In Figure 12 we present traversals for *CLAP* trained only on the prediction part of the loss in Eq. 3.

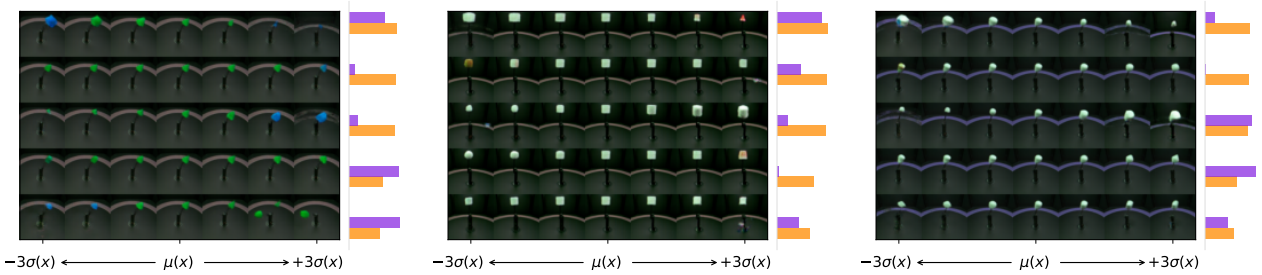


Figure 12: Prediction-only *CLAP* traversals of $\mathbf{Z}_c$ on the MPI3D dataset.

E.3. MPI3D with One Label Only

In Figure 13 we present traversals for *CLAP* trained on the MPI3D dataset where only the first label is made available for supervision.

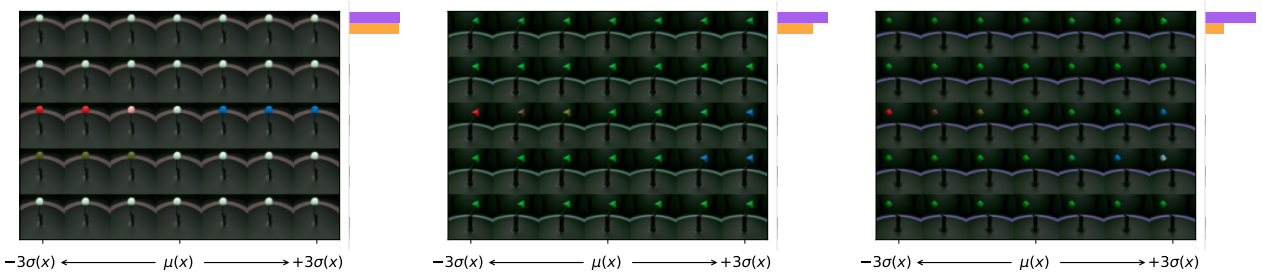


Figure 13: Prediction-only *CLAP* traversals of $\mathbf{Z}_c$ on the MPI3D dataset with only one label.

F. CCVAE on Shapes3D

In Figure 14 we show the traversals of CCVAE on the Shapes3D dataset.

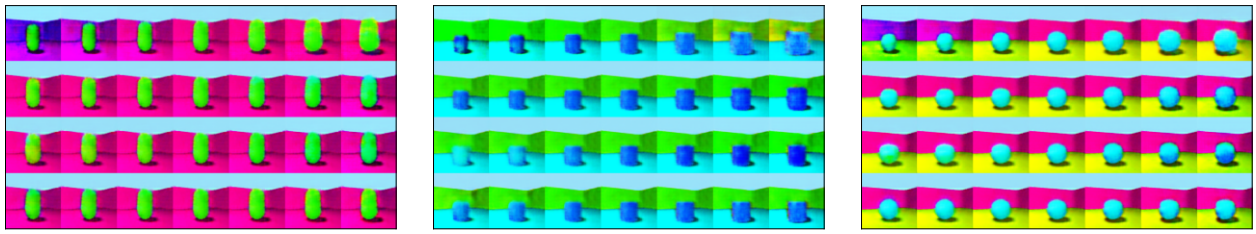


Figure 14: CCVAE traversals of $\mathbf{Z}_c$ on the Shapes3D dataset.

G. ChestXRay traversals

We include the complete *CLAP* traversals on the ChestXRay dataset in Figure 15.

