Learning Manifold Dimensions with Conditional Variational Autoencoders

Yijia Zheng[1]* **Tong He**[2] **Yixuan Qiu**[3] **David Wipf**[2]
[1] Purdue University
[2] Amazon Web Services
[3] School of Statistics and Management, Shanghai University of Finance and Economics
zheng709@purdue.edu, {htong, daviwipf}@amazon.com, qiuyixuan@sufe.edu.cn

Abstract

Although the variational autoencoder (VAE) and its conditional extension (CVAE) are capable of state-of-the-art results across multiple domains, their precise behavior is still not fully understood, particularly in the context of data (like images) that lie on or near a low-dimensional manifold. For example, while prior work has suggested that the globally optimal VAE solution can learn the correct manifold dimension, a necessary (but not sufficient) condition for producing samples from the true data distribution, this has never been rigorously proven. Moreover, it remains unclear how such considerations would change when various types of conditioning variables are introduced, or when the data support is extended to a union of manifolds (e.g., as is likely the case for MNIST digits and related). In this work, we address these points by first proving that VAE global minima are indeed capable of recovering the correct manifold dimension. We then extend this result to more general CVAEs, demonstrating practical scenarios whereby the conditioning variables allow the model to adaptively learn manifolds of varying dimension across samples. Our analyses, which have practical implications for various CVAE design choices, are also supported by numerical results on both synthetic and real-world datasets.

1 Introduction

Variational autoencoders (VAE) [6, 14] and conditional variants (CVAE) [17] are powerful generative models that produce competitive results in various domains such as image synthesis [21, 5, 13], natural language processing [16], time-series forecasting [9, 19], and trajectory prediction [8]. As a representative example, when equipped with an appropriate deep architecture, VAE models have recently achieved state-of-the-art performance generating large-scale images [11]. And yet despite this success, there remain VAE/CVAE behaviors in certain regimes of interest where we lack a precise understanding or a supporting theoretical foundation.

In particular, when the data lie on or near a low-dimensional manifold, as occurs with real-world images [12], it is meaningful to have a model that learns the manifold dimension correctly. The latter can provide insight into core properties of the data and be viewed as a necessary, albeit not sufficient, condition for producing samples from the true distribution. Although it has been suggested in prior work [4, 3] that a VAE model can learn the correct manifold dimension when globally optimized, this has only been formally established under the assumption that the decoder is linear or affine [2]. And the potential ability to learn the correct manifold dimension becomes even more nuanced when a conditioning variable is introduced. In this regard, a set of discrete conditions (e.g., MNIST image digit labels) may correspond with different "slices" through the data space, with each

*Work completed during internship at the AWS Shanghai AI Labs.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

inducing a manifold with varying dimension (intuitively, the manifold dimension of images labelled "1" is likely smaller than those of "5"). Alternatively, it is possible to have data expand fully in the ambient space but lie on a low-dimensional manifold when continuous conditional variables are present. Such a situation can be trivially constructed by simply treating some data dimensions, or transformations thereof, as the conditioning variable. In both scenarios, the role of CVAE models remains under-explored.

Moreover, unresolved CVAE properties in the face of low-dimensional data structure extend to practical design decisions as well. For example, there has been ongoing investigation into the choice between a fixed VAE decoder variance and a learnable one [4, 10, 15, 18, 3], an issue of heightened significance when conditioning variables are involved. And there exists similar ambiguity regarding the commonly-adopted strategy of sharing weights between the prior and encoder/posterior in CVAEs [7, 17]. Although perhaps not obvious at first glance, in both cases these considerations are inextricably linked to the capability of learning data manifold dimensions.

Against this backdrop our paper makes the following contributions:

(i) In Section 2.1 we provide the first demonstration of general conditions under which VAE global minimizers provably learn the correct data manifold dimension.

(ii) We then extend the above result in Section 2.2 to address certain classes of CVAE models with either continuous or discrete conditioning variables, the latter being associated with data lying on a union of manifolds.

(iii) Later, Section 3 investigates common CVAE model designs and training practices, including the impact of strategies for handling γ and the impact of weight sharing between condition prior and posterior networks.

(iv) Section 4 supports our theoretical conclusions and analysis with numerical experiments on both synthetic and real-world datasets.

2 Learning the Dimension of Data Manifolds

We begin with the definition of data:

Definition 1 (Data lying on a Manifold) *Suppose r and d are two positive integers with $r < d$. Then $\mathcal{X}$ is a simple r-Riemannian manifold embedded in $\mathbb{R}^d$ when there exists a diffeomorphism φ between $\mathcal{X}$ and $\mathbb{R}^r$. Specifically, for every $x \in \mathcal{X}$, there exists a $u = \varphi(x) \in \mathbb{R}^r$, where φ is invertible and both φ and φ^{-1} are differentiable.*

Given infinite observable variables $x \in \mathcal{X}$, where $\mathcal{X}$ is a r-dimensional manifold embedded in $\mathbb{R}^d$ whose definition is in 1, and latent variable $z \in \mathcal{Z}^\kappa \subseteq \mathbb{R}^\kappa$ which serves as a low-dimensional representation, a VAE model is built to approximate the ground truth measure with a parametric density $p_\theta(x) = \int p_\theta(x|z)p(z)dz$, where prior $p(z) = N(0, I)$. Denote the ground-truth measure on $\mathcal{X}$ as ω_{gt}. The probability mass of an infinitesimal dx on the manifold is $\omega_{gt}(dx)$ and $\int_{\mathcal{X}} \omega_{gt}(dx) = 1$. The canonical cost of VAE model is a bound on the average negative log-likelihood

$$\mathcal{L}(\theta, \phi) = \int_{\mathcal{X}} \{-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathbb{KL}[q_\phi(z|x)||p(z)]\}\omega_{gt}(dx) \tag{1}$$

Before starting our main theorems, we first convert the similar idea from [4] and formalize a κ-simple VAE model with explicit Gaussian and Lipschitz assumptions, as well as the way of parameterizations.

Definition 2 (κ-simple VAE) *A κ-simple VAE is defined as a VAE model with dim[z] = κ latent dimensions, the Gaussian encoder $q_\phi(z|x) = N(z|\mu_z(x;\phi), \mathrm{diag}\{\sigma_z^2(x;\phi)\})$, the Gaussian decoder $p_\theta(x|z) = N(x|\mu_x(z;\theta), \gamma I)$, and the prior $p(z) = N(z|0, I)$. Here $\gamma > 0$ is a trainable scalar in θ. The mean functions of the encoder and the decoder, i.e. $\mu_z(x;\phi)$ and $\mu_x(z;\theta)$ are arbitrary L-Lipschitz continuous functions.*

2.1 Learning Manifold Dimension in VAE

We start by analyzing the dimension of manifold learning in VAE models. As it is pointed out in [4], if γ of a VAE model is trainable and not constrained, γ must be arbitrarily small but explicitly

nonzero to minimize the cost and reconstruct data from $\mathcal{X}$. Here we define *active latent dimensions* as the dimensions that have a decaying encoder variance related to γ.

Definition 3 (Active Latent Dimensions in VAE) *In a κ-simple VAE, let $\{\theta^*_\gamma, \phi^*_\gamma\}$ denote any global optimal κ-simple VAE solutions as a function of any fixed γ, then for any $j = 1, \ldots, \kappa$, a dimension j of latent variable z is defined as an active dimension if the optimal encoder variance $\sigma_z(x; \phi^*_\gamma)^2_j \to 0$ as $\gamma \to 0$.*

Theorem 1 (Learning the Data Manifold Dimension Using VAEs) *Suppose training data lies on a r-dimensional manifold embedded in $\mathbb{R}^d$. Then for any $\kappa \geq r$, when a κ-simple VAE model achieves its any global optimums, we have*

(i) $\mathcal{L}(\theta^*_\gamma, \phi^*_\gamma) = (d - r)\log\gamma + O(1)$*, and*

(ii) The number of active latent dimensions almost surely equals r, and

(iii) The optimal variance of each active dimension satisfies $\sigma_z(x; \phi^*_\gamma)^2_j = O(\gamma)$.

Theorem 4 aims to state that when the decoder is an L-Lipschitz continuous function and learns the ground truth diffeomorphism, the number of active dimensions determines intrinsic dimensionality of the generative model in data space.

A rigorous proof of Theorem 4 is provided in the appendices. As a summary, our proof follows this path: first, we prove the existence of active dimensions and the rate by contradiction. The conclusion is that there must exist at least r active dimensions whose encoder variances going to zero at a rate of $O(\gamma)$ for reconstruction tasks, otherwise the reconstruction term will grow at a rate of $O(\frac{1}{\gamma})$ as γ goes to zero. Next, we get an upper bound and a lower bound of ELBO and show that their difference is a constant, both in the form of $(d - r)\log\gamma + O(1)$. Last, we get the exact number of active dimensions by showing that adding unnecessary active dimensions decreases the coefficient of $\log\gamma$, which may increase loss value.

This result implies that we can explicitly obtain the manifold dimension and the state of convergence by viewing the performance of σ_z^2 and γ. Also, both in theory and experiments, the ratio of the norm term in reconstruction and γ is a constant, i.e. the data dimensions d. Specifically, when a VAE model converges to its optimum, γ will go to zero, as well as r dimensions of σ_z^2, which comes from the optimal $\gamma^* = \arg\min_\gamma \tilde{\mathcal{L}}(\theta, \phi) = \frac{L^2}{d}\mathbb{E}_{\varepsilon\sim N(0,I)}[||\sigma_z(x)_{1:r}\varepsilon||^2]$. To achieve the optimal solution, the active latent dimensions play a crucial role in reconstruction and the inactive ones help to push KL regularization factor to its minimum along these dimensions. By definition, inactive dimensions are with a variance such that $\lim_{\gamma\to 0}\sigma_z(x; \phi^*_\gamma) = C > 0$, where C is some constant. When such latent dimensions are greater than $d - r$, no matter how many inactive dimensions there are, the model cannot perfectly reconstruct even a single manifold dimension.

Note that previous work [4] has demonstrated that global minima of VAE models can achieve zero reconstruction error for all samples lying on the data manifold. However, it was not previously demonstrated in a general setting that this perfect reconstruction was possible using a minimal number of active latent dimensions, and hence, it is conceivable for generated samples involving a larger number of active dimensions to stray from this manifold. In contrast, to achieve perfect reconstruction using the minimal number of active latent dimensions, as we will demonstrate under the stated assumptions, implies that generated samples must also lie on the manifold. The noisy signals from inactive dimensions are blocked by the decoder and therefore cannot produce deviations from the manifold.

2.2 Learning Manifold Dimension in Conditional VAE

In this section, we extend the analysis on κ-simple VAE models to κ-simple CVAE models. We first the manifold with condition. Specifically, suppose $\mathcal{X}$ is a Riemannian manifold and there exists a set $\mathcal{C}$ where for each $x \in \mathcal{X}$, there is a corresponding $c \in \mathcal{C}$, and we call that c is the condition of x.

Definition 4 (κ-simple CVAE) *For any $x \in \mathcal{X}$ and its corresponding $c \in \mathcal{C}$, a κ-simple CVAE is an extension of κ-simple VAE with the Gaussian encoder $q_\phi(z|x, c) = N(z|\mu_z(x, c; \phi), \mathrm{diag}\{\sigma_z^2(x, c; \phi)\})$, the Gaussian decoder $p_\theta(x|z, c) = N(x|\mu_x(z, c; \theta), \gamma I)$, and*

the prior $p_\theta(z|c) = N(z|\mu_z(c;\theta), \mathrm{diag}\{\sigma_z^2(c;\theta)\})$. $\mu_z(x,c;\phi)$ *and* $\mu_x(z,c;\theta)$ *are arbitrarily L-Lipschitz continuous, while* $\mu_z(c;\theta)$ *can be arbitrary function.*

The canonical CVAE cost with conditional variable being c is

$$\mathcal{L}_c(\theta,\phi) = \int_{\mathcal{X}} \{-\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] + \mathbb{KL}[q_\phi(z|x,c)||p_\theta(z|c)]\}\omega_{gt}(dx) \tag{2}$$

When prior is independent of c, i.e. $p(z|c) = p(z)$, the cost still holds.

In a CVAE model, besides samples $x \in \mathcal{X}$, the conditioning variable c may also provide information relevant to x. Thus we define the *effective dimension* to measure the contribution of the conditioning variable.

Definition 5 (Effective Dimension of the Conditioning Variable) *For any* $x \in \mathcal{X}$ *and its corresponding* $c \in \mathcal{C}$*, we call integer* t *the number of effective dimensions of* c*, if 1) there exists a function* $g : \mathcal{C} \to \mathbb{R}^t$ *and exists* t *dimensions of* $\varphi(x)$*, denoted as* $\varphi(x)_t$*, such that* $g(c) = \varphi(x)_t$*, where* φ *is the diffeomorphism in Definition 1, and 2) there doesn't exist such a function* g *for* $t+1$*. Further, let* $\mathcal{C}_t \subset \mathcal{C}$ *be the subset for all* c *with effective dimensions of* t*.*

Definition 5 shows that any $c \in \mathcal{C}_t$ can reconstruct at most $t \le r$ dimensions of the manifold.

Also, we extend our definition of *active latent dimensions* to CVAE.

Definition 6 (Active Latent Dimensions in CVAE) *In a* κ*-simple CVAE, denote* $\mathrm{diag}\{\sigma_{z_p}(c;\theta)\}$ *and* $\mathrm{diag}\{\sigma_{z_q}(x,c;\phi)\}$ *as the variance of the prior and encoder. For any* $j = 1,\ldots,\kappa$*, a dimension* j *of latent variable* z *is defined as an active dimension if the ratio* $\sigma_{z_q}(x,c;\phi^*_\gamma)_j^2/\sigma_{z_p}(c;\theta^*_\gamma)_j^2 \to 0$ *as* $\gamma \to 0$*.*

As special cases, for VAE or CVAE with a standard Gaussian prior we have $\sigma_{z_p}(x)_j^2 = 1$, and that indicates Definition 3 is a special case of Definition 6.

