

---

# Information-theoretic Generalization Analysis for VQ-VAEs: A Role of Latent Variables

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Latent variables (LVs) play a crucial role in encoder–decoder models by enabling  
2 effective data compression, prediction, and generation. Although their theoretical  
3 properties, such as generalization, have been extensively studied in supervised  
4 learning, similar analyses for unsupervised models such as variational autoencoders  
5 (VAEs) remain insufficiently underexplored. In this work, we extend information-  
6 theoretic generalization analysis to vector-quantized (VQ) VAEs with discrete  
7 latent spaces, introducing a novel data-dependent prior to rigorously analyze the  
8 relationship among LVs, generalization, and data generation. We derive a novel  
9 generalization error bound of the reconstruction loss of VQ-VAEs, which depends  
10 solely on the complexity of LVs and the encoder, independent of the decoder.  
11 Additionally, we provide the upper bound of the 2-Wasserstein distance between  
12 the distributions of the true data and the generated data, explaining how the regu-  
13 larization of the LVs contributes to the data generation performance.

## 14 1 Introduction

15 Encoder–decoder (ED) models have demonstrated remarkable performance [23] in (un)supervised  
16 tasks such as classification [2, 4] and data generation [39, 68], compressing input data into latent  
17 variables (LVs) via an encoder. The success of ED models hinges on how effectively the encoder can  
18 represent essential features of the input in LVs, stimulating analyses of the relationship between LVs  
19 and ED model performance, as well as developing algorithms designed to appropriately control LVs.

20 In supervised learning, the information bottleneck (IB) hypothesis [65, 59] has gained significant  
21 attention for proposing that minimizing the mutual information (MI) between input data and LVs  
22 enhances generalization by ensuring LVs retain the minimal information necessary for prediction.  
23 This hypothesis has motivated numerous learning algorithms for deep neural networks and empirical  
24 studies exploring their performance [64, 60, 55, 22, 1, 2]. Moreover, theoretical research about how  
25 LVs contribute to generalization has been actively pursued [69, 28, 38, 70] within the IB hypothesis.  
26 Recently, Sefidgaran et al. [56] has highlighted the limitations of these analyses, particularly in  
27 terms of assumptions and the sample complexity represented by the MI. To address these limitations,  
28 they proposed extending the supsample setting of information-theoretic (IT) analysis [62]. Their  
29 approach induces a *symmetric, data-dependent prior over LVs* that facilitates rigorous analysis, which  
30 successfully characterizes generalization performance using the Kullback–Leibler (KL) divergence  
31 between the posterior distribution of the LVs and this prior. These results suggest that, by carefully  
32 constructing the data-dependent prior distribution, we can obtain **a decoder-independent bound**,  
33 which illustrates clearly how LVs contribute to the generalization for ED models in classification.  
34 Their analysis has recently been extended to multi-view learning settings [57, 58].

35 LVs play a key role in deep generative models for unsupervised learning tasks such as data compres-  
36 sion and generation. For example, variational autoencoders (VAEs) [39] are trained by optimizing an

objective function that includes the KL divergence of the posterior from the prior in the LV space as a regularization term. Extended methods such as  $\beta$ -VAE [34] highlight the importance of appropriately tuning the strength of KL regularization to improve LV representations. Additionally, methods like vector-quantized VAEs (VQ-VAEs) [68], which discretize the latent space, have been developed to address posterior collapse. Numerous empirical studies have also evaluated model performance based on the MI, such as the IB hypothesis and rate-distortion theory [3, 9, 67, 12].

In contrast to supervised learning, theoretical insights into the relationship between the generalization of ED models and LVs in unsupervised learning remain limited. Although Chérif-Abdellatif et al. [13] has employed *probably approximately correct* (PAC) Bayes analysis [47, 6] to investigate the generalization error defined in terms of reconstruction loss, they consider the posterior and prior distributions over the *encoder and decoder parameters*. Similarly, Epstein & Meir [19] focused on the complexity of encoder and decoder parameters to analyze the generalization capability. Therefore, these studies lack the analysis of the relationship between LVs and generalization capability. Mbacke et al. [46] attempted to address this problem by deriving PAC-Bayes bounds based on the KL divergence within prior and posterior distributions over LVs; however, their analysis relies on the impractical assumption that decoders are not trained, leaving significant challenges in achieving a practical understanding of the role of LVs in generalization performance.

To address these challenges, we provide the first rigorous theoretical analysis of the relationship among LVs, generalization, and data generation in ED models, with a focus on VQ-VAEs [68]. Motivated by Sefidgaran et al. [56], we construct a data-dependent prior over LVs using the supersample setting from IT analysis [62, 30, 32]. This approach yields a generalization error bound for the reconstruction loss, characterized by the KL divergence between the prior and the posterior over LVs (Theorem 2). Similar to Sefidgaran et al. [56], our bound remains independent of decoder complexity even when the encoder and decoder are trained jointly, underscoring the critical role of designing the encoder network for the generalization.