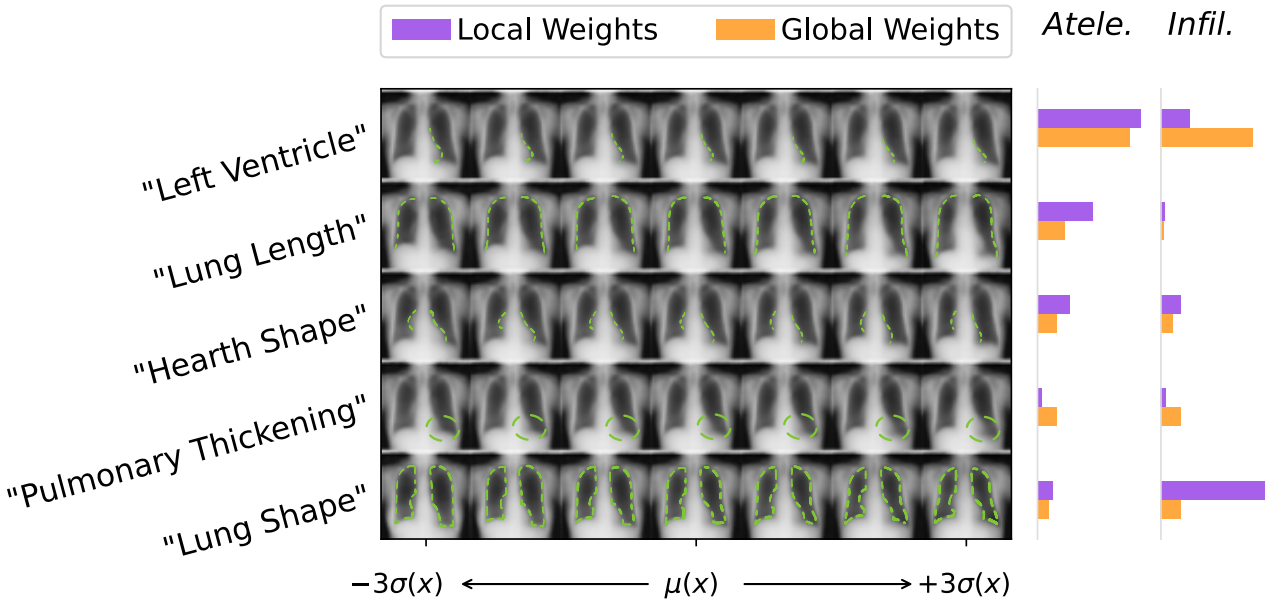


Figure 15: Complete *CLAP* traversals, corresponding to those shown in Figure 4(a).

We present additional ChestXRay traversals of *CLAP* for $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ (second row) in Figure 16.