Next, we extend Theorem 4 to CVAE models, which describes how the manifold dimension relates to conditioning variables.

Theorem 2 (Learning Manifold Dimension in CVAE) *Suppose* $\mathcal{X}$ *is a* r*-dimensional manifold embedded in* $\mathbb{R}^d$ *with its corresponding condition* $c \in \mathcal{C}_t$*. Then for any* $\kappa \ge r$*, when a* κ*-simple CVAE model achieves its global optimum, we have*

(i) $\mathcal{L}_c(\theta^*_\gamma,\phi^*_\gamma) = (d-r+t)\log\gamma + O(1)$*, and*

(ii) The number of active latent dimensions almost surely equals $r-t$*, and*

(iii) The variance of each active dimension satisfies $\sigma_{z_q}(x,c;\phi^*_\gamma)_j^2/\sigma_{z_p}(c;\theta^*_\gamma)_j^2 = O(\gamma)$*.*

This theorem illustrates the significance of conditioning variable that it can push the cost to a smaller value than a VAE model without condition, which is favorable. It also points out that the condition will reduce the number of active latent dimensions without disrupting the data fit. It worth mentioning that even though the data is not on the manifold initially, once a conditioning variable with a positive effective dimension is incorporated, the conditioned data then become on a manifold, which can be induced directly in our theorem.

Pushing further, we have the following result describing CVAE's capacity in learning data lies on a union of manifolds.

Corollary 2.1 (Adaptive Active Dimension) *Suppose there exist* $x_1, x_2 \in \mathcal{X}$ *with* r*-dimensional manifold and their corresponding conditioning variables* $c_1 \in \mathcal{C}_{t_1}$ *and* $c_2 \in \mathcal{C}_{t_2}$*, i.e.* c_1*,* c_2 *have* t_1 *and* t_2 *effective dimensions respectively. Given a CVAE model with an optimal solution* $\{\theta^*,\phi^*\}$*, we have the number of active dimensions of* $z_1 \sim q_{\phi^*}(z|x,c_1)$ *equals to* $r-t_1$*, and that of* $z_2 \sim q_{\phi^*}(z|x,c_2)$ *equals to* $r-t_2$*.*

This result indicates that CVAE can adaptively learn different data manifolds under conditioning variables with different effective dimensions. When c is discrete, it is equivalent to training separate

VAEs where each model corresponds to a discrete condition value. On the other hand, collecting all possible conditions in a single CVAE model enables it to compress the dimension dynamically. For example, c describes the set of manifolds that share the same dimension, thus for samples in each set, the corresponding manifold dimensions are the same. When c is continuous, conditioning variables may contain intact information of certain manifold dimensions, which substitute for corresponding active dimensions. In Section 4, we demonstrate numerical evidence of this behavior.

3 On Common Model Design Choices

In this section, we review some model designs and training practices that are commonly seen in prior works but could potentially negatively impact the model convergence and performance, especially within the present context of learning data manifold dimensionality.

3.1 On the Equivalence of Conditioned and Unconditioned Prior

In prior works, it is popular to design a parameterized prior $p_\theta(z|c)$[2] in CVAE [24, 7, 8, 20, 1], which is a trainable module alongside the encoder $q_\phi(z|x,c)$ and the decoder $p_\theta(x|z,c)$.

Remark 1 (Remove conditioning variable from the Prior) *Consider a κ-simple CVAE model with prior $p_\theta(z|c)$, encoder $q_\phi(z|x,c)$ and decoder $p_\theta(x|z,c)$. We can always find another κ-simple CVAE model with prior $p(z) \sim \mathcal{N}(0,I)$, encoder $q_{\phi'}(z|x,c)$ and decoder $p_{\theta'}(x|z,c)$, such that $\mathcal{L}_c(\theta,\phi) = \mathcal{L}_c(\theta',\phi')$.*

Remark 1 indicates that a conditioned prior is not necessary. Specifically, as shown in its proof, we can always explicitly convert an existing κ-simple CVAE model with prior $p_\theta(z|c)$ into another κ-simple CVAE model with prior being $\mathcal{N}(0,I)$. In practice, when we have the flexibility in architecture design, it is equivalent to use a standard Gaussian prior or a parameterized one.

3.2 The Initial γ Impacts Model Convergence and Dimension Learning

Recall in Theorem 4, we have $\mathcal{L}(\theta^*_\gamma,\phi^*_\gamma) = (d-r)\log\gamma + O(1)$, and that implies when $\mathcal{L}(\theta^*_\gamma,\phi^*_\gamma)$ approaches to $-\infty$, γ is expected to be closer to 0. This behavior makes γ important in CVAE training. A direct conclusion is that the model would be compromised if γ is set to be a fixed positive constant, as analysed and empirically studied in literature [4, 10, 15, 18, 3].

Although a model converging towards global optimum would push γ to 0 regardless of its initial value, in practice, we observe that this initial value of γ still significantly impacts the model performance. In the cost function, we have

$$-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] = \frac{1}{\gamma}\int_{\mathcal{Z}} q_\phi(z|x)||x-\mu_x(z)||_2^2 dz + \log(2\pi\gamma)$$

By this formulation we see that γ acts as the weight between data likelihood and KL-divergence. A large γ downgrades the weight for data likelihood, and encourages the model to optimize the KL-divergence term first, and vice versa. Making γ a learnable parameter is equivalent to adaptively balancing the cost. When the initial value is set to be on an extreme end, the model is at risk of falling into a local minimum, as empirically in Section 4.4.

3.3 Weight Sharing Keeps Sequence Model from Global Optimal Convergence

The cost of VAE balances the three modules the encoder, decoder and prior with the two terms, the data likelihood and KL-divergence. It has been argued in literature [3] that such a cost might lead to unstable training dynamics. Therefore, one widely adapted trick[17] in practice is to share model weights between prior and posterior.

Weight sharing is conducted in various forms. For data from a general domain where x and c are differently distributed, the weight sharing couples with the design in which the posterior takes in x

[2]Note that we slightly abused the notation of θ as parameters for both the prior and decoder.

and c as two input and process differently at the beginning, and the prior processes c as its input with the same module and weight in the posterior.

In this section we take the sequential data as a representation for the weight sharing issue because it takes a more thorough approach of weight sharing. Sequential data poses the task of using early parts of a sequence to predict the later parts. Formulating in CVAE, we have c being the early parts, and that conditions the to-predict later parts x.

Assuming x is a sequential data, i.e. $\{x_l\}$ with $l > 0$ denoted as the index, and each x_l is paired with its condition $c = x_{<l}$. As a result, the encoder becomes $q_\phi(z_l|x_{\leq l})$, and the prior becomes $p(z_l|x_{<l})$ for any l greater than the observed sequence length. It has been proposed in literature [7, 8] to share weights between the encoder and prior, i.e. $p(z_l|x_{<l}) = q_\phi(z_l|x_{<l})$.

Theorem 3 (Weight Sharing Compromises Performance of Sequential Modeling) *By sharing weights between the encoder and prior, the conditional VAE model has its cost being*

$$\mathcal{L}_c(\theta, \phi) = \int_{\mathcal{X}} \sum_l \{-\mathbb{E}_{q_\phi(z|x_{\leq l})}[\log p_\theta(x_l|z, x_{<l})] + \mathbb{KL}[q_\phi(z|x_{\leq l})||q_\phi(z|x_{<l})]\}\omega_{gt}(\mathrm{d}x) \quad (3)$$

Then we have $\mathcal{L}_c(\theta, \phi) = \Omega(1)$ for any θ and ϕ.

Theorem 6 implies that a fully weight-sharing pair of prior and encoder prevents the model from achieving its global optimum. As the numerical study in Section 4.5, this weight sharing strategy could severely constraint the performance.

4 Experiments

In this section we present numerical results to support our analysis. We first verify model behavior in the synthetic environment for a controllable set of experiments, and extend to real-world datasets to further validate our conclusions. [3]

4.1 Dataset and Metrics

Synthetic Dataset The basic data samples are generated from a mixture of a finite number of Gaussian distributions. Specifically, we first generate a 5-dimensional categorical selecting vector from uniform distribution, and normalize it with norm equaling one. Then we generate two r-dimensional parameter vectors μ_u and $\log \sigma_u$ from a standard Gaussian distribution in $\mathbb{R}^r$. By repeating the second step 5 times, we get 5 independent Gaussian distributions. Our synthetic data $u \in \mathbb{R}^r$ is generated by selecting one of the 5 candidate distributions with a probability of the selection vector, and next generate samples from the selected distribution. To get the manifold $\mathcal{X}$ in $\mathbb{R}^d$, we design a function $h : \mathbb{R}^r \to \mathbb{R}^d$, and the data sample $x = h(u)$. In our experiments, $h(u) = \text{Sigmoid}(G_x u)$, where $G_x \in \mathbb{R}^{d\times r}$ is randomly initialized controlled by a certain random seed. The conditioning variable c is transformed by another transform $h'(u_{1:t}) = G' u_{1:t}$, where $G' \in \mathbb{R}^{t\times t}$. Given a fixed seed, the generation results is controlled by parameters r, d and t. For all the experiments, the training size of synthetic data is 100,000.

Real-world dataset Besides synthetic dataset, we also investigate the model behavior on MNIST[23] and Fashion MNIST [22], two image datasets[4]. These two datasets contain x with complex mappings between the manifold and ambient space. At the same time, they are not too complex thus the training are made reasonably easy with common neural architecture, otherwise our focus could be driven to specific architecture choices other than VAE and CVAE model behavior. By carefully tuning the pair (x, c), we can control how c impacts the model.

Metrics The main metric that we use to evaluate VAEs and CVAEs is the negative Evidence Lower BOund (ELBO), as in Eq 1 and Eq 2, and note that it is reported on convention and not for comparison. There are also auxiliary metrics that we use to compare and diagnose models, including Active Dimensions (AD), Reconstruction error $\mathbb{E}||x - \mu_x(z)||^2$ (Recon), the KL-divergence ($\mathbb{KL}$), and the data variance γ.

[3] Code is available at https://github.com/zhengyjzoe/manifold-dimensions-cvae

[4] MNIST is under the CC BY-SA 3.0 license, and Fashion MNIST is MIT-Licensed.

Table 1: VAE latent compression in synthetic dataset. The first three blocks of rows varied d and r. VAE shows powerful capacity to learn the manifold dimension. With a surplus in latent dimension, i.e. $\kappa > r$, the number of active dimensions is exactly r and a small negative ELBO is achievable. As showed in the last row, given $\kappa < r$, model capacity is constrained since the active dimensions are not enough for reconstruction, which is indicated in the proof of Theorem 4.

κ	d	r	AD	Recon	$\mathbb{KL}$	γ	-ELBO
20	10	2	2	3×10^{-4}	18.31	1.625×10^{-5}	-58.26
		4	4	2.6×10^{-3}	24.22	5.654×10^{-5}	-29.83
		6	6	9.2×10^{-3}	24.14	3×10^{-4}	-17.39
		8	7	1.27×10^{-2}	27.91	1.4×10^{-3}	-10.38
		10	8	5.99×10^{-2}	16.39	2.5×10^{-3}	-6.40
	20	2	2	1.6×10^{-3}	17.98	5.052×10^{-5}	-114.52
		4	4	1.75×10^{-2}	23.11	2×10^{-4}	-60.90
		6	6	3.09×10^{-2}	28.96	6×10^{-4}	-43.75
		8	8	3.42×10^{-2}	33.83	1.2×10^{-3}	-36.82
		10	10	4.74×10^{-2}	35.81	1.1×10^{-3}	-28.34
	30	2	2	2.6×10^{-3}	18.42	7.221×10^{-5}	-176.74
		4	4	2.73×10^{-2}	24.60	2×10^{-4}	-100.28
		6	6	4.74×10^{-2}	31.89	9×10^{-4}	-76.46
		8	8	5.68×10^{-2}	37.28	1.6×10^{-3}	-65.66
		10	10	1.13×10^{-1}	35.13	2.5×10^{-3}	-47.00
5	20	6	5	1.299×10^{-1}	22.53	2.1×10^{-3}	-36.97
		8	5	3.719×10^{-1}	16.618	8.8×10^{-3}	-22.60
		10	5	3.564×10^{-1}	15.966	1.113×10^{-2}	-16.96

Table 2: VAE latent compression in real dataset. Support of Theorem 4.

Dataset	κ	AD	Recon	-ELBO
MNIST	5	5	14.899	-842.286
	16	12	9.749	-1065.83
	32	13	7.469	-1224.37
Fashion MNIST	5	5	13.163	-935.127
	16	9	9.026	-1216.68
	32	9	7.820	-1327.26

4.2 Manifold Dimensions Learning in VAE

Here, we provide direct numerical support to Theorem 4. Specifically, we show that the model captures the correct dimensions of manifold regardless of the ambient data dimensions. To indicate how the active dimension is determined, we provide an example in Section B of the appendix. Although in numerical experiments a variance cannot be exact zero, we can still observe from a converged model that each variance of encoder is either close to 1, or close to zero.

With $d = \kappa = 20$ and varying r, we demonstrate that the VAE model can learn the active latent dimensions when γ converges to a small value, regardless of the value of r. The results are showed in Table 1 and all the numbers are values at convergence.

Experimental results on MNIST and Fashion MNIST are reported in Table 2. Without knowing the true manifold dimensions, a small latent dimension discourages the model from catching up the information for every dimension, while once given enough latent dimensions, the number of active dimensions equals to data's intrinsic dimension and the rest is for the regularization. The two parts of latent dimensions cooperate to push the VAE model to its global optimum.