However, we observe that the bound based on the supersample setting does not necessarily converge to 0 asymptotically with respect to the number of samples. To address this issue, we extend the supersample framework by introducing a novel data-dependent prior, called the *permutation symmetric prior distribution*, which explicitly accounts for the inherent symmetries specific to unsupervised learning tasks (Theorem 3). This formulation enables us to derive a generalization error bound that asymptotically converges to 0 as the number of samples increases and is independent of the decoder.

Finally, we investigate the data generation capability of VQ-VAEs by deriving the upper bound on the 2-Wasserstein distance between the true data and the generated data distributions (Theorem 5). Our analyses reveal that the generalization and data-generating capabilities of VQ-VAEs depend solely on the parameters of the encoder and LVs, *remaining entirely independent of the decoder*.

## 2 Background

In this section, we introduce the VQ-VAE and define the reconstruction-based generalization error, which forms the basis of our analysis (Sections 2.1 and 2.2). We then present the IT analysis using *supersamples* (Section 2.3), highlighting its limitations in unsupervised settings (Section 2.4).

**Notations:** We use uppercase letters for random variables and lowercase letters for their realizations. The distribution of  $X$  is denoted by  $p(X)$ , and the conditional distribution of  $Y$  given  $X$  by  $p(Y|X)$ . Expectations are written as  $\mathbb{E}_{p(X)}$  or  $\mathbb{E}_X$ . The MI and conditional MI (CMI) are denoted by  $I(X; Y)$  and  $I(X; Y|Z)$ , respectively. The KL divergence from  $p(X)$  to  $p(Y)$  is written as  $\text{KL}(p(X)||p(Y))$ . For  $a \in \mathbb{N}$ , we define  $[a] := \{1, \dots, a\}$ .

### 2.1 VQ-VAE and its stochastic extensions

Let  $\mathcal{X} \subset \mathbb{R}^d$  denote the data space, and assume an unknown data-generating distribution  $\mathcal{D}$ . The latent space is represented as  $\mathcal{Z} \subset \mathbb{R}^{d_z}$ , where both  $\mathcal{X}$  and  $\mathcal{Z}$  are equipped with the Euclidean metric  $\|\cdot\|$ . The discrete latent space comprises  $K$  distinct points, collectively referred to as the *codebook*, denoted by  $\mathbf{e} = \{e_j\}_{j=1}^K \in \mathcal{Z}^K$ , which are learned from the training data.

The VQ-VAE model consists of the encoder network  $f_\phi: \mathcal{X} \rightarrow \mathcal{Z}$  and the decoder network  $g_\theta: \mathcal{Z} \rightarrow \mathcal{X}$  responsible for (i) data compression and (ii) reconstruction, where  $\phi \in \Phi \subset \mathbb{R}^{d_\phi}$  and  $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ .

denote the parameters of the encoder and decoder, respectively. In the compression phase, a data point  $x$  is mapped to  $f_\phi(x)$ , and the discrete representation  $e_j$  is selected from the codebook  $\mathbf{e}$ . Then, the posterior distribution of the discrete representation indexed by  $j$  is denoted as  $q(J = j|\mathbf{e}, \phi, x)$  for all  $j = 1, \dots, K$ . In the original VQ-VAE [68], the following deterministic posterior is used:

$$q(J = j|\mathbf{e}, \phi, x) = \begin{cases} 1 & \text{for } j = \operatorname{argmin}_{k \in [K]} \|f_\phi(x) - e_k\|, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where the distance between the encoder output and the codebook entries determines the posterior. Recent extensions of VQ-VAE [79, 61, 54, 63] introduce a stochastic posterior defined by

$$q(J = j|\mathbf{e}, \phi, x) \propto \exp(-\beta \|f_\phi(x) - e_j\|^2), \quad (2)$$

where a softmax is applied over codebook indices, and the temperature parameter  $\beta \in \mathbb{R}^+$  controls the level of stochasticity. The data is then reconstructed by passing the selected latent representation  $e_{J=j}$  through the decoder, resulting in  $g_\theta(e_{J=j})$ . The fidelity of the reconstruction to the original input is measured by the *reconstruction loss*, defined as  $l(x, g_\theta(e_{J=j}))$ , where  $l : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ .