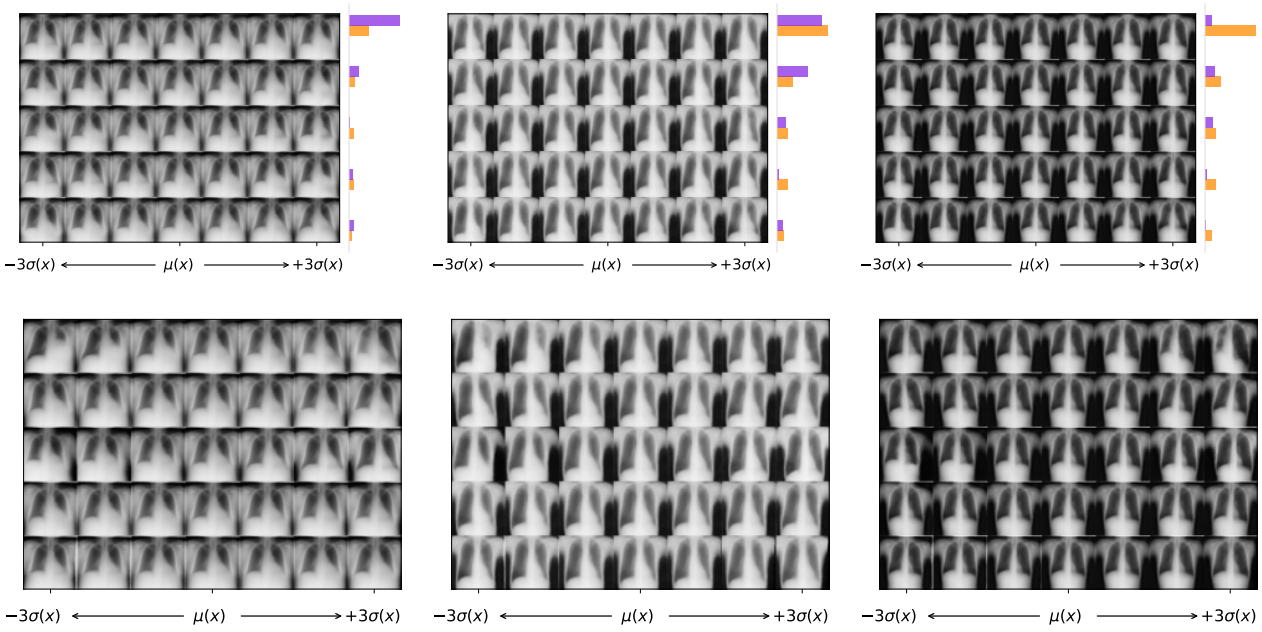
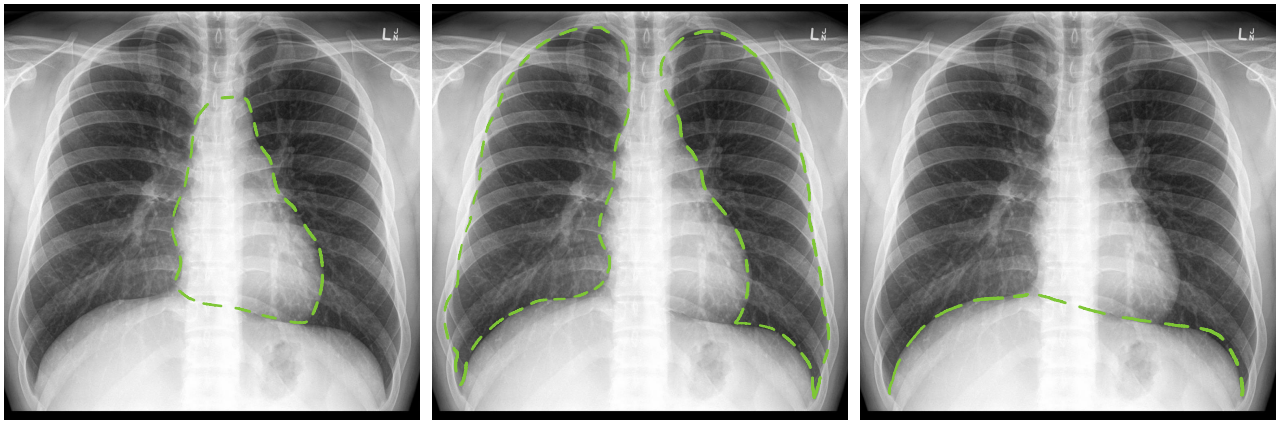


Figure 16: *CLAP* traversals of $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ on the ChestXRay dataset.

G.1. Glossary and Reading of ChestXray

In Figure 17 we show the main thorax parts used for the analysis of the *CLAP* traversals. We also remark that left and right are intended from the patient's viewpoint, which is reversed with respect to that of the reader.



(a) Heart. (b) Lungs. (c) Liver and stomach.

Figure 17: ChestXRay annotations.

H. PlantVillage traversals

In Figure 18 we present traversals of *CLAP* for $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ (second row) on the PlantVillage dataset (Hughes et al., 2015). The dataset includes various plants' leaves, we utilize the 10 binary labels (9 diseases plus healthy or not) of the tomato leaves. We include a magnified figure with human interpretations in Figure 19.