Table 3: CVAE latent compression showing AD $= r - t$ exactly, which supports our Theorem 5. c is the first t dimensions of u, i.e. $G' = I_t$, $d = \kappa = 20$, $r = 10$.

t	-ELBO	Recon	$\mathbb{KL}$	γ	AD
1	-31.41	4.61×10^{-2}	33.26	2.4×10^{-3}	9
3	-36.67	4.66×10^{-2}	27.78	2.4×10^{-3}	7
5	-42.78	4.86×10^{-2}	20.81	2.6×10^{-3}	5
7	-52.39	4.29×10^{-2}	13.72	2.2×10^{-3}	3
9	-62.25	3.84×10^{-2}	6.07	2×10^{-3}	1

Table 4: CVAE latent compression in real datasets. $\kappa = 32$. AD is averaged over all the classes. This result supports Theorem 5.

	AD	Recon	$\mathbb{KL}$	γ	-ELBO
MNIST	12	6.044	81.672	0.0063	-1489.42
Fashion MNIST	9	8.773	54.552	0.0102	-1239.09

4.3 CVAE Adaptive Active Dimensions Varied With c

In this section, we design experiments to show that a CVAE model can learn the dimension of manifold with respect to the effective latent dimensions of the condition. Further, it can also learn the adaptive active dimension on a union of manifolds.

By Theorem 5 and Remark 1, a CVAE model with an unconditioned prior and condition $c \in \mathcal{C}_t$ is equivalent to a general one in design, whose active dimensions equals $r - t$. In this experiments, we investigate the behavior of a CVAE model when data lies on a single manifold and its relation to condition is static. Specifically, we set c to be the first t dimensions of u, i.e. take G' as I_t, and let $d = \kappa = 20$, $r = 10$. Table 3 shows the results.

On MNIST and Fashion MNIST, we train CVAEs with the class label of each image as the conditioning variable, and the results with $\kappa = 32$ are in Table 4. By comparing it to Table 2, we notice that the CVAE model for MNIST has a lower active dimension and better ELBO, while for FashionMNIST the CVAE performs similarly. This implies that when data points in MNIST are conditioning on the label, they lie on a lower-dimensional manifold, while for data points in FashionMNIST where among classes the manifold dimensions are similar, the class label doesn't reduce the manifold dimension. This characterization of the dataset complexity between classes is consistent with their intuitive visual complexity. We also provide an example of the encoder variance on the MNIST dataset in Section B of the appendix.

Discrete Condition To verify Corollary 2.1, we train a CVAE model on d-dimensional data that lie on a union of manifolds with different dimensions, with the conditioning variable c being an indicator of the source manifold. Table 5 shows that the model successfully captures the active latent dimensions with an attention layer in the decoder, while without the attention layer the model learns a static active dimension of 4. The result indicates two aspects of CVAE: 1) CVAE models have capacity to capture a union of manifolds simultaneously and converge on each manifold (indicated by -ELBO); 2) In practice, we need careful model architecture design to help the model to distinguish manifolds. In this experiment $d = 20$ and $\kappa = 40$.

Continuous Condition When c is a continuous variable, we show that the model can also adaptively learn the active dimensions. In our setting, a continuous condition $c \in \mathcal{C}_t$ contains t-dimensional information of samples, thus under different conditions the data may lie on different manifolds. We set $r = 12$, $d = 20$ and $\kappa = 20$, then we set the first $2, 4, 6, 8, 10$ dimensions of u as the conditions. Since the model needs dimension of c to be consistent, we padding them to $\mathbb{R}^{10}$ with 0. With the attention layer, the model can use the condition to collapse certain dimensions instead of activating all the possible dimensions and increasing unnecessary loss in KL term. The results are included in Table 6.

Table 5: Adaptively learnt active dimension with discrete c. $d = 20, \kappa = 40$.

r	True AD	AD without attention	AD with attention	-ELBO without attention	-ELBO with attention
1	1	4	1	-102.25	-114.22
2	2	4	2	-62.42	-99.81
3	3	4	3	-60.13	-74.28
4	4	4	4	-28.06	-50.36
5	5	4	5	-7.58	-59.25

Table 6: Adaptively learnt active dimension with continuous c. $r = 12, d = 20, \kappa = 90$.

t	True AD	AD without attention	AD with attention	-ELBO without attention	-ELBO with attention
2	10	10	10	-9.69	-41.49
4	8	10	8	-33.71	-20.52
6	6	10	6	-27.10	-73.26
8	4	10	4	-50.70	-80.64
10	2	10	2	-61.77	-55.14

4.4 Initialization of γ Impacts Model Training

Here we check how the initialization of γ would impact model convergence, as mentioned in Section 3.2. Our results are in Table 7. With different initial $\log \gamma$ values, we observe the convergence of VAE is affected significantly, as a result, the model can't capture the correct manifold dimension. In comparison, CVAE models are less sensitive, as CVAEs have more parameters and the conditioned data lie on a lower-dimension manifold thus it actually poses an easier task. Here the CVAE model with an unparameterized prior $p(z)$ shows that in practice, given enough model capacity and good training parameters, one can directly train an unparameterized CVAE and achieve similar performance, instead of starting from training a parameterized one. Still, when $\log \gamma = 20$, they converge to local optima. From Table 7, we can conclude that in practice the performance at convergence heavily depends on both model architecture design and hyperparameters. Results with less careful architecture or hyper-parameters may not reflect the true global optimal. In this experiment, $d = 20, \kappa = 20, r = 10$. In CVAE, $t = 5$.

4.5 Shared Weights Between Encoder and Prior

The experiments in this section is to support Theorem 6, i.e. sharing weights between the encoder and prior is not a satisfactory way for training in sequential data, we generate a sequence by AutoRegressive-Moving-Average (ARMA) model with fixed parameters, then use rolling windows to make our samples. Specifically, the window size is 5, i.e. each sample we use the past 5 points as our condition, and reconstruct the sixth one which is viewed as current sample. The results in Table 8 show that when weight shared, the norm term of reconstruction cannot go down, which may result in a restriction in the capacity of the model, and this observation is coherent to the key contradiction used in our proof of Theorem 6.

Table 7: Active dimensions under different initial γ. $d = 20, r = 10, t = 5, \kappa = 20$.

Init $\log \gamma$	VAE		CVAE $p(z)$		CVAE $p(z\|c)$	
	AD	-ELBO	AD	-ELBO	AD	-ELBO
-20	10	-28.39	5	-41.20	5	-40.72
-10	9	-28.57	5	-44.53	5	-45.25
0	8	-27.56	5	-44.38	5	-45.2
10	3	-13.89	5	-43.72	5	-43.66
20	1	-1.7	5	-45.22	4	-37.85

Table 8: Weight sharing on sequential data between the encoder and prior.

Weight Shared	-ELBO	Recon	$\mathbb{KL}$	γ
True	-2.49	0.374	18.09	0.012
False	-45.015	1.81×10^{-5}	175.99	7.252×10^{-7}

5 Conclusion

In this paper we provide insights on the behavior of VAEs and CVAEs for data lying on low-dimensional manifolds. In the theoretic analysis, we show that VAE has the ability to learn the manifold dimension. More interestingly, we extend the conclusion to Conditional VAE that works with conditioned data. This additional conditioning term impacts the relation between data and manifold, as a result it also affects the model design choices in practice. With the theoretic results, we further examine some design choices. A dedicated section of experiments supports our statements and conclusions numerically on both synthetic and real-world data. Finally, we believe this work can in general help to improve the design and diagnose techniques of VAEs and CVAEs.

Limitations Our work may still have limitations in empirical verification for real-world datasets at a larger scale or from other modalities than images. Although our contributions are theoretically grounded, in practice it still takes non-trivial efforts to design model architectures that both align with our guideline and fit the data.

References

[1] Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[2] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Hidden talents of the variational autoencoder. *arXiv preprint arXiv:1706.05148*, 2018.

[3] Bin Dai, Li Wenliang, and David Wipf. On the value of infinite gradients in variational autoencoder models. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

[5] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taïga, Francesco Visin, David Vázquez, and Aaron C. Courville. Pixelvae: A latent variable model for natural images. In *5th International Conference on Learning Representations*, 2017.

[6] Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[7] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697. PMLR, 2018.

[8] Longyuan Li, Jian Yao, Li Wenliang, Tong He, Tianjun Xiao, Junchi Yan, David Wipf, and Zheng Zhang. Grin: Generative relation and intention network for multi-agent trajectory prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proceedings of the First Conference on Causal Learning and Reasoning*. PMLR, 2022.

[10] Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[12] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[13] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[15] Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

[16] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[17] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[18] Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, 2018.

[19] Binh Tang and David S Matteson. Probabilistic transformer for time series analysis. *Advances in Neural Information Processing Systems*, 34:23592–23608, 2021.

[20] Jakub Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[21] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[22] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[23] Corinna Cortes Yann LeCun. THE MNIST DATABASE of handwritten digits. `http://yann.lecun.com/exdb/mnist/`.

[24] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

A Details on model architecture and computation resources

Synthetic Dataset For the experiments with VAE models, both the encoder and decoder are defined as a Multi Layer Perceptron (MLP) with a single hidden layer. For all the experiments with CVAE models except Sec 4.5, the encoder first processes conditioning variable c via a MLP, then concatenate the output and sample x as the input of another MLP, and the decoder processes c in the same way then use a MLP to decode the latent variable. In Sec 4.5, both the encoder and decoder are LSTM with one hidden layer.

Real Dataset Both of the encoder and decoder use two ResNet Blocks to process MNIST/ Fashion MNIST images. The encoder block is made up of two ConvNet in residual function and a $1 * 1$ ConvNet for the shortcut to reshape. The decoder block uses one ConvNet in the residual function and inverse ConvNet for the shortcut.

Resources We conduct our experiments on an Amazon Web Services g4dn.12xlarge EC2 instance, which provides 4 T4 GPUs. We estimate that the time to run through all experiments in this paper once would cost 20 GPU-hours. The research activity for this paper cost around 100 GPU-hours in total.

B Example of the encoder variance

To show the active dimensions visually, here we report the encoder variances both on synthetic data and MNIST dataset.

Table 9: Encoder variance matrix of VAE on synthetic data in Table 1 of the main text, where $\kappa = 20, d = 30, r = 6$ and we find the number of active dimensions is 6. The active dimension figures are in blue.

0.0080	0.0018	1.0000	1.0000	1.0000
0.0027	0.0031	1.0000	1.0000	1.0000
1.0000	0.0087	1.0000	0.0141	1.0000
1.0000	1.0000	1.0000	1.0000	1.0000

Table 10: Encoder variance matrix of CVAE on MNIST data. the encoder variance of a CVAE model on MNIST dataset in Table 4 of the main text, where $\kappa = 32$ and we find the number of active dimensions is 12. The active dimension figures are in blue.

3.6159e-03	9.6320e-01	7.6566e-04	3.5173e-04
9.8518e-01	9.6739e-01	9.6077e-01	8.1020e-04
9.8065e-01	9.7336e-01	3.7781e-03	7.1394e-04
9.6985e-01	6.1294e-03	9.7449e-01	9.8012e-01
7.8233e-04	9.7318e-01	9.8596e-01	2.4359e-04
9.7785e-01	9.7737e-01	9.7315e-01	9.8431e-01
9.2616e-01	9.8335e-01	9.6775e-01	1.2756e-03
1.0324e-03	9.6723e-01	9.6046e-01	2.1289e-03

C Proof of Theorem 1

Recap of Theorem 1

Theorem 4 (Learning the Data Manifold Dimension Using VAEs) *Suppose training data lies on a r-dimensional manifold embedded in $\mathbb{R}^d$. Then for any $\kappa \geq r$, when a κ-simple VAE model achieves its any global optimums, we have*

(i) $\mathcal{L}(\theta^*_\gamma, \phi^*_\gamma) = (d - r) \log \gamma + O(1)$*, and*

(ii) The number of active latent dimensions almost surely equals r, and

(iii) The optimal variance of each active dimension satisfies $\sigma_z(x;\phi^*_\gamma)^2_j = O(\gamma)$.

Summary of the Proof In this section, we define three categories based on the number of active dimensions and the rate of their encoder variance. Note that any possible VAE optimum has to fall into the following three categories: the number of active dimensions whose encoder variance $\sigma_z^2(x,\phi^*_\gamma) = O(\gamma)$ is either greater than r, equal to r or less than r. The proof's logic flow is:

1. When the number of active dimensions whose encoder variance $\sigma_z^2(x) = O(\gamma)$ is less than r, the reconstruction error will increase at a rate of $O(\frac{1}{\gamma})$, thus the cost cannot reach the optimum. This is proven in Sec C.1;

2. When such dimensions' number equals r, the cost is exactly $(d-r)\log\gamma + O(1)$. The corresponding proof is in Sec C.2;

3. When such dimensions' number is greater than r, denoted as $m > r$, the cost is $(d-m)\log\gamma + O(1) > (d-r)\log\gamma + O(1)$. It is showed in Sec C.3.

C.1 The number of active dimensions whose encoder variance $\sigma_z^2(x) = O(\gamma)$ is less than r

The main idea is to link the gap between a large σ_z and large reconstruction error. For a given z_0, $\mu_x(z_0)$ will equal some x_0 such that $||x_0 - \mu_x(z_0)||^2 = 0$. But for other choices from $\mathcal{X}$ where $x \neq x_0$, we have $||x-\mu_x(z_0)||^2 > 0$ leading to the positive expectation term $\int_z q(z|x)||x-\mu_x(z)||^2 dz$. To minimize such positive error, we need to lower the density $q(z|x)$ where $x \neq x_0$, which is a function of σ_z.