## 2.2 Generalization error based on reconstruction loss

Hereafter, let the set of parameters be denoted as  $W := \{\mathbf{e}, \phi, \theta\} \in \mathcal{W} := \mathcal{Z}^K \times \Phi \times \Theta$ . Given the training dataset  $S = (S_1, \dots, S_n) \in \mathcal{X}^n$  consisting of independently and identically distributed (i.i.d.) data points sampled from the data distribution  $\mathcal{D}$ , these parameters are learned jointly using a randomized algorithm  $\mathcal{A} : \mathcal{X}^n \rightarrow \mathcal{W}$  that minimizes the reconstruction loss between a data point  $x$  and a reconstructed data  $g_\theta(e_j)$ , i.e.,  $l(x, g_\theta(e_j))$ . Consequently, the learned parameters  $\mathbf{e}, \phi, \theta$  follow the conditional distribution  $q(\mathbf{e}, \phi, \theta|S)$ . For simplicity, we define the expected reconstruction loss for an input  $x$  and  $w$  as  $l_0 : \mathcal{W} \times \mathcal{X} \rightarrow \mathbb{R}$ , where  $l_0(w, x) := \mathbb{E}_{q(J|\mathbf{e}, \phi, x)}[l(x, g_\theta(e_J))]$ . In this study, we consider the squared distance as  $l$ . Accordingly, our objective is to minimize  $l_0(w, x) := \mathbb{E}_{q(J|\mathbf{e}, \phi, x)}[\|x - g_\theta(e_J)\|^2]$  over the training dataset  $x \in S$ . We introduce the following assumption about the data space imposed on our analysis.

**Assumption 1.** *There exists a positive constant  $\Delta$  such that  $\sup_{x, x' \in \mathcal{X}} \|x - x'\| < \Delta^{1/2}$ .*

This assumption ensures that the reconstruction loss  $l(x, g_\theta(e_j))$  is bounded by  $\Delta$  for all  $x, e_j$ , and  $\theta$ . Our goal is to theoretically characterize the relationship between generalization performance and LVs in VQ-VAEs. To this end, we analyze the following generalization error:

$$\operatorname{gen}(n, \mathcal{D}) := \left| \mathbb{E}_{S, X} \mathbb{E}_{q(W|S)} l_0(W, X) - \frac{1}{n} \sum_{m=1}^n l_0(W, S_m) \right|, \quad (3)$$

where the first term denotes the expected test reconstruction loss, and the second term is the empirical training loss. Following the success of Sefidgaran et al. [56], we also consider analyzing Eq. (3) under the IT analysis framework with the *supersample* (or ghost sample) setting [62, 30, 32].

## 2.3 Supersample settings for IT analysis

Now, we introduce the supersample setting for IT analysis. The supersample  $\tilde{X} \in \mathcal{X}^{n \times 2}$  is defined as an  $n \times 2$  matrix, where each entry is drawn i.i.d. from  $\mathcal{D}$ . Each column of  $\tilde{X}$  is associated with an index set  $\{0, 1\}$ , determined by  $U = (U_1, \dots, U_n) \sim \operatorname{Uniform}(\{0, 1\}^n)$ , independent of  $\tilde{X}$ . In this setting, we consider  $\tilde{X}_U := (\tilde{X}_{m, U_m})_{m=1}^n$  as the training dataset and  $\tilde{X}_{\bar{U}} := (\tilde{X}_{m, \bar{U}_m})_{m=1}^n$  as the test dataset, where  $\bar{U}_m = 1 - U_m$ . The  $m$ -th row of  $\tilde{X}$  is denoted as  $\tilde{X}_m$  with entries  $(\tilde{X}_{m,0}, \tilde{X}_{m,1})$ . Using the supersamples,  $l_0(\mathcal{A}(\tilde{X}_U), \tilde{X})$  denotes an  $n \times 2$  loss matrix, where  $l_0(\mathcal{A}(\tilde{X}_U), \cdot)$  is applied elementwise to  $\tilde{X}$ . The IT analysis of Eq. (3) under the supersample setting gives the following result.

**Theorem 1** (Hellström & Durisi [32]). *Under Assumption 1 and the supersample setting, we have*

$$\operatorname{gen}(n, \mathcal{D}) \leq \Delta \sqrt{2I(l_0(W, \tilde{X}); U|\tilde{X})/n}. \quad (4)$$

The complete proof is provided in Appendix C. We refer to this bound as the **basic IT-bound**, as it arises from the direct application of existing IT analysis [32] developed for supervised learning. Unfortunately, we find that the basic IT-bound is insufficient to fully understand the role of LVs in the generalization performance of VQ-VAE. The next section elaborates on this limitation.



Figure 1: Graphical models illustrating different dependency structures for LVs. The left panel shows the structure considered in the standard supersample setting (Theorem 1). The right panel depicts our proposed structure tailored for unsupervised learning. See Appendix B.3 for further details.

## 2.4 Limitation of the direct application of IT analysis

The limitation of the basic IT-bound is that it does not offer a clear interpretation of how the LVs contribute to the generalization performance independently of other random variables. Specifically, let  $\tilde{J}$  denote the random variable that follows the distribution  $q(\tilde{J}|\mathbf{e}, \phi, \tilde{X})$ , which is defined by applying  $q(J|\mathbf{e}, \phi, \cdot)$  elementwise to  $\tilde{X}$ . With this definition, we can upper bound Eq. (4) as

$$I(l_0(W, \tilde{X}); U|\tilde{X}) \leq I(\theta; U|\tilde{X}) + I(\tilde{J}; U|\tilde{X}, \theta). \quad (5)$$

See Appendix C.2 for the proof. This result implies that the generalization of VQ-VAE can be bounded by the CMI related to the decoder parameter  $\theta$  and the selected index  $\tilde{J}$ . Note that selecting  $J$  corresponds to selecting an LV  $e_J$  from the codebook. Therefore, the second term above illustrates how LVs contribute to generalization. However, since conditioning on  $\theta$  is taken, it does not allow the independent analysis of  $e_J$  and  $\theta$ . This dependence hinders a precise theoretical analysis of how LVs affect generalization performance.