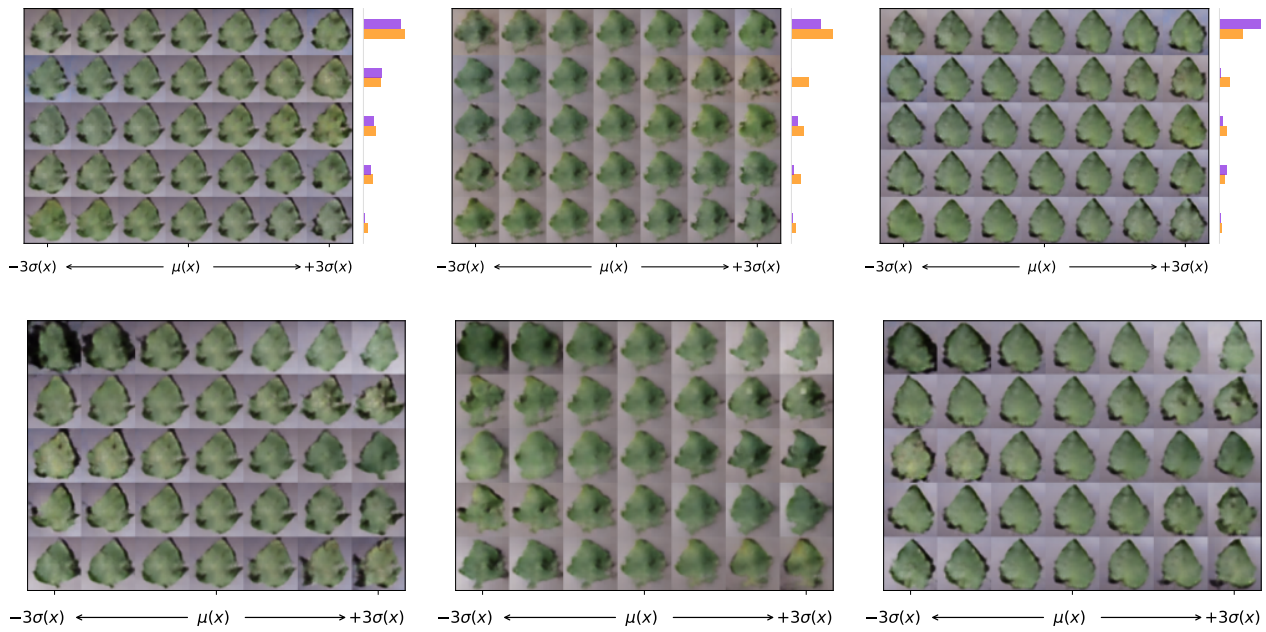


Figure 18: *CLAP* traversals of $\mathbf{Z}_c$ (first row) and $\mathbf{Z}_s$ on the PlantVillage dataset.

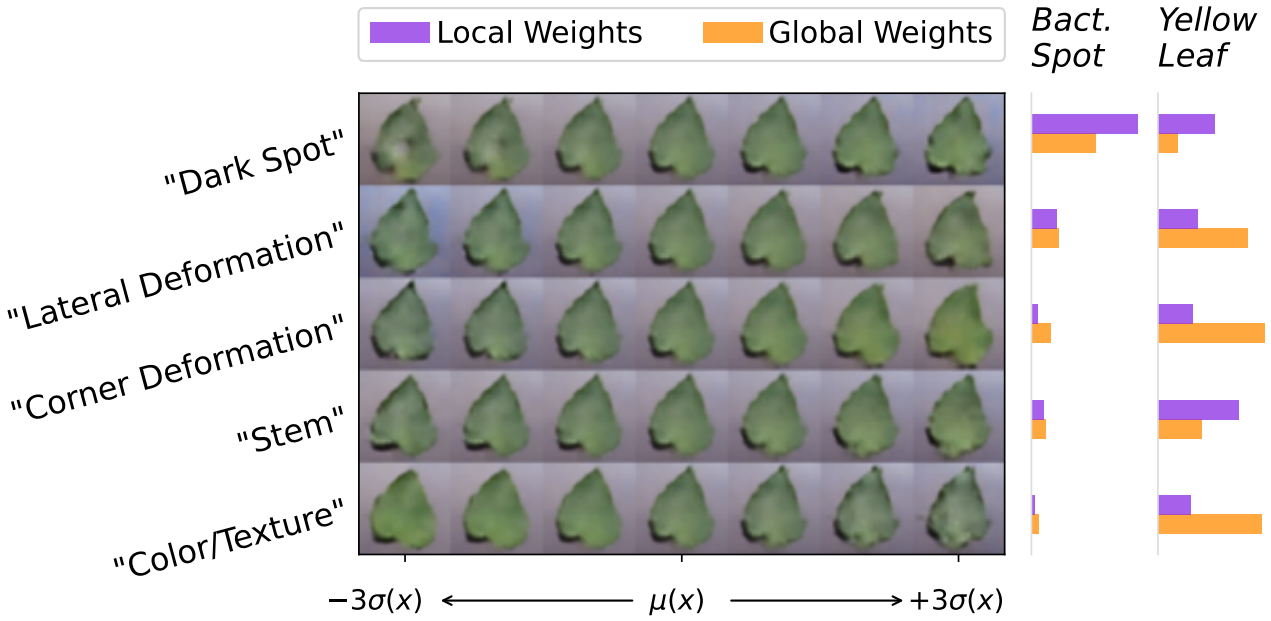


Figure 19: Traversals and human intepretations of *CLAP* on the PlantVillage dataset. We include prediction weights for the *Bacterial Spot* and *Yellow Leaf* diseases.