Suppose that the number of active dimensions whose encoder variance $\sigma_z^2(x) = O(\gamma)$ is less than r. In this section, we will show that under this assumption the model can't reach its global optimum, i.e. $\mathcal{L} - \infty$. Remind that the cost of VAE is

$$\mathcal{L}(\theta,\phi) = \int_{\mathcal{X}} \{-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathbb{KL}[q_\phi(z|x)||p(z)]\}\omega_{gt}(dx)$$

We have

$$\begin{aligned}
2\mathcal{L}(\theta,\phi) &= \int_{\mathcal{X}} \{-2\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + 2\mathbb{KL}[q_\phi(z|x)||p(z)]\}\omega_{gt}(dx) \\
&= d\log(2\pi\gamma) + \int_{\mathcal{X}} \left\{\gamma^{-1}\mathbb{E}_{q_\phi(z|x)}[||x-\mu_x(z)||^2] + 2\mathbb{KL}(q_\phi(z|x)||p(z))\right\}\omega_{gt}(dx) \\
&= d\log(2\pi\gamma) + \gamma^{-1}\int_{\mathcal{X}}\int_z q_\phi(z|x)[||x-\mu_x(z)||^2]dz\,\omega_{gt}(dx) \\
&\quad + \int_{\mathcal{X}} 2\mathbb{KL}(q_\phi(z|x)||p(z))\omega_{gt}(dx)
\end{aligned} \tag{4}$$

Following the two facts:

1. Lebesgue measure on the real numbers is σ-finite.

2. $z \in \mathbb{R}^\kappa$ and $x \in \mathcal{X}$, where $\mathcal{X}$ is a r-dimensional manifold embedded in $\mathbb{R}^d$.

and referring to Fubini's theorem, we can switch the integration order of $\omega_{gt}(dx)$ and dz. Assume the components of $z \in \mathbb{R}^\kappa$ is permutable. For a r-dimensional manifold, we can always use the first r dimensions of z to get $\varphi(x)$, i.e. once given r-dimensional information, there always exists a decoder,

s.t. $\mu_x(z_{1:r}) = \mu_x(z)$. Denote by $\mu_z(x)_{1:r}$ and $\sigma_z^2(x)_{1:r}$ the mean and covariance matrix of the first r dimension of z. After switching the integration order, we have

$$
\begin{aligned}
&\int_{\mathcal{X}} \frac{1}{\gamma} \int_z q_\phi(z|x)[||x-\mu_x(z)||^2]dz\, \omega_{gt}(dx) + \int_{\mathcal{X}} [d\log(2\pi\gamma) + 2\mathbb{KL}(q_\phi(z|x)||p(z))]\, \omega_{gt}(dx) \\
=&\frac{1}{\gamma} \int_z \int_{\mathcal{X}} q_\phi(z|x)[||x-\mu_x(z_{1:r})||^2]\omega_{gt}(dx)dz + \int_{\mathcal{X}} [d\log\gamma + 2\mathbb{KL}(q_\phi(z|x)||p(z)) + O(1)]\, \omega_{gt}(dx) \\
=&\frac{1}{\gamma} \int_{z\in\mathcal{Z}^r} \int_{\mathcal{X}} \frac{1}{\sqrt{(2\pi)^r|\sigma_z^2(x)_{1:r}|}} e^{-\frac{1}{2}(z-\mu_z(x)_{1:r})^T\sigma_z^{-2}(x)_{1:r}(z-\mu_z(x)_{1:r})}[||x-\mu_x(z)||^2]\omega_{gt}(dx)dz + \\
&\int_{\mathcal{X}} [d\log\gamma + 2\mathbb{KL}(q_\phi(z|x)||p(z)) + O(1)]\, \omega_{gt}(dx)
\end{aligned} \tag{5}
$$

C.1.1 Analyze the density with respect to $\sigma_z(x)_{1:r}$ and $z_{1:r} - \mu_z(x)_{1:r}$

Next, for the integral over $\mathcal{X}$ in the first term in Eq 5, we examine a certain $z_{1:r} \in \mathcal{Z}^r$ and view it as a constant. Since μ_x is a deterministic function, $\mu_x(z_{1:r})$ is also constant. The log-density on $z_{1:r}$ is

$$
\frac{r}{2}\log(\frac{1}{2\pi}) + \frac{1}{2}\log\frac{1}{|\sigma_z^2|} - \frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T\sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})
$$

Take the derivative of σ_z^2, we have

$$
-\frac{\sigma_z^{-2}}{2} + \frac{1}{2}\sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T\sigma_z^{-2}
$$

When σ_z^2 is smaller than $(z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T$, the second term's rate is larger. Thus the density is monotonically increasing when $\sigma_z^2 \prec (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T$ and monotonically decreasing when $\sigma_z^2 \succ (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T$. Note that μ_x is L-Lipschitz continuous, so we have $L|z_{1:r}-\mu_z(x)_{1:r}| \geq |\mu_x(z_{1:r}) - \mu_x(\mu_z(x)_{1:r})| = |\mu_x(z_{1:r}) - x|$. The equality comes from the fact that we can choose optimal μ_z and μ_x, s.t. $\mu_x(\mu_z(x)_{1:r}) = x$.

Now we can divide $x \in \mathcal{X}$ into four cases and we assume all the four disjoint cases exist when analyzing, otherwise the integration over corresponding domain is 0 which would not affect our result. The four cases are as follows:

1. $\mathcal{X}_1(z_{1:r}) = \{x : \sigma_z^2(x)_{1:r} \prec (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T\} \cap \{x : ||z_{1:r} - \mu_z(x)_{1:r}|| = +\infty\}$
2. $\mathcal{X}_2(z_{1:r}) = \{x : \sigma_z^2(x)_{1:r} \prec (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T\} \cap \{x : ||z_{1:r} - \mu_z(x)_{1:r}|| < +\infty\}$
3. $\mathcal{X}_3(z_{1:r}) = \{x : (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T \preceq \sigma_z^2(x)_{1:r} < \infty\}$
4. $\mathcal{X}_4(z_{1:r}) = \{x : (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T \preceq \sigma_z^2(x)_{1:r} = \infty\}$

We have $\mathcal{X}_1(z_{1:r}) \cup \mathcal{X}_2(z_{1:r}) = \{x : \sigma_z^2(x)_{1:r} \prec (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T\}$ and $\mathcal{X}_3(z_{1:r}) \cup \mathcal{X}_4(z_{1:r}) = \{x : \sigma_z^2(x)_{1:r} \succeq (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T\}$. Thus $\mathcal{X}_1(z_{1:r}) \cup \mathcal{X}_2(z_{1:r}) \cup \mathcal{X}_3(z_{1:r}) \cup \mathcal{X}_4(z_{1:r})$ cover the whole space of x related to $z_{1:r}$, i.e. $\mathcal{X}(z_{1:r})$.

(i) $\mathcal{X}_1(z_{1:r}) = \{x : \sigma_z^2(x)_{1:r} \prec (z_{1:r}-\mu_z(x)_{1:r})(z_{1:r}-\mu_z(x)_{1:r})^T\} \cap \{x : ||z_{1:r} - \mu_z(x)_{1:r}|| = +\infty\}$

Denote σ_l^2 as the lower bound of σ_z^2's eigenvalues, which cannot approach 0 by our assumption. $\sigma_z < +\infty$. The integral over $\mathcal{X}_1(z_{1:r})$

$$
\begin{aligned}
&\int_{\mathcal{X}_1(z_{1:r})} \frac{1}{\sqrt{|\sigma_z^2|}} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T\sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})}[||x-\mu_x(z_{1:r})||^2]\omega_{gt}(dx)_r \\
\leq &\int_{\mathcal{X}_1(z_{1:r})} \frac{1}{\sigma_l^r} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T\sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})}[L^2||z_{1:r}-\mu_z(x)_{1:r})||^2]\omega_{gt}(dx)_r
\end{aligned}
$$

will approach 0 as $||z_{1:r} - \mu_z(x)_{1:r}|| = +\infty$. Thus $\int_{z_{1:r}} 0dz = 0$

(ii) $\mathcal{X}_2(z_{1:r}) = \{x : \sigma_z^2(x)_{1:r} \prec (z_{1:r} - \mu_z(x)_{1:r})(z_{1:r} - \mu_z(x)_{1:r})^T\} \cap \{x : ||z_{1:r} - \mu_z(x)_{1:r}|| < +\infty\}$

Denote $N = \max_x\{||z_{1:r} - \mu_z(x)_{1:r}||^2\}$ and $\mathcal{X}_2^\alpha(z_{1:r}) = \{x : ||x - \mu_x(z_{1:r})||^2 > \alpha\} \cap \mathcal{X}_2(z_{1:r})$, where $\alpha 0$. If for any α, $\mathcal{X}_2^\alpha(z_{1:r}) = \emptyset$, we have for all $x \in \mathcal{X}_2(z_{1:r})$, $x = \mu_x(z_{1:r})$. However, if $\mu_x(x)_r \in \mathcal{X}_2(z_{1:r})$, i.e. $\mu_x(z_{1:r})$ satisfies $\sigma_z^2(\mu_x(z_{1:r})) \prec (z_{1:r} - \mu_z(\mu_x(z_{1:r})))(z_{1:r} - \mu_z(\mu_x(z_{1:r})))^T$. We can find a pair of μ_x and μ_z, e.g. identity mapping, s.t. $\mu_z(\mu_x(z_{1:r})) = z_{1:r}$ and $\sigma_z^2(\mu_x(z_{1:r})) < 0$, which is impossible. Thus $\mu_x(z_{1:r}) \notin \mathcal{X}_2(z_{1:r})$ and $\mathcal{X}_2(z_{1:r}) = \emptyset$. Thus, once $\mathcal{X}_2(z_{1:r}) \neq \emptyset$, there exists an α, s.t. $\mathcal{X}_2^\alpha(z_{1:r}) \neq \emptyset$. The integral over $\mathcal{X}_2(z_{1:r})$

$$
\begin{aligned}
&\int_{\mathcal{X}_2(z_{1:r})} \frac{1}{\sqrt{|\sigma_z^2|}} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T \sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r \\
\geq &\int_{\mathcal{X}_2(z_{1:r})} \frac{1}{\sigma_l^r} e^{-\frac{1}{2}\sigma_l^{-2r} N} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r \\
= &\frac{1}{\sigma_l^r} e^{-\frac{1}{2}\sigma_l^{-2r} N} [\int_{\mathcal{X}_2^\alpha(z_{1:r})} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r + \int_{(\mathcal{X}_2^\alpha(z_{1:r}))^c} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r] \\
\geq &\frac{\alpha}{\sigma_l^r} e^{-\frac{1}{2}\sigma_l^{-2r} N}
\end{aligned}
$$

The last inequality comes from the fact that $\varpi(\mathcal{X}_2^\alpha(z_{1:r})) \geq 1$ where ϖ is a counting measure and the non-negativity of the second term.

(iii) $\mathcal{X}_3(z_{1:r}) = \{x : (z_{1:r} - \mu_z(x)_{1:r})(z_{1:r} - \mu_z(x)_{1:r})^T \preceq \sigma_z^2(x)_{1:r} < \infty\}$

In this case the density is monotonically decreasing with σ_z. Since $\sigma_z \neq +\infty$, denote σ_u^2 as the upper bound of the eigenvalues of σ_z^2. It can also bound $||z_{1:r} - \mu_z(x)_{1:r}||^2$. Use the same strategy in (ii), define $\mathcal{X}_3^{\alpha'}(z_{1:r}) = \{x : ||x - \mu_x(z_{1:r})||^2 > \alpha'\} \cap \mathcal{X}_3(z_{1:r})$. If $\mathcal{X}_3(z_{1:r}) \neq \emptyset$, we have

$$
\begin{aligned}
&\int_{\mathcal{X}_3(z_{1:r})} \frac{1}{\sqrt{|\sigma_z^2|}} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T \sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r \\
\geq &\int_{\mathcal{X}_3(z_{1:r})} \frac{1}{\sigma_u^r} e^{-\frac{r}{2}\sigma_u^{-2r+2}} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r \\
= &\frac{1}{\sigma_u^r} e^{-\frac{r}{2}\sigma_u^{-2r+2}} [\int_{\mathcal{X}_3^{\alpha'}(z_{1:r})} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r + \int_{(\mathcal{X}_3^{\alpha'}(z_{1:r}))^c} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r] \\
\geq &\frac{\alpha'}{\sigma_u^r} e^{-\frac{r}{2}\sigma_u^{-2r+2}}
\end{aligned}
$$

(iv) $\mathcal{X}_4(z_{1:r}) = \{x : (z_{1:r} - \mu_z(x)_{1:r})(z_{1:r} - \mu_z(x)_{1:r})^T \preceq \sigma_z^2(x)_{1:r} = \infty\}$

In this case the density is monotonically decreasing with σ_z, and the dominant factor is $\frac{1}{\sqrt{|\sigma_z^2(x)_{1:r}|}}$. Since σ_z is arbitrarily large, it is obvious that $\sqrt{|\sigma_z^2|} > tr(\sigma_z^2) \geq ||z_{1:r} - \mu_z(x)_{1:r}||^2 \geq \frac{1}{L^2}||x - \mu_x(z_{1:r})||^2$. Note that $|\cdot| = det(\cdot)$.

The integral over $\mathcal{X}_4(z_{1:r})$

$$
\begin{aligned}
&\int_{\mathcal{X}_4(z_{1:r})} \frac{1}{\sqrt{|\sigma_z^2|}} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T \sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})} [||x - \mu_x(z_{1:r})||^2] \omega_{gt}(dx)_r \\
\leq &\int_{\mathcal{X}_4(z_{1:r})} \frac{L^2||z_{1:r} - \mu_z(x)_{1:r})||^2}{\sqrt{|\sigma_z^2(x)_{1:r}|}} \omega_{gt}(dx)_r
\end{aligned}
$$

will approach 0 as $\sigma_z^2 \to \infty$.

C.1.2 Analyze the existence of the above cases and get a lower bound

We have $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{X}_3 \cup \mathcal{X}_4$, where $\mathcal{X}_i = \cup_{z_{1:r}} \mathcal{X}_i(z_{1:r}), i = 1, 2, 3, 4$. In $\mathcal{X}_1 \cup \mathcal{X}_4$, the integral is 0. To get a lower bound of Eq 5, we need to prove $\mathcal{X}_2 \cup \mathcal{X}_3 \neq \emptyset$, i.e. there must exists $z_{1:r}$ such that $x \in \{\sigma_z^2(x)_{1:r} < \infty\} \cap \{||z_{1:r} - \mu_z(x)_{1:r}|| < \infty\}$ exists.