We can better understand this difficulty by considering how IT-based generalization analysis is typically formulated: it is framed as the problem of inferring which samples were used for training, given a random supersample index,  $U$ , that determines the shuffling of the dataset. The randomness introduced by this shuffling is governed by the design of the prior, which plays a central role in applying the Donsker–Varadhan inequality to derive an upper bound on the generalization error. In the basic IT-bound (Theorem 1), shuffling via  $U$  leads to randomly altering the training dataset, producing a bound that jointly depends on both model parameters and LVs, thereby entangling  $\theta$  and  $J$ . This illustrates that a straightforward extension of standard IT analysis is insufficient to isolate the contribution of LVs to generalization, motivating the development of a new analytical framework.

## 3 Proposed IT analysis under supersamples and its limitations

In this section, we first present the results of our generalization analysis for VQ-VAE (Section 3.1). We then offer a detailed interpretation of the resulting generalization error bound and discuss its limitations (Section 3.2). All corresponding proofs are provided in Appendix D.

### 3.1 Our supersample setting and result

As discussed in Section 2.4, the naive application of the existing supersample setting in IT analysis is insufficient to capture the role of LVs. To address this limitation, we introduce posterior and prior distributions over  $J$  that explicitly encode the dependence between the supersample index  $U$  and the LVs, on the basis of the approach of Sefidgaran et al. [56].

To this end, we define the following posterior distributions based on both  $\tilde{X}_U$  and  $\tilde{X}_{\bar{U}}$ :  $q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U) := \prod_{m=1}^n q(J_m|\mathbf{e}, \phi, \tilde{X}_{m,U_m})$  and  $q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}}) := \prod_{m=1}^n q(\bar{J}_m|\mathbf{e}, \phi, \tilde{X}_{m,\bar{U}_m})$ . For notational simplicity, we write  $\mathbf{Q}_{\mathbf{J},U} := q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U)$ . We then define the following joint distribution to capture the dependence of the LVs on both  $\tilde{X}_U$  and  $\tilde{X}_{\bar{U}}$ :  $\mathbf{Q}_{\bar{\mathbf{J}},U} := q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}}) \cdot q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U)$ .

We consider two types of prior distribution to facilitate the analysis of VQ-VAEs: a *data-independent prior*  $\mathbf{P}$  and a *data-dependent prior*  $\mathbf{Q}_{\bar{\mathbf{J}}}$  defined as

$$\mathbf{P} := \prod_{m=1}^n \pi(J_m|\mathbf{e}, \phi), \quad \text{and} \quad \mathbf{Q}_{\bar{\mathbf{J}}} := \mathbb{E}_U \mathbf{Q}_{\bar{\mathbf{J}},U} = \mathbb{E}_U q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}}) q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U), \quad (6)$$

where  $\pi(J_m|\mathbf{e}, \phi)$  denotes an *arbitrary* distribution over LVs that is independent of both  $\tilde{X}$  and the supersample index  $U$ . For the data-dependent prior, we adopt the supersample setting specifically

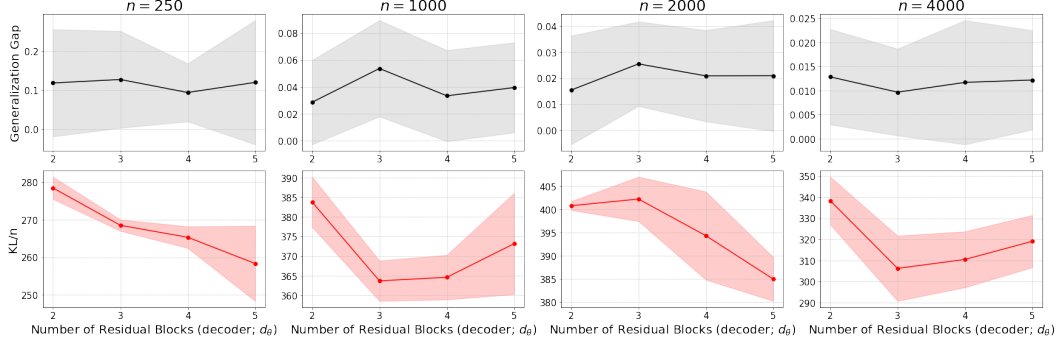


Figure 2: The behavior of the generalization gap and the empirical KL term ( $\text{KL}(\mathbf{Q}_{\mathbf{J},U}||\mathbf{P})/n$ ) on the MNIST dataset when increasing the number of residual blocks to enlarge the decoder dimension  $d_\theta$  ( $K = 128$ ,  $d_z = 64$ ). See Appendix G for detailed experimental settings.

tailored to the LVs. The basis for introducing both types of prior is discussed following the main theorem. Figure 1 illustrates the distinction in LV dependencies between the conventional supersample setting (as used in Theorem 1) and our approach. The central idea is to apply supersample-based shuffling to the LVs directly. Under these settings, the following is our main result.

**Theorem 2.** *Under Assumption 1 and the supersample setting, we have*

$$\text{gen}(n, \mathcal{D}) \leq 2\Delta \sqrt{\frac{\mathbb{E}_{\tilde{X},U} \mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)} (\text{KL}(\mathbf{Q}_{\mathbf{J},U}||\mathbf{P}) + \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}},U}||\mathbf{Q}_{\tilde{\mathbf{J}}}))}{n}} + \frac{\Delta}{\sqrt{n}}. \quad (7)$$

The upper bound comprises two distinct complexity terms. The first,  $\text{KL}(\mathbf{Q}_{\mathbf{J},U}||\mathbf{P})$ , captures the complexity of the LVs. The second,  $\text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}},U}||\mathbf{Q}_{\tilde{\mathbf{J}}})$ , reflects the complexity of the LVs and the degree of overfitting when learning parameters  $\mathbf{e}$  and  $\phi$ , as we will further discuss in Section 3.2.

Consistent with the findings of Sefidgaran et al. [56], our bound is independent of the decoder  $g_\theta$ . This indicates that increasing the complexity of  $g_\theta$  has a limited effect on the generalization performance. Empirical results in the top line of Figure 2 support this implication: adding a single ResBlock—introducing approximately 74,000 additional parameters—has a negligible effect on the generalization gap. Further experiments across various datasets and decoder architectures in Appendix G reinforce this observation.

We emphasize that our results *do not imply that the decoder is unimportant*. Although our generalization bound is independent of decoder complexity, a sufficiently expressive decoder is still required to fit the training data. Otherwise, the test loss may remain high since  $\text{Test Loss} \leq \text{Training Loss} + \text{Generalization Gap}$ . Our analysis specifically focuses on bounding the generalization gap, under the implicit assumption that the decoder can adequately fit the training data. In practice, this suggests that improving generalization in VQ-VAEs hinges more on careful encoder design, since overly complex encoders can increase the KL divergence of the LVs. We discuss this point further in Section 6.

**Why two types of prior are required:** Our proof reveals that isolating the LVs from the decoder parameter and obtaining a decoder-independent generalization bound requires the prior to satisfy two essential conditions: (A) it allows random shuffling without changing the LV distribution, and (B) it supports a swap between training and test samples to assess overfitting. From this perspective, the shuffling induced by  $U$  in the basic IT-bound (Theorem 1) satisfies condition (B) but violates condition (A), as it changes the distribution of LVs. To address this issue, the proof of Theorem 2 decomposes the generalization gap into two components: the term associated with condition (A), which is controlled using a data-independent prior  $\mathbf{P}$ , and the term associated with condition (B), which is controlled using a data-dependent prior  $\mathbf{Q}_{\tilde{\mathbf{J}}}$ . By combining both priors, we can derive the final upper bound in Eq. (7). For a detailed explanation, see Appendices B.3 and D.1.

**Remark 1.** When  $K = 1$ , VQ-VAEs map all input data to the same LV, effectively estimating the low-dimensional mean of the data distribution. In this case, the generalization error should not



depend on the decoder. It is straightforward to show—without using our IT-based analysis—that  $\text{gen}(n, \mathcal{D}) = O(1/\sqrt{n})$ . Notably, our bound in Eq. (7) correctly reflects this behavior, as the square root term vanishes when  $K = 1$  (see Appendix C.3 for details).

### 3.2 Further analyses of our bound and limitations on convergence

In this section, we further analyze the properties of the two KL divergence terms in Theorem 2 and discuss their asymptotic behavior as the sample size  $n$  increases.

**Regarding  $\text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}},U} \parallel \mathbf{Q}_{\tilde{\mathbf{J}}})$ :** We can derive the following upper bound:

$$\mathbb{E}_{\tilde{X},U} \mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)} \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}},U} \parallel \mathbf{Q}_{\tilde{\mathbf{J}}}) \leq I(\mathbf{e}, \phi; U | \tilde{X}) + I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X}). \quad (8)$$

Since  $\tilde{X}_U = S$ , the data processing inequality implies that  $I(\mathbf{e}, \phi; U | \tilde{X}) \leq I(\mathbf{e}, \phi; S)$ . This quantity captures how much information about the training data is retained in the encoder, thereby reflecting the degree of overfitting of the encoder parameters. The term  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$  can be viewed as a regularization term for the LVs, analogous to the IB hypothesis; see Appendix D.6 for further details.