For $\sigma_z(x)_r$, if $\sigma_z(x)_r = \infty$, then in the KL term the trace $tr(\sigma_z^2(x)_{1:r}) = +\infty$ which cannot be offset by $-\log|\sigma_z^2(x)_{1:r}|$. Thus to minimize loss, $\sigma_z < \infty$.

For $||z_{1:r} - \mu_z(x)_{1:r}|| < \infty$, with L-Lipschitz continuity, for any $z^*_{1:r}$, we can find a $x^* \in \mathcal{X}$, s.t. $||z^*_{1:r} - \mu_z(x^*)_{1:r}|| = 0$. Denote $U_\delta(x)_r$ as a neighborhood of x with the radius of δ. For any $x \in U_\delta(x^*)$, we have

$$||\mu_z(x)_{1:r} - z^*_{1:r}|| = ||\mu_z(x)_{1:r} - \mu_z(x^*)_{1:r}|| \leq L||x - x^*|| \leq L\delta$$

So $U_\delta(x^*) \subset \mathcal{X}_2 \cup \mathcal{X}_3$. To get a positive lower bound, we need to prove there exists x' and δ, s.t. the image of $\mu_z(x')$ for $x \in U_\delta(x')$ is with positive measure. If for any $x \in U_\delta(x^1)$, $\mu_z(x)_{1:r} = z^1_{1:r}$, and for any $x \in U_\delta(x^2)$, $\mu_z(x)_{1:r} = z^2_{1:r}$, which satisfy $\delta < ||x^2 - x^1|| \leq \frac{3}{2}\delta$, and $z^1_{1:r} \neq z^2_{1:r}$. Note that can always choose a larger δ to get such pair of $\{x^1, x^2\}$. Then there exists $x^3 \in U_\delta(x^1) \cap U_\delta(x^2)$, $\mu_z(x^3)$ should equals $z^1_{1:r}$ and $z^2_{1:r}$ simultaneously which is impossible. Thus, there must exists x^*, s.t. $\mu_z(U_\delta(x^*))$ has a positive measure.

With the existence of x^*, such that $U_\delta(x^*) \subset \mathcal{X}_2 \cup \mathcal{X}_3$ and positive measured $\mu_z(U_\delta(x^*))$, we have

$$\begin{aligned}
&\frac{1}{\gamma}\int_{z_{1:r}}\int_{\mathcal{X}} \frac{1}{\sqrt{(2\pi)^r|\sigma_z^2|}} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T\sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})}[||x-\mu_x(z_{1:r})||^2]\omega_{gt}(dx)_r dz_{1:r} + \\
&\int_{\mathcal{X}} [d\log\gamma + 2\mathbb{KL}(q_\phi(z|x)||p(z)) + O(1)]\,\omega_{gt}(dx) \\
=&\frac{1}{\gamma}\int_{z_{1:r}}\int_{\mathcal{X}_2\cup\mathcal{X}_3} \frac{1}{\sqrt{(2\pi)^r|\sigma_z^2|}} e^{-\frac{1}{2}(z_{1:r}-\mu_z(x)_{1:r})^T\sigma_z^{-2}(z_{1:r}-\mu_z(x)_{1:r})}[||x-\mu_x(z_{1:r})||^2]\omega_{gt}(dx)_r dz_{1:r} + \\
&\int_{\mathcal{X}} [d\log\gamma + 2\mathbb{KL}(q_\phi(z|x)||p(z)) + O(1)]\,\omega_{gt}(dx) \\
\geq&\frac{1}{\gamma}\int_{z_{1:r}} \min\{\frac{\alpha}{\sigma_l^r}e^{-\frac{1}{2}\sigma_l^{-2r}N}, \frac{\alpha'}{\sigma_u^r}e^{-\frac{r}{2}\sigma_u^{-2r+2}}\}dz_{1:r} + \\
&\int_{\mathcal{X}} [d\log\gamma + 2\mathbb{KL}(q_\phi(z|x)||p(z)) + O(1)]\,\omega_{gt}(dx) \\
=&\frac{C}{\gamma} + \int_{\mathcal{X}} \left[d\log\gamma - \log|\sigma_z^2(x)| + O(1)\right]\omega_{gt}(dx)
\end{aligned} \tag{6}$$

Here denote $C = \int_{z_{1:r}} \min\{\frac{\alpha}{\sigma_l^r}e^{-\frac{1}{2}\sigma_l^{-2r}N}, \frac{\alpha'}{\sigma_u^r}e^{-\frac{r}{2}\sigma_u^{-2r+2}}\}dz_{1:r}$ for simplicity.

C.1.3 Analyze the rate of the lower bound

The first term $\frac{C}{\gamma}$ grows at a rate of O($\frac{1}{\gamma}$). Because σ_z^2 is at a lower rate than γ, we have

$$O(d\log\gamma - \log|\sigma_z^2(x)|) < O(-(d-\kappa)\log\frac{1}{\gamma})$$

and the right part decreases at a rate of $\log\frac{1}{\gamma}$. When $\gamma \to 0$, $O(\frac{1}{\gamma}) > O(\log\frac{1}{\gamma})$, which means the increase from reconstruction term cannot be offset by the decrease from the KL term. Moreover, from the fact that $O(\frac{1}{\gamma}) > O(\log\frac{1}{\gamma})$, when γ is small enough, the loss is monotonically decreasing with γ.

Therefore, when $\gamma \to 0$, the lower bound cannot approach $-\infty$, which means at this case, the model can never achieve optimum. Thus, there must exist some active dimensions whose variance $\sigma_z^2(x)$ satisfies $\sigma_z^2(x) = O(\gamma)$ as $\gamma \to 0$ to reach the global optimum. We can learn from the expression of C that as long as the number of such active dimensions whose encoder variance $\sigma_z^2(x) = O(\gamma)$ exceeds r, as γ approaches zero, the reconstruction term is at most at the rate of $O(1)$. Next, we will show that when there exist at least r such active dimensions, the VAE model's optimum is achievable.

C.2 The number of active dimensions whose encoder variance $\sigma_z^2(x) = O(\gamma)$ equals r

In this section, we get an upper bound and a lower bound and show that both case the cost is $(d-r)\log\gamma + O(1)$.

C.2.1 An Upper Bound of ELBO

Get an upper bound by Lipschitz We can write $z = \mu_z(x) + \varepsilon * \sigma_z(x)$, where $\varepsilon \sim N(0, I)$. Since decoder mean function $\mu_x(z;\theta)$ is L-Lipschitz continuous, we have:

$$\begin{aligned}
&\mathbb{E}_{z\sim q_{\phi_\gamma}(z|x)}[||x-\mu_x(z)||^2] \\
=&\mathbb{E}_{\varepsilon\sim N(0,I)}[||\mu_x(\mu_z(x)_{1:r}) - \mu_x(z_{1:r})||^2] \\
\leq&\mathbb{E}_{\varepsilon\sim N(0,I)}[||L(\mu_z(x)_{1:r} - \mu_z(x)_{1:r} - \sigma_z(x)_{1:r}\varepsilon)||^2] \\
=&\mathbb{E}_{\varepsilon\sim N(0,I)}[||L\sigma_z(x)_{1:r}\varepsilon||^2]
\end{aligned} \tag{7}$$

where the first equality comes from the fact that we can choose optimal encoder-decoder pairs such that $\mu_x(\mu_z(x)_{1:r}) = x$. Take it into $\mathcal{L}$,

$$\begin{aligned}
&2\mathcal{L}(\sigma_z(x)_{1:r}, \gamma) \\
=&\int_{\mathcal{X}} \left[\mathbb{E}_{z\sim q_\phi(z|x)}[\frac{1}{\gamma}||x-\mu_x(z)||_2^2] + d\log 2\pi\gamma - \log|\sigma_z^2(x)_{1:r}| - \log|\sigma_z^2(x)_{r+1:\kappa}| + O(1)\right]\omega_{gt}(dx) \\
\leq&\frac{L^2}{\gamma}\int_{\mathcal{X}} \left[\mathbb{E}_{\varepsilon\sim N(0,I)}[||\sigma_z(x)_{1:r}\varepsilon||^2] + d\log 2\pi\gamma - \log|\sigma_z^2(x)_{1:r}| - \log|\sigma_z^2(x)_{r+1:\kappa}| + O(1)\right]\omega_{gt}(dx)
\end{aligned} \tag{8}$$

We get an upper bound of $\mathcal{L}$, denoted as $\tilde{\mathcal{L}}$.

Analysis of the Upper Bound $\tilde{\mathcal{L}}$ Now we only pay attention to the upper bound $\tilde{\mathcal{L}}$ and try to prove that it is at a rate of $O((d-r)\log\gamma)$.

We can get implicit optimal values of $\tilde{\mathcal{L}}$: γ^* and $\sigma^*(x)_{1:r}^2$ by taking the derivative of $\tilde{\mathcal{L}}$ separately.

We have optimal γ^*

$$\gamma^* = \arg\min_\gamma \tilde{\mathcal{L}}(\theta,\phi) = \frac{L^2}{d}\mathbb{E}_{\varepsilon\sim N(0,I)}[||\sigma_z(x)_{1:r}\varepsilon||^2] \tag{9}$$

and

$$\begin{aligned}
\frac{\partial 2\tilde{\mathcal{L}}(\sigma_z(x)_{1:r},\gamma)}{\partial\sigma_z(x)_{1:r}} &= \frac{2L^2\sigma_z(x)_{1:r}}{\gamma}\mathbb{E}_{\varepsilon\sim N(0,I)}[\varepsilon\varepsilon^T] - 2\sigma_z(x)_{1:r}^{-1} \\
&= \frac{2L^2\sigma_z(x)_{1:r}}{\gamma} - 2\sigma_z(x)_{1:r}^{-1} = 0
\end{aligned}$$

we have the optimal variance of $\tilde{\mathcal{L}}$:

$$\sigma_z^*(x)_{1:r}^2 = \gamma\frac{I}{L^2} \tag{10}$$

It shows that $\frac{1}{\sqrt{\gamma}}\sigma_z^*(x)_{1:r} = O(1)$ when it reaches the optimal value.

Take the optimal values into $\tilde{\mathcal{L}}$, then we get $\tilde{\mathcal{L}}$ as a function of γ^* and $\sigma_z^*(x)_{1:r}$:

$$
\begin{aligned}
&2\tilde{\mathcal{L}}(\gamma^*, \sigma_z^*(x)_{1:r}) \\
=&\int_{\mathcal{X}} \left[\frac{L^2}{\gamma^*} \mathbb{E}_{\varepsilon \sim N(0,I)}[||\sigma_z(x)_{1:r}\varepsilon||^2] + d\log 2\pi\gamma^* - \log|\sigma_z^*(x)_{1:r}^2| - \log|\sigma_z(x)_{r+1:\kappa}^2| + O(1) \right] \omega_{gt}(dx) \\
=&\int_{\mathcal{X}} \left[d + d\log(2\pi\gamma^*) - \log|\gamma^* \frac{I}{L^2}| - \log|\sigma_z(x)_{r+1:\kappa}^2| + O(1) \right] \omega_{gt}(dx) \\
=&d\log(2\pi\gamma^*) - r\log\gamma^* - \log|\sigma_z(x)_{r+1:\kappa}^2| + O(1)
\end{aligned}
\tag{11}
$$

Define $\{\lambda_i\}_{i=1}^{\kappa}$ as the eigenvalues of $\sigma_z(x)$. Denote $\tilde{r}$ as the number of $\{\lambda_i\}_{i=r+1}^{\kappa}$ that will go to 0 as $\gamma^* \to 0$. We have

$$
2\tilde{\mathcal{L}}(\gamma^*) = d\log\gamma^* - r\log\gamma^* - 2\sum_{i=r+1}^{r+\tilde{r}} \log\lambda_i - 2\sum_{i=\tilde{r}+r+1}^{\kappa} \log\lambda_i + O(1) \tag{12}
$$

To minimize Eq 12, we want $\tilde{r}$ to be as small as possible and at the best equals 0 which is achievable. Since the rest $\kappa - r - \tilde{r} = \kappa - r$ dimensions are irrelevant to γ, at least will not approach 0 when $\gamma \to 0$, we can view them as constants. We have the loss equals

$$
(d-r)\log\gamma + O(1) \tag{13}
$$

C.2.2 A Lower Bound of ELBO

From C.1 we have analyzed the loss performance when there are less than r active dimensions whose variance goes to zero at a rate no lower than γ. In this part, we focus on the case that r latent dimensions are such active dimensions whose encoder variance goes to zero at a rate of $O(\gamma)$. Without loss of generality, we assume the first r latent dimensions satisfy $\sigma_z^2(x)_{1:r} = O(\gamma)$ as $\gamma \to 0$. We have

$$
\begin{aligned}
2\mathcal{L}(\theta,\phi) =& \int_{\mathcal{X}} \{-2\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + 2\mathbb{KL}[q_\phi(z|x)||p(z)]\}\omega_{gt}(dx) \\
=& \int_{\mathcal{X}} \left\{ \frac{1}{\gamma}\mathbb{E}_{q_\phi(z|x)}[||x-\mu_x(z)||^2] + d\log(2\pi\gamma) + 2\mathbb{KL}(q_\phi(z|x)||p(z)) \right\} \omega_{gt}(dx) \\
\geq& \int_{\mathcal{X}} \left\{ d\log(2\pi\gamma) - \log|\sigma_z^2(x)| + O(1) \right\} \omega_{gt}(dx) \\
=& \int_{\mathcal{X}} \left\{ d\log\gamma - \log|\sigma_z^2(x)_{1:r}| - \log|\sigma_z^2(x)_{r+1:\kappa}| + O(1) \right\} \omega_{gt}(dx) \\
\geq& \int_{\mathcal{X}} \left\{ (d-r)\log\gamma - \log|\sigma_z^2(x)_{r+1:\kappa}| + O(1) \right\} \omega_{gt}(dx) \\
=&(d-r)\log\gamma + O(1)
\end{aligned}
\tag{14}
$$

The inequalities come from the fact that the norm term is non-negative and the active dimensions' rate is no less than γ. For the last equality, we can use the strategy in Eq 12. To minimize the lower bound, there should not be any active dimensions in these $r+1:\kappa$ dimensions.