Next, we investigate whether each term in Eq. (8) exhibits asymptotic convergence as the sample size  $n$  increases, which is a key requirement for a valid generalization error bound. We begin by analyzing the asymptotic behavior of  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$ .

**Lemma 1.** *Let the posterior distribution over  $J$  be deterministic as defined in Eq. (1), and we denote the composition of this mapping with the encoder  $f_\phi$  by  $f'_{\mathbf{e},\phi} : \mathcal{X} \rightarrow [K]$ . If the function class to which  $f'_{\mathbf{e},\phi}$  belongs has a finite Natarajan dimension, then  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})/n = O(\log n/n)$ .*

This result implies that if the encoder is appropriately regularized, the quantity  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})/n$  converges asymptotically to zero. We also empirically evaluated this term in practical settings (see Appendix G) and observed that it indeed decreases as the sample size  $n$  increases.

Next, the CMI term  $I(\mathbf{e}, \phi; U | \tilde{X})$  has been extensively analyzed under the standard supersample setting of IT analysis [62]. Prior works have established its asymptotic convergence through various approaches, including algorithmic stability [62], analyses of specific optimization methods such as stochastic gradient descent (SGD) [75] and stochastic gradient Langevin dynamics (SGLD) [20], and complexity-based arguments using covering numbers [80], all showing that  $I(\mathbf{e}, \phi; U | \tilde{X})/n \rightarrow 0$  as  $n \rightarrow \infty$ . In conclusion, the term  $\mathbb{E}_{\tilde{X},U} \mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)} \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}},U} \parallel \mathbf{Q}_{\tilde{\mathbf{J}}})/n$  can be shown to converge asymptotically under certain algorithmic conditions. For a detailed discussion, see Appendix D.8.

**Regarding  $\text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P})$ :** This term can be rewritten as  $\text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P})/n = \frac{1}{n} \sum_{m=1}^n \text{KL}(q(J_m | \mathbf{e}, \phi, S_m) \parallel \pi(J_m | \mathbf{e}, \phi))$ , where the training data is selected via  $U$ , i.e.,  $\tilde{X}_U = S = (S_1, \dots, S_n)$ . This quantity corresponds to the *empirical KL divergence*, which also appears in the analysis of Mbacke et al. [46], and reflects the complexity of the LVs. Such a term is commonly used as the regularization term appearing in many VAE training procedures [39, 33, 63].

A key factor in minimizing  $\text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P})$  is the choice of the prior  $\mathbf{P}$ . In VQ-VAEs, a uniform distribution is typically adopted [63]; however, is this choice optimal for minimizing the KL divergence? The following lemma addresses this question.

**Lemma 2.** *Assume that for any fixed training dataset  $S = (s_1, \dots, s_n)$  and any permutation  $\tau$ , the posterior satisfies permutation invariance, i.e.,  $q(\mathbf{e}, \phi, \theta | S) = q(\mathbf{e}, \phi, \theta | S^\tau)$ , where  $S^\tau = (s_{\tau_1}, \dots, s_{\tau_n})$ . Then, the optimal prior that minimizes  $\mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi|S)} \text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P})$  is given by  $\mathbf{P}^* = \prod_{m=1}^n \mathbb{E}_{q(S_m|\mathbf{e},\phi)} q(J_m | \mathbf{e}, \phi, S_m)$ . Moreover, under this prior, we obtain  $\mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi|S)} \text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P}^*) = \sum_m I(J_m; S_m | \mathbf{e}, \phi)$ .*

This connection provides insight into the choice of prior distributions in practical implementations—for instance, encouraging the use of mixture priors similar to the VampPrior [66] (see Appendix D.2 for further discussion). We also note that the assumption in Lemma 2, namely permutation invariance of the posterior, is standard in the analysis of randomized algorithms [42], and is satisfied by commonly used training methods such as SGD and SGLD [78].

Next, we present the asymptotic behavior of the empirical KL divergence term as follows:

247 **Lemma 3.** *Suppose the assumptions in Lemma 1 hold. Then, even under the optimal prior  $\mathbf{P}^*$  given*  
 248 *in Lemma 2, we have  $\mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi|S)} \text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P}^*)/n = \mathcal{O}(1)$ .*

249 This result indicates that asymptotic convergence cannot be achieved, even when using the optimal  
 250 prior  $\mathbf{P}^*$ , which minimizes  $\text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P})$ , to regularize the complexity of the encoder. To see why,  
 251 suppose that  $\text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P}^*) \rightarrow 0$  as  $n \rightarrow \infty$ . By Lemma 2, this would imply that the mutual  
 252 information  $\sum_m I(J_m; S_m | \mathbf{e}, \phi)$  also vanishes asymptotically. In turn, this would mean that the  
 253 encoder’s representations become independent of the training data as  $n$  increases, contradicting  
 254 the fundamental requirement for learning. In the supervised learning context, it has similarly been  
 255 observed that empirical KL terms analogous to  $\text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P})/n$  do not necessarily converge, even  
 256 for models that generalize well [21, 56]. Our findings are consistent with these results.

257 **Remark 2.** *Even when the posterior of  $J$  is defined by Eq. (2), a comparable upper bound on the KL*  
 258 *regularization term can still be derived by analyzing the encoder’s complexity via metric entropy. For*  
 259 *further details, see Section 4.2 and Appendix E.4.*

## 260 4 Proposed IT analysis under the new permutation symmetric setting

261 The observations presented in the previous section motivate the derivation of a generalization error  
 262 bound that avoids explicit dependence on  $\text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P})$ . We conjecture that the appearance of this  
 263 term in Theorem 2 arises from a fundamental limitation of the supersample setting, which necessitates  
 264 the use of a data-independent prior  $\mathbf{P}$  (as defined in Eq. (6)) to satisfy the necessary conditions  
 265 (A) and (B) described in Section 3.1. To overcome this limitation, in this section, we introduce an  
 266 extension of the supersample framework—namely, a novel *permutation symmetric setting*. This new  
 267 setting enables the construction of a data-dependent prior that satisfies both conditions simultaneously,  
 268 thereby yielding a generalization error bound that achieves asymptotic convergence. All the proofs of  
 269 this section are provided in Appendix E.

### 270 4.1 Permutation symmetric setting

271 To simultaneously satisfy the two conditions in Section 3.1, we propose randomly shuffling all  $2n$   
 272 data points in  $\tilde{X}$  using a uniform distribution and taking their expectation as the data-dependent prior  
 273 distribution. By definition, this distribution is permutation-invariant, thereby satisfying conditions  
 274 (A) and (B), allowing us to obtain the improved bound.

275 Formally, let us denote a random permutation of  $[2n]$  as  $\mathbf{T} = \{T_1, \dots, T_{2n}\}$ , where each permutation  
 276 appears with uniform probability,  $P(\mathbf{T}) = 1/(2n)!$ . Given a supersample  $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_{2n}) \in$   
 277  $\mathcal{X}^{2n}$ , a set of  $2n$  RVs drawn i.i.d. from  $\mathcal{D}$ , we reorder the samples using  $\mathbf{T}$  expressed as  $\tilde{X}_{\mathbf{T}} =$   
 278  $(\tilde{X}_{T_1}, \dots, \tilde{X}_{T_{2n}})$ . The first  $n$  samples  $(\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n})$  are used for the test dataset and the remaining  
 279  $n$  samples  $(\tilde{X}_{T_{n+1}}, \dots, \tilde{X}_{T_{2n}})$  are used for the training dataset. We further express  $\mathbf{T} = \{\mathbf{T}_0, \mathbf{T}_1\}$ ,  
 280 and  $\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n})$  and  $\tilde{X}_{\mathbf{T}_1} = (\tilde{X}_{T_{n+1}}, \dots, \tilde{X}_{T_{2n}})$  represent the test and training datasets,  
 281 respectively.

282 Given  $\tilde{X}$  and  $\mathbf{T}$ , we define the posterior distributions over the LVs of the test and train-  
 283 ing data, respectively, as  $q(\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0}) := \prod_{m=1}^n q(\tilde{J}_m | \mathbf{e}, \phi, \tilde{X}_{T_m})$ ,  $q(\mathbf{J} | \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1}) :=$   
 284  $\prod_{m=1}^n q(J_m | \mathbf{e}, \phi, \tilde{X}_{T_{n+m}})$ . We then define the joint posterior distribution as  $\mathbf{Q}_{\tilde{\mathbf{J}}, \mathbf{T}} :=$   
 285  $q(\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0})q(\mathbf{J} | \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1})$ .

286 Finally, we define our new data-dependent prior as

$$\mathbf{Q}_{\tilde{\mathbf{J}}} := \mathbb{E}_{\mathbf{T}} \mathbf{Q}_{\tilde{\mathbf{J}}, \mathbf{T}} = \mathbb{E}_{\mathbf{T}} q(\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0})q(\mathbf{J} | \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1}). \quad (9)$$

287 We refer to these settings as **the permutation symmetric (supersample) setting**. The following is  
 288 our main result.

289 **Theorem 3.** *Under Assumptions 1 and the permutation symmetric setting, we have*

$$\text{gen}(n, \mathcal{D}) \leq 3\Delta \sqrt{\frac{\mathbb{E}_{\tilde{X}, \mathbf{T}} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_{\mathbf{T}_1})} \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}}, \mathbf{T}} \| \mathbf{Q}_{\tilde{\mathbf{J}}})}{n}} + \frac{\Delta}{\sqrt{n}}.$$

290 **Remark 3.** *Unlike the existing supersample setting, where  $\{U_m\}$ s are independent, the elements of*  
 291  *$\mathbf{T}$  are dependent, which makes the analysis more complicated.*

**Explanation of Theorem 3:** Similar to Theorem 2, this bound is *independent of the decoder*  $g_\theta$ . The key difference is that the empirical KL term,  $\text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P})$ , is eliminated owing to our new data-dependent prior distribution  $\mathbf{Q}_{\tilde{\mathbf{J}}}$ . The proposed permutation satisfies both conditions (A) and (B) in Section 3.1, eliminating the need for a data-independent prior  $\mathbf{P}$ .