We get an upper bound and a lower bound at the same rate, i.e. $\log\gamma$ with r active latent dimensions. Therefore, the original loss is also with r active dimensions. We have the optimal cost for each x equals

$$
(d-r)\log\gamma + O(1) \tag{15}
$$

So far, we have get the conclusion in Thm 4 about the form of ELBO when $\gamma \to 0$, as well as the number and rate of active dimensions. Next, we show that the number of active dimensions can't be greater than r.

C.3 When the number of active dimensions is greater than r

Denote now there are m active dimensions and $m > r$. From Eq 11, in this case $\tilde{r} = m - r$, and the loss is

$$\frac{1}{\gamma}\mathbb{E}_{q_{\phi_\gamma}(z|x)}[||x - \mu_x(z)||^2] + d\log(2\pi\gamma) - 2\sum_{i=1}^{r}\log\lambda_i - 2\sum_{i=r+1}^{r+\tilde{r}}\log\lambda_i + O(1) \tag{16}$$

since here $\lim_{\gamma\to 0}\lambda_i = 0$, $-2\sum_{i=r+1}^{r+\tilde{r}}\log\lambda_i$ is monotonically increases as $\tilde{r}$ increases at the rate of $\Omega(\log\frac{1}{\gamma})$. For the reconstruction term, it is unaffected since we only use the first r latent dimensions for reconstruction. Therefore, the loss will increase at the rate of $\Omega(\log\frac{1}{\gamma})$, which is larger than the loss when $m = r$.

In conclusion, only when the number of active dimensions equals r, and these active dimensions' encoder variance $\sigma_z^2(x) = O(\gamma)$ as $\gamma \to 0$, the optimal cost is $(d - r)\log\gamma + O(1)$.

D Proof of Theorem 2

Recap of Theorem 2

Theorem 5 (Learning Manifold Dimension in CVAE) *Suppose $\mathcal{X}$ is a r-dimensional manifold embedded in $\mathbb{R}^d$ with its corresponding condition $c \in \mathcal{C}_t$. Then for any $\kappa \geq r$, when a κ-simple CVAE model achieves its global optimum, we have*

(i) $\mathcal{L}_c(\theta^*_\gamma, \phi^*_\gamma) = (d - r + t)\log\gamma + O(1)$, and

(ii) The number of active latent dimensions almost surely equals $r - t$, and

(iii) The variance of each active dimension satisfies $\sigma_{z_q}(x, c; \phi^*_\gamma)^2_j / \sigma_{z_p}(c; \theta^*_\gamma)^2_j = O(\gamma)$.

Summary of Proof In this section, we first make the proof when $p_\theta(z|c) = p(z)$, i.e. a parameter free prior, and then extend to the case when the prior involves the conditioning variable. The logic flow is as follows:

1. The prior is independent of c
   (a) Following the same proof idea in Theorem 1, when the number of active dimensions whose encoder variance $\sigma^2_{z_q}(x, c) = O(\gamma)$ is less than $r - k$, the reconstruction error will grow at a rate of $O(\frac{1}{\gamma})$. This is proven in Sec D.1.1;
   (b) In Sec D.1.2 we show that both the upper bound and the lower bound are $(d-r+t)\log\gamma$ and the exact number of active dimensions is $r - t$
2. The prior is a function of c
   (a) Since involving c in the prior will not affect the reconstruction term, we extend the conclusion in Sec D.1.1 to the general case;
   (b) Show that both the upper bound and the lower bound are $(d - r + t)\log\gamma$ and the exact number of active dimensions is $r - t$. The proof is in Sec D.2.

Under CVAE setting, we first make some denotations for proof. Since the encoder, prior, and decoder share the same condition c, the model has flexibility to use part of each c from the three networks. Denote t as the number of effective dimensions of c, and k as the number of effective dimensions of c used in the decoder $p_\theta(x|z, c)$, i.e. there exists a pair of encoder and decoder, s.t. $\mu_x(c) = \varphi^{-1}(u_{1:k})$, where $0 \leq k \leq t$ and is a learnable parameter. The encoder and prior use the rest $t - k$ effective dimensions, i.e. $\mu_x(\mu_z(c)) = \varphi^{-1}(u_{k+1:t})$, and this part of information will be included in the latent variable z.

D.1 When the prior is independent of c, i.e. $p_\theta(z|c) = p(z)$

In this case, we can write the cost as:

$$
\begin{aligned}
2\mathcal{L}_c(\theta,\phi) =& 2\int_{\mathcal{X}} \{-\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] + \mathbb{KL}[q_\phi(z|x,c)||p(z)]\}\omega_{gt}(dx) \\
=& \int_{\mathcal{X}} \frac{1}{\gamma}\mathbb{E}_{q_\phi(z|x,c)}[||x-\mu_x(z,c)||^2] + d\log(2\pi\gamma) + 2\mathbb{KL}(q_\phi(z|x,c)||p(z))\omega_{gt}(dx) \\
=& \frac{1}{\gamma}\int_{\mathcal{X}} \mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||\varphi^{-1}(u_{1:k}) - \mu_x(c)||^2 + ||\varphi^{-1}(u_{k+1:r}) - \mu_x(z_{k+1:r})||^2 + \\
& d\log(2\pi\gamma) + 2\mathbb{KL}(q_\phi(z|x,c)||p(z))\omega_{gt}(dx) \\
=& \frac{1}{\gamma}\int_{\mathcal{X}} \mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||\varphi^{-1}(u_{k+1:r}) - \mu_x(z_{k+1:r})||^2 + d\log(2\pi\gamma) + 2\mathbb{KL}(q_\phi(z|x,c)||p(z))\omega_{gt}(dx)
\end{aligned}
$$

D.1.1 The number of active dimensions whose encoder variance $\sigma^2_{z_q}(x,c) = O(\gamma)$ is less than $r-k$

Following the same proof idea in Theorem 1, assume there is no active dimension in $\sigma_z(x)$. For the reconstruction term, it is equivalent to reconstruct a $(r-k)$-dimensional manifold. Thus we can find an lower bound of the cost

$$
\begin{aligned}
2\mathcal{L}_c(\theta,\phi) \geq& \frac{C'}{\gamma} + \int_{\mathcal{X}} \left[d\log\gamma + 2\mathbb{KL}(q_\phi(z|x,c)||p(z)) + O(1)\right]\omega_{gt}(dx) \\
=& \frac{C'}{\gamma} + \int_{\mathcal{X}} \left[d\log\gamma - \log|\sigma^2_{z_q}(x,c)_{1:k}| - \log|\sigma^2_{z_q}(x,c)_{k+1:\kappa-t+k}| + O(1)\right]\omega_{gt}(dx)
\end{aligned} \tag{17}
$$

where $C' = \int_{z_{1:r-k}} \min\{\frac{\alpha}{\sigma_l^{r-k}} e^{-\frac{1}{2}\sigma_l^{-2(r-k)}N}, \frac{\alpha'}{\sigma_u^{r-k}} e^{-\frac{r-k}{2}\sigma_u^{-2(r-k)+2}}\} dz_{1:r-k}$.

The first term $\frac{C'}{\gamma}$ grows at a rate of O($\frac{1}{\gamma}$). Because σ^2_z is at a lower rate than γ, we have

$$
O(d\log\gamma - \log|\sigma^2_z(x,c)|) < O(-(d-\kappa)\log\frac{1}{\gamma})
$$

and the right part decreases at a rate of $\log\frac{1}{\gamma}$. When $\gamma \to 0$, $O(\frac{1}{\gamma}) > O(\log\frac{1}{\gamma})$, which means the increase from reconstruction term cannot be offset by the decrease from the KL term. Moreover, from the fact that $O(\frac{1}{\gamma}) > O(\log\frac{1}{\gamma})$, when γ is small enough, the loss is monotonically decreasing with γ.

Therefore, when $\gamma \to 0$, the lower bound cannot approach $-\infty$, which means at this case, the model can never achieve optimum. Thus, there must exist some active dimensions whose variance satisfies $\sigma^2_{z_q}(x,c)_i = O(\gamma), i = 1,\ldots,\kappa$ as $\gamma \to 0$ to reach the global optimum, and it is showed in C' that as long as the number of such active dimensions exceeds $r-k$, as γ approaches zero, the reconstruction term is at most at the rate of $O(1)$. Next, we will show that when there exist at least $r-k$ such active dimensions, the CVAE model's optimum is achievable.

D.1.2 Bounds of CVAE cost

The upper bound We have $z_{1:t-k} = \mu_{z_q}(x,c)_{1:t-k} + \sigma_{z_q}(x,c)_{1:t-k}\varepsilon_1$ and $z_{t-k+1:r} = \mu_{z_q}(x,c)_{t-k+1:r-k} + \sigma_{z_q}(x)_{t-k+1:r-k}\varepsilon_2$, where $\varepsilon_1 \sim N(0, I^{t-k}), \varepsilon_2 \sim N(0, I^{r-t})$.

The loss is:

$$
\begin{aligned}
&\frac{1}{\gamma}\mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||x-\mu_x(z,c)||^2]+d\log(2\pi\gamma)+2\mathbb{KL}(q_\phi(z|x,c)||p(z))\\
=&\frac{1}{\gamma}\mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||\varphi^{-1}(u_{1:k})-\mu_x(c)||^2+||\varphi^{-1}(u_{k+1:t})-\mu_x(z_{1:t-k})||^2+\\
&||\varphi^{-1}(u_{t+1:r})-\mu_x(z_{t-k+1:r-k})||^2]+d\log(2\pi\gamma)+2\mathbb{KL}(q_\phi(z|x,c)||p(z))\\
\leq&\frac{1}{\gamma}\mathbb{E}_{\varepsilon_1\sim N(0,I^{t-k})}[||L\sigma_{z_q}(x,c)_{1:t-k}\varepsilon_1||^2]+\frac{1}{\gamma}\mathbb{E}_{\varepsilon_2\sim N(0,I^{r-t})}[||L\sigma_{z_q}(x,c)_{t-k+1:r-k}\varepsilon_2||^2]+\\
&d\log\gamma-\log|\sigma^2_{z_q}(x,c)_{1:t-k}|-\log|\sigma^2_{z}(x,c)_{t-k+1:r-k}|+O(1)
\end{aligned}
\tag{18}
$$

Denote $\sigma^2_{z_q}(x,c)_{1:t-k}$ as $\sigma^2_{z_q}(c)$ for simplicity, and denote the upper bound of loss as $\mathcal{L}^u_c$. Take the derivative of $\sigma_{z_q}(c)$ and $\sigma_{z_q}(x,c)_{t-k+1:r-k}$ separately. Because the diagonal elements in σ_{z_q} are independent, we can make both achieve optimum. We have:

$$
\mathcal{L}^u_c(\gamma,k)=-(t-k)\log\gamma-(r-t)\log\gamma+d\log\gamma+O(1) \tag{19}
$$

To minimize $\mathcal{L}^u_c(\gamma,k)$, the optimal k is t, thus the upper bound is:

$$
\mathcal{L}^u_c(\gamma)=(d-r+t)\log\gamma+O(1) \tag{20}
$$

The lower bound We have show that there must be at least $r-k$ active dimension at a rate of $O(\gamma)$, otherwise the loss will increase at a rate of $O(\gamma)$. We can get a lower bound

$$
\begin{aligned}
&\frac{1}{\gamma}\mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||x-\mu_x(z,c)||^2]+d\log(2\pi\gamma)+2\mathbb{KL}(q_\phi(z|x,c)||p(z))\\
\geq&d\log\gamma-\log|\sigma^2_{z_q}(x,c)_{1:r-k}|-\log|\sigma^2_{z_q}(x,c)_{r-k+1:\kappa}|+O(1)\\
\geq&d\log\gamma-(r-k)\log\gamma-\log|\sigma^2_{z_q}(x,c)_{r-k+1:\kappa}|+O(1)\\
\geq&(d-r+k)\log\gamma+O(1)
\end{aligned}
\tag{21}
$$

Denote the lower bound as $\mathcal{L}^l_c(\gamma,k)$, to minimize it, we have $k=t$, thus the lower bound is

$$
\mathcal{L}^l_c(\gamma)=(d-r+t)\log\gamma+O(1)
$$

Both $\mathcal{L}^u_c$ and $\mathcal{L}^l_c$ are at a rate of $O(\log\gamma)$, we come to the conclusion that the ELBO is $(d-r+t)\log\gamma+O(1)$ and the number of active dimensions is $r-t$ when $p_\theta(z|c)=p(z)$.

D.2 The general case

Define a trainable parametric prior of z, i.e. $z\sim N(\mu_{z_p}(c),\sigma^2_{z_p}(c))$. Since involving c in the prior doesn't affect the reconstruction term, we have the conclusion in Section D.1.1 that there are at least $r-k$ active latent dimensions at a rate of $O(\gamma)$. Without loss of generality, we assume the first $r-k$ dimension of $\sigma^2_{z_q}(x,c)$, i.e. $\sigma^2_{z_q}(x,c)_{1:r-k}=O(\gamma)$.