Next, we analyze the KL term in the bound. Similar to Eq. (8), we have

$$\mathbb{E}_{\tilde{X}, \mathbf{T}} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_{\mathbf{T}_1})} \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}}, \mathbf{T}} \parallel \mathbf{Q}_{\tilde{\mathbf{J}}}) \leq I(\mathbf{e}, \phi; \mathbf{T} | \tilde{X}) + I(\tilde{\mathbf{J}}; \mathbf{T} | \mathbf{e}, \phi, \tilde{X}).$$

Since  $\tilde{X}_{\mathbf{T}_1}$  corresponds to the training dataset  $S$ ,  $I(\mathbf{e}, \phi; \mathbf{T} | \tilde{X}) \leq I(\mathbf{e}, \phi; S)$  holds. Then, we can show that our generalization bound becomes

$$\text{gen}(n, \mathcal{D}) \leq 3\Delta \sqrt{\frac{I(\mathbf{e}, \phi; S) + I(\tilde{\mathbf{J}}; \mathbf{T} | \mathbf{e}, \phi, \tilde{X})}{n}} + \frac{\Delta}{\sqrt{n}}. \quad (10)$$

Our bound consists of the complexity of LV ( $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$ ) and the overfitting caused by learning the encoder parameters ( $I(\mathbf{e}, \phi; S)$ ) similar to Theorem 2. This implies the two key factors identified in Theorem 2 of Kawaguchi et al. [38]: how much information the LV retains from the input data and how much information from the training dataset is used to train the encoder.

As discussed in Section 3.2, when using a sufficiently regularized deterministic encoder,  $f'_{\mathbf{e}, \phi} : \mathcal{X} \rightarrow [K]$ , the CMI term satisfies  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})/n = \mathcal{O}(\log n/n)$ ; see Appendix D.7 for details. The parameter overfitting term can be controlled by specifying the training algorithm, as discussed in Section 3.2. Under these conditions, the generalization bound decreases as  $n \rightarrow \infty$ , meaning that Theorem 3 successfully characterizes generalization.

**Comparison with Theorem 2:** Although Theorem 3 shares a similar structure with Theorem 2, it introduces a refined shuffling strategy with  $\mathbf{T}$ , which resolves the issues of the supersample settings as discussed in Section 3.2. This shuffling is based on the fact that the marginal distribution of the dataset, which is invariant under permutation, can be expressed by the LV model. This new symmetry allows defining a data-dependent prior that satisfies necessary conditions while preserving decoder independence. On the other hand, the shuffling in Theorem 2 is based on the supersample setting and suitable for supervised learning, where overfitting is measured by swapping test and training data points. Practically, however, Theorem 2 relies on an  $n$ -dimensional variable,  $U$  (with independent components), which facilitates CMI estimation and algorithm design. In contrast, Theorem 3 uses a  $2n$ -dimensional variable,  $\mathbf{T}$  (with dependent components), which is theoretically more preferable but more difficult to estimate the CMI.

## 4.2 Generalization bound based on metric entropy

When using a softmax distribution in Eq. (2) for  $J$ , we show that the generalization bound is governed by the *metric entropy* under the permutation symmetric setting. Consequently, it does not require specifying a learning algorithm, which is required to discuss the convergence of Theorem 3 and provides a *uniform convergence bound* that depends solely on the function class of the encoder.

Let  $\mathcal{F}$  be the encoder function class equipped with the metric  $\|\cdot\|_\infty$ . Given  $x^n := (x_1, \dots, x_n) \in \mathcal{X}^n$ , define the pseudo-metric  $d_n$  on  $\mathcal{F}$  as  $d_n(f, g) := \max_{i \in [n]} \|f(x_i) - g(x_i)\|_\infty$  for  $f, g \in \mathcal{F}$ . The  $\delta$ -covering number of  $\mathcal{F}$  with respect to  $d_n$  is denoted as  $\mathcal{N}(\delta, \mathcal{F}, x^n)$ , and we define  $\mathcal{N}(\delta, \mathcal{F}, n) := \sup_{x^n \in \mathcal{X}^n} \mathcal{N}(\delta, \mathcal{F}, x^n)$ .

**Theorem 4.** Assume that there exists a positive constant  $\Delta_z$  such that  $\sup_{z, z' \in \mathcal{Z}} \|z - z'\| < \Delta_z$ . Then, when using Eq. (2) and under the same setting as Theorem 3, for any  $\delta \in (0, 1]$ , we have

$$\text{gen}(n, \mathcal{D}) \leq 4\Delta \sqrt{2\beta n \delta \Delta_z} + 3\Delta \sqrt{\frac{2 \log \mathcal{N}(\delta, \mathcal{F}, 2n)}{n}} + \frac{