The upper bound The loss is:

$$
\begin{aligned}
&\frac{1}{\gamma}\mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||x-\mu_x(z,c)||^2]+d\log(2\pi\gamma)+2\mathbb{KL}(q_\phi(z|x,c)||p(z|c))\\
\leq&\frac{1}{\gamma}\mathbb{E}_{\varepsilon_1\sim N(0,I^{t-k})}[||L\sigma_{z_q}(c)\varepsilon_1||^2]+\frac{1}{\gamma}\mathbb{E}_{\varepsilon_2\sim N(0,I^{r-k})}[||L\sigma_{z_q}(x,c)_{t-k+1:r-k}\varepsilon_2||^2]+d\log(2\pi\gamma)-\\
&\log|\sigma^2_{z_q}(c)|-\log|\sigma^2_{z_q}(x,c)_{t-k+1:r-k}|-\log|\sigma^2_{z_q}(x,c)_{r-k+1:\kappa}|+\log|\sigma^2_{z_p}(c)_{1:t-k}|+\log|\sigma^2_{z_p}(c)_{t-k+1:\kappa}|\\
&+(\mu_{z_q(1:t-k)}-\mu_{z_p(1:t-k)})^T\sigma^2_{z_p}(c)^{-1}_{1:t-k}(\mu_{z_q(1:t-k)}-\mu_{z_p(1:t-k)})\\
&+(\mu_{z_q(t-k+1:\kappa)}-\mu_{z_p(t-k+1:\kappa)})^T\sigma^2_{z_p}(c)^{-1}_{t-k+1:\kappa}(\mu_{z_q(t-k+1:\kappa)}-\mu_{z_p(t-k+1:\kappa)})\\
&-\kappa+tr(\sigma^2_{z_q}(c)/\sigma^2_{z_p}(c)_{1:t-k})+tr(\sigma^2_{z_q}(x)/\sigma^2_{z_p}(c)_{t-k+1:\kappa})
\end{aligned}
\tag{22}
$$

Since we can only control k dimensions of the prior when training, take the derivative of $\mu_{z_p}(c)_{1:k}$ and $\sigma_{z_p}(c)_{1:k}$, we have

$$
\begin{aligned}
&\mu_{z_p}(c)^*_{1:t-k}=\mu_{z_q}(c)\\
&\sigma^2_{z_p}(c)^*_{1:t-k}=(\mu_{z_q}(c)-\mu_{z_p}(c)^*_{1:t-k})(\mu_{z_q}(c)-\mu_{z_p}(c)^*_{1:t-k})^T+\sigma^2_{z_q}(c)=\sigma^2_{z_q}(c)
\end{aligned}
\tag{23}
$$

Let them achieve the optimal values. The loss becomes

$$
\begin{aligned}
&\frac{1}{\gamma}\mathbb{E}_{\varepsilon_1\sim N(0,I^{t-k})}[||L\sigma_{z_q}(c)\varepsilon_1||^2]+\frac{1}{\gamma}\mathbb{E}_{\varepsilon_2\sim N(0,I^{r-k})}[||L\sigma_{z_q}(x,c)_{t-k+1:r-k}\varepsilon_2||^2]+d\log(2\pi\gamma)\\
&-\log|\sigma^2_{z_q}(x,c)_{t-k+1:r-k}|-\log|\sigma^2_{z_q}(x,c)_{r-k+1:\kappa}|+\log|\sigma^2_{z_p}(c)_{r-k+1:\kappa}|+tr(\sigma^2_{z_q}(x)/\sigma^2_{z_p}(c)_{k+1:\kappa})\\
&+(\mu_{z_q(t-k+1:\kappa)}-\mu_{z_p(t-k+1:\kappa)})^T\sigma^2_{z_p}(c)^{-1}_{t-k+1:\kappa}(\mu_{z_q(t-k+1:\kappa)}-\mu_{z_p(t-k+1:\kappa)})+t-k-\kappa
\end{aligned}
\tag{24}
$$

Eq 24 shows that if we have a flexible enough prior, there are $t-k$ latent dimensions that won't provide any loss both in reconstruction and kl term. To minimize Eq 24, $\sigma^2_{z_q}(c)^*=0$ and $\sigma^2_{z_q}(x)^*_{t-k+1:r-k}=\gamma\frac{I}{L^2}$. Let them be the optimums, and view the terms that are irrelevant with γ when it approaches 0 as constants, we have

$$
\mathcal{L}^{u'}_c=(d-r+t)\log\gamma+O(1)
\tag{25}
$$

The lower bound to get the lower bound, we have

$$
\begin{aligned}
&\frac{1}{\gamma}\mathbb{E}_{q_{\phi_\gamma}(z|x,c)}[||x-\mu_x(z,c)||^2]+d\log(2\pi\gamma)+2\mathbb{KL}(q_\phi(z|x,c)||p(z|c))\\
\geq&d\log\gamma-\log|\sigma^2_{z_q}(x,c)_{t-k+1:r-k}|-\log|\sigma^2_{z_q}(x,c)_{r-k+1:\kappa}|+\log|\sigma^2_{z_p}(c)_{r-k+1:\kappa}|+\\
&tr(\sigma^2_{z_q}(x)/\sigma^2_{z_p}(c)_{k+1:\kappa})+(\mu_{z_q(t-k+1:\kappa)}-\mu_{z_p(t-k+1:\kappa)})^T\sigma^2_{z_p}(c)^{-1}_{t-k+1:\kappa}(\mu_{z_q(t-k+1:\kappa)}-\mu_{z_p(t-k+1:\kappa)})+O(1)\\
=&d\log\gamma-\log|\sigma^2_{z_q}(x,c)_{t-k+1:r-k}|+O(1)\\
\geq&d\log\gamma-(r-t)\log\gamma+O(1)\\
=&(d-r+t)\log\gamma+O(1)
\end{aligned}
\tag{26}
$$

The last inequality comes from the conclusion that there are at least $r-k$ active dimensions at a rate of $O(\gamma)$ and the loss is monotonously increase with γ. In this case k can be any integer in $[0,t]$, thus we cannot determine how many dimensions are used by the encoder and decoder separately. But no matter what value k is, the cost of CVAE is

$$
(d-r+t)\log\gamma+O(1)
$$

E Proof of Theorem 3

Recap of Theorem 3

Theorem 6 (Weight Sharing Compromises Performance of Sequential Modeling) *By sharing weights between the encoder and prior, the conditional VAE model has its cost being*

$$\mathcal{L}_c(\theta,\phi) = \int_{\mathcal{X}} \sum_l \{-\mathbb{E}_{q_\phi(z|x_{\leq l})}[\log p_\theta(x_l|z,x_{<l})] + \mathbb{KL}[q_\phi(z|x_{\leq l})||q_\phi(z|x_{<l})]\}\omega_{gt}(\mathrm{d}x) \quad (27)$$

Then we have $\mathcal{L}_c(\theta,\phi) = \Omega(1)$ for any θ and ϕ.

Summary of Proof In this section, we first define a space of sequences, and then separate the sequences into two categories according to the performance of the KL term. In Sec E.1, we analyze the case when the Kl term equals $O(\log\frac{1}{\gamma})$, and in Sec E.2, the rate of KL term is higher than $O(\log\frac{1}{\gamma})$. In both categories, we prove that the whole cost cannot go to $-\infty$.

Let $\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathcal{L}_c(\theta,\phi)$. Define $S \subset \mathcal{X}$ as the set of the sequences, and the sequence is defined as $\{x_l\}_{l=1}^{\infty} \in S$.

Consider when l equals to a constant l_0, we have the prior as $q_{\phi^*}(z|x_{<l_0})$, and encoder as $q_{\phi^*}(z|x_{\leq l_0})$. Next, consider $l = l_0+1$, we have the prior as $q_{\phi^*}(z|x_{\leq l_0})$, which is exactly the same as the encoder at $l = l_0$, and the encoder as $q_{\phi^*}(z|x_{\leq l_0+1})$. The cost function at these two points are

$$\mathcal{L}_c^{(l_0)}(\theta^*,\phi^*) = -\mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0})}[\log p_{\theta^*}(x_{l_0}|z,x_{<l_0})] + \mathbb{KL}[q_{\phi^*}(z|x_{\leq l_0})||q_{\phi^*}(z|x_{<l_0})]$$

and

$$\mathcal{L}_c^{(l_0+1)}(\theta^*,\phi^*) = -\mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0+1})}[\log p_{\theta^*}(x_{l_0+1}|z,x_{\leq l_0})] + \mathbb{KL}[q_{\phi^*}(z|x_{\leq l_0+1})||q_{\phi^*}(z|x_{\leq l_0})]$$

respectively.

We don't include the sequences within which all the samples share the same values into consideration in this theorem. Next, we separate the sequences with varied values into **two** cases.

E.1 KL term is at a rate of $O(\log\frac{1}{\gamma})$ when $\gamma \to 0$.

Denote $\log q_\phi(z|x_{\leq l}) - \log q_\phi(z|x_{<l}) = f_l(\gamma) = O(\log\frac{1}{\gamma})$. In this setting, we analyze the reconstruction term

$$\begin{aligned}
&-\mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0})}[\log p_{\theta^*}(x_{l_0}|z,x_{<l_0})] \\
=&\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0}) \log p_{\theta^*}(x_{l_0}|z,x_{<l_0})dz \\
=&\frac{1}{\gamma}\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0})[||x_{l_0} - \mu_x^*(z)||^2]dz + d\log(2\pi\gamma)
\end{aligned} \quad (28)$$

Similarly we have

$$\begin{aligned}
&-\mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0+1})}[\log p_{\theta^*}(x_{l_0+1}|z,x_{\leq l_0})] \\
=&\frac{1}{\gamma}\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0+1})[||x_{l_0+1} - \mu_x^*(z)||^2]dz + d\log(2\pi\gamma)
\end{aligned} \quad (29)$$

With the condition that $\log q_\phi(z|x_{\leq l}) - \log q_\phi(z|x_{<l}) = f_l(\gamma) = O(\log\frac{1}{\gamma})$, we have

$$\begin{aligned}
&\mathbb{KL}[q_\phi(z|x_{\leq l})||q_\phi(z|x_{<l})] \\
=&\mathbb{E}_{q_\phi(z|x_{\leq l})}\left[\log q_\phi(z|x_{\leq l}) - \log q_\phi(z|x_{<l})\right] \\
\leq&\mathbb{E}_{q_\phi(z|x_{\leq l})}[f_l(\gamma)] = f_l(\gamma)
\end{aligned} \quad (30)$$

That shows KL term is either small than a constant or goes to infinity at a slower rate than rate of $\log \frac{1}{\gamma}$ when $\gamma \to 0$. We can also get $q_\phi(z|x_{<l}) \geq \frac{1}{e^{f_l(\gamma)}} q_\phi(z|x_{<l})$, from which we have

$$q_\phi(z|x_{<l}) - q_\phi(z|x_{\leq l}) \geq q_\phi(z|x_{\leq l})(\frac{1}{e^{f_l(\gamma)}} - 1) \tag{31}$$

Together we have

$$\begin{aligned}
&- \mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0})}[\log p_{\theta^*}(x_{l_0}|z, x_{<l_0})] - \mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0+1})}[\log p_{\theta^*}(x_{l_0+1}|z, x_{\leq l_0})] \\
=&\frac{1}{\gamma}\left[\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0})||x_{l_0} - \mu_x^*(z)||^2 dz + \int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0+1})||x_{l_0+1} - \mu_x^*(z)||^2 dz\right] + 2d\log(2\pi\gamma) \\
=&\frac{1}{\gamma}[\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0+1}) \left[||x_{l_0} - \mu_x^*(z)||^2 + ||x_{l_0+1} - \mu_x^*(z)||^2\right] dz + \\
&\int_{\mathcal{Z}} [q_{\phi^*}(z|x_{\leq l_0}) - q_{\phi^*}(z|x_{\leq l_0+1})] \, ||x_{l_0} - \mu_x^*(z)||^2 dz] + 2d\log(2\pi\gamma) \\
&\int_{\mathcal{Z}} [q_{\phi^*}(z|x_{\leq l_0}) - q_{\phi^*}(z|x_{\leq l_0+1})] \, ||x_{l_0} - \mu_x^*(z)||^2 dz] + 2d\log(2\pi\gamma) \\
\geq&\frac{1}{\gamma}[\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0+1}) \left[||x_{l_0} - \mu_x^*(z)||^2 + ||x_{l_0+1} - \mu_x^*(z)||^2\right] dz + \\
&(\frac{1}{e^{f_{l_0}(\gamma)}} - 1) \int_{\mathcal{Z}} q_\phi(z|x_{\leq l_0+1})||x_{l_0} - \mu_x^*(z)||^2] dz + 2d\log(2\pi\gamma) \\
=&\frac{1}{\gamma}\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0+1}) \left[\frac{1}{e^{f_{l_0}(\gamma)}}||x_{l_0} - \mu_x^*(z)||^2 + ||x_{l_0+1} - \mu_x^*(z)||^2\right] dz + 2d\log(2\pi\gamma)
\end{aligned} \tag{32}$$

For any $l_0 = 1, 2, \ldots$ and all $z \in \mathcal{Z}$, we have the following cases:

1. **For any** $z \in \mathcal{Z}_1$**,** $\mu_x^*(z) = x_{l_0}$ **and** $\mu_x^*(z) \neq x_{l_0+1}$. We have $\frac{||x_{l_0} - \mu_x^*(z)||^2}{\gamma} \to \infty$ at a rate of $O(\frac{1}{\gamma})$.
2. **For any** $z \in \mathcal{Z}_2$**,** $\mu_x^*(z) = x_{l_0+1}$ **and** $\mu_x^*(z) \neq x_{l_0}$. We have $\frac{||x_{l_0+1} - \mu_x^*(z)||^2}{\gamma e^{f_{l_0}(\gamma)}} \to \infty$ at a rate of $O(\frac{1}{\gamma e^{f_{l_0}(\gamma)}})$.
3. **For any** $z \in \mathcal{Z}_3$**,** $\mu_x^*(z) \neq x_{l_0}$ **and** $\mu_x^*(z) \neq x_{l_0+1}$**.** Both cases above cause the norm term equal $\Omega(1)$.

With the setting, the lower bound of reconstruction term is

$$\begin{aligned}
&\frac{1}{\gamma}\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l_0+1}) \left[\frac{1}{e^{f_{l_0}(\gamma)}}||x_{l_0} - \mu_x^*(z)||^2 + ||x_{l_0+1} - \mu_x^*(z)||^2\right] dz + 2d\log(2\pi\gamma) \\
=&\frac{1}{\gamma}\int_{\mathcal{Z}_1} q_{\phi^*}(z|x_{\leq l_0+1})||x_{l_0+1} - \mu_x^*(z)||^2 dz + \frac{1}{\gamma e^{f_{l_0}(\gamma)}}\int_{\mathcal{Z}_2} q_{\phi^*}(z|x_{\leq l_0+1})||x_{l_0} - \mu_x^*(z)||^2 dz + \\
&\frac{1}{\gamma}\int_{\mathcal{Z}_3} q_{\phi^*}(z|x_{\leq l_0+1}) \left[\frac{1}{e^{f_{l_0}(\gamma)}}||x_{l_0} - \mu_x^*(z)||^2 + ||x_{l_0+1} - \mu_x^*(z)||^2\right] dz + 2d\log(2\pi\gamma) \\
\geq&\frac{1}{\gamma}\int_{\mathcal{Z}_1 \cup \mathcal{Z}_3} q_{\phi^*}(z|x_{\leq l_0+1})||x_{l_0+1} - \mu_x^*(z)||^2 dz + 2d\log(2\pi\gamma)
\end{aligned} \tag{33}$$

If (i) $\mathcal{Z}_1 \cup \mathcal{Z}_3$ is zero measured, (ii) for any $z \in \mathcal{Z}_1 \cup \mathcal{Z}_3$, $q_{\phi^*}(z|x_{\leq l_0+1}) = 0$ or (iii) $||x_{l_0+1} - \mu_x^*(z)||^2$ approaches zero, all the density will collapse to a zero-measured set, which may cause

over-fitting. In this case, there must exist a sequence $\{x_l\}^{i_0}$, in which $\sum_{l=1}^{\infty}\int_{\mathcal{Z}} q_{\phi^*}(z|x_{\leq l+1})||x_{l+1}-\mu_x^*(z)||^2dz \geq C$, where C is a constant, otherwise all the sequences $\{x_l\}^i \in S, i=1,2,\dots$ share the same values.

Then, for Eq 33, there must exist a constant C', such that

$$\int_{\mathcal{Z}_1\cup\mathcal{Z}_3} q_{\phi^*}(z|x_{\leq l_0+1})||x_{l_0+1}-\mu_x^*(z)||^2dz \geq C'$$

Thus, the lower bound of the cost is

$$\frac{C'}{\gamma} - 2d\log\frac{1}{2\pi\gamma}$$

When γ goes to zero, $O(\frac{1}{\gamma}) > O(\log\frac{1}{\gamma})$. We get the conclusion that $\mathcal{L}_c(\theta,\phi) = \int_{\mathcal{X}}\Omega(1)\omega_{gt}(dx)$ for any θ and ϕ.

E.2 KL term goes to infinity at a rate higher than $O(\log\frac{1}{\gamma})$.

In this case, we have

$$\begin{aligned}
&2\mathbb{KL}[q_{\phi^*}(z|x_{\leq l_0+1})||q_{\phi^*}(z|x_{\leq l_0})]\\
=&\log\frac{|\sigma_z^2(x_{\leq l_0})|}{|\sigma_z^2(x_{\leq l_0+1})|} - \kappa + (\mu_z(x_{\leq l_0+1})-\mu_z(x_{\leq l_0}))^T\sigma_z^{-2}(x_{\leq l_0})(\mu_z(x_{\leq l_0+1})-\mu_z(x_{\leq l_0}))\\
&+tr(\sigma_z^{-2}(x_{\leq l_0})\sigma_z^2(x_{\leq l_0+1}))
\end{aligned} \tag{34}$$

Thus it can only happen when there are some dimensions where $\sigma_z^2(x_{\leq l_0})$ is active while $\sigma_z^2(x_{\leq l_0+1})$ is not, which indicate that $tr(\sigma_z^{-2}(x_{\leq l_0})\sigma_z^2(x_{\leq l_0+1})) \to \infty$ at a rate of $\Omega(\frac{1}{\gamma})$. We have

$$\begin{aligned}
&-2\mathbb{E}_{q_{\phi^*}(z|x_{\leq l_0+1})}[\log p_{\theta^*}(x_{l_0+1}|z,x_{\leq l_0})] + 2\mathbb{KL}[q_{\phi^*}(z|x_{\leq l_0+1})||q_{\phi^*}(z|x_{\leq l_0})]\\
\geq& d\log(2\pi\gamma) + tr(\sigma_z^{-2}(x_{\leq l_0})\sigma_z^2(x_{\leq l_0+1})) + \log\frac{|\sigma_z^2(x_{\leq l_0})|}{|\sigma_z^2(x_{\leq l_0+1})|}\\
&-\kappa + (\mu_z(x_{\leq l_0+1})-\mu_z(x_{\leq l_0}))^T\sigma_z^{-2}(x_{\leq l_0})(\mu_z(x_{\leq l_0+1})-\mu_z(x_{\leq l_0}))
\end{aligned} \tag{35}$$

Because $d\log(2\pi\gamma) + \log\frac{|\sigma_z^2(x_{\leq l_0})|}{|\sigma_z^2(x_{\leq l_0+1})|} \to -\infty$ at a rate of $O(\log\gamma)$ while $tr(\sigma_z^{-2}(x_{\leq l_0})\sigma_z^2(x_{\leq l_0+1})) \to \infty$ at a rate of $\Omega(\frac{1}{\gamma})$, the whole loss will go to infinity.

In summary, in both cases, when summing over l, $\mathcal{L}_c(\theta,\phi)$ will go to infinity.

F Proof of Corollary 2.1

Recap of Corollary 2.1

Corollary 2.1 (Adaptive Active Dimension) *Suppose there exists conditioning variables $c_1 \in \mathcal{C}_{t_1}$ and $c_2 \in \mathcal{C}_{t_2}$, i.e. c_1, c_2 have t_1 and t_2 effective dimensions respectively. Given a CVAE model with optimal solution $\{\theta^*,\phi^*\}$, we have the number of active dimensions of $z_1 \sim q_{\phi^*}(z|x,c_1)$ equals to $r-t_1$, and that of $z_2 \sim q_{\phi^*}(z|x,c_2)$ equals to $r-t_2$.*

Under the setting of Theorem 5, the CVAE model is arbitrarily complex. Denote $\mathcal{X}_{t_1}$ and $\mathcal{X}_{t_2}$, where $\mathcal{X}_{t_1}\cup\mathcal{X}_{t_2}=\mathcal{X}$, s.t. for any $x\in\mathcal{X}_{t_1}$, the c of pair $\{x,c\}$ is from $\mathcal{C}_{t_1}$, and the same setting for $\mathcal{X}_{t_2}$. Then for any $x\in\mathcal{X}_{t_1}$ and $x\in\mathcal{X}_{t_2}$. Denote $t = t_1\omega(\mathcal{X}_{t_1}) + t_2\omega(\mathcal{X}_{t_2})$, which is a weighted-average of t_1 and t_2. We have

Thus we have

$$
\begin{aligned}
\mathcal{L}(\theta^*, \phi^*) &= \int_{\mathcal{X}} (d - r + t) \log \gamma + O(1)\omega(dx) \\
&= \int_{\mathcal{X}} (d - r) \log \gamma + O(1)\omega(dx) + (t_1 \omega(\mathcal{X}_{t_1}) + t_2 \omega(\mathcal{X}_{t_2})) \log \gamma \\
&= \int_{\mathcal{X}_{t_1}} (d - r + t_1) \log \gamma + O(1)\omega(dx) + \int_{\mathcal{X}_{t_2}} (d - r + t_2) \log \gamma + O(1)\omega(dx)
\end{aligned} \tag{36}
$$

Thus the active dimension of $z_1 \sim q_{\phi^*}(z|x, c_1)$ equals $r - t_1$, and that of $z_2 \sim q_{\phi^*}(z|x, c_2)$ equals $r - t_2$.

G Justification of Remark 1

Consider a κ-simple CVAE model with encoder $q_\phi(z|x, c)$, prior $p_\theta(z|c)$ and decoder $p_\theta(x|z, c)$. Further, let $\mu_q(x, c), \sigma_q(x, c)$ be the distributional parameters for $z \sim q_\phi(z|x, c)$, $\mu_p(c), \sigma_p(c)$ be the distributional parameters for $z \sim p_\theta(z|c)$. Name this model as M, and we have its cost with regard to (θ, ϕ) being

$$
\begin{aligned}
2\mathcal{L}_c(M; \theta, \phi) =& \int_{\mathcal{X}} \{-2\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)] + 2\mathbb{KL}[q_\phi(z|x, c)||p_\theta(z|c)]\}\omega_{gt}(dx) \\
=& \frac{1}{\gamma} \int_{\mathcal{X}} \int_{\mathcal{Z}} \mathcal{N}(z; \mu_q(x, c), \sigma_q(x, c))||x - \mu_x(z)||^2 dz \omega_{gt}(dx) \\
&+ \log(2\pi\gamma) + \int_{\mathcal{X}} 2\mathbb{KL}[q_\phi(z|x, c)||p_\theta(z|c)]\omega_{gt}(dx) \\
=& \frac{1}{\gamma} \int_{\mathcal{X}} \int_{\mathcal{Z}} \frac{1}{\sqrt{(2\pi\gamma)^d}} \exp\{-\frac{||z - \mu_q||^2}{2\sigma_q^2}\}||x - \mu_x(z)||^2 dz \omega_{gt}(dx) \\
&+ \log(2\pi\gamma) + \int_{\mathcal{X}} [\log \sigma_p - \log \sigma_q - \kappa + ||\mu_q - \mu_p||^2/\sigma_p + \mathrm{tr}(\sigma_q/\sigma_p)]\omega_{gt}(dx)
\end{aligned} \tag{37}
$$

Next, we construct another κ-simple CVAE, M', with a standard Gaussian prior, only using computation modules in M. Specifically, the new prior, decoder, and encoder are defined as:

- Prior: $p'(z') = \mathcal{N}(0, \mathrm{I})$
- Decoder: $p'(x|z', c) = p_\theta(x|z' * \sigma_p(c) + \mu_p(c), c)$
- Encoder: $q'(z|x, c) = \mathcal{N}(\mu'_q, \sigma'_q)$, where
  - $\mu'_q = (\mu_q(x, c) - \mu_p(c))/\sigma_p(c)$
  - $\sigma'_q = \sigma_q(x, c)/\sigma_p(c)$

With M' defined, we are going to show that it has the exact same cost value as the above one during training, i.e. $\mathcal{L}(M'; \theta, \phi) = \mathcal{L}(M; \theta, \phi)$, and the generated data distribution during generation, i.e. $p'(x|z', c)\mathcal{N}(z'; 0, \mathrm{I}) \equiv p_\theta(x|z, c)p_\theta(z|c)$.

During training, we have $z' \sim \mathcal{N}(\mu'_q, \sigma'_q)$, thus $z' * \sigma_p(c) + \mu_p(c) \sim \mathcal{N}((\mu_q(x, c) - \mu_p(c))/\sigma_p(c) * \sigma_p(c) + \mu_p(c), \sigma_q(x, c)/\sigma_p(c) * \sigma_p(c)) = \mathcal{N}(\mu_q(x, c), \sigma_q(x, c))$. Thus, we have

$$
\begin{aligned}
&\mathbb{E}_{q(z'|x,c)}[\log p_\theta(x|z' * \sigma_p(c) + \mu_p(c), c)] \\
=& \frac{1}{\gamma} \int_{\mathcal{Z}} \mathcal{N}(z'; \mu', \sigma')||x - \mu_x(z' * \sigma_p(c) + \mu_p(c))||^2 dz' + \log(2\pi\gamma) \\
=& \frac{1}{\gamma} \int_{\mathcal{Z}} \mathcal{N}(z; \mu_q(x, c), \sigma_q(x, c))||x - \mu_x(z)||^2 dz + \log(2\pi\gamma) \\
=& \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)]
\end{aligned} \tag{38}
$$

Besides, we also have

$$
\begin{aligned}
&\mathbb{KL}[q'(z'|x,c)||\mathcal{N}(0,\mathrm{I})] \\
=&\mathbb{KL}[\mathcal{N}(\mu_q(x,c)-\mu_p(c))/\sigma_p, \sigma_q(x,c)/\sigma_p(c)||\mathcal{N}(0,\mathrm{I})] \\
=&\frac{1}{2}[||\mu_q(x,c)-\mu_p(c))/\sigma_p||^2 + \mathrm{tr}(\sigma_q/\sigma_p) - \kappa - \log(\sigma_q/\sigma_p)] \\
=&\frac{1}{2}[\log\sigma_p - \log\sigma_q - \kappa + ||\mu_q-\mu_p||^2/\sigma_p + \mathrm{tr}(\sigma_q/\sigma_p)] \\
=&\mathbb{KL}[q_\phi(z|x,c)||p_\theta(z|c)]
\end{aligned}
\tag{39}
$$

In terms of generation equivalence, for any $z'_p \sim \mathcal{N}(0,\mathrm{I})$, we have

$$
\begin{aligned}
&p'(x|z',c)p(z';0,\mathrm{I}) \\
=&p_\theta(x|z'*\sigma_p(c)+\mu_p(c))p(z';0,\mathrm{I}) \\
=&p_\theta(x|z)p(z;\mu_p(c),\sigma_p(c)) \\
=&p_\theta(x|z)p_\theta(z|c)
\end{aligned}
\tag{40}
$$

Therefore we conclude that M' and M share the same cost value, i.e. $\mathcal{L}_c(M';\theta,\phi) = \mathcal{L}_c(M;\theta,\phi)$, and equivalent data generation distributions.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4 and Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Section 4
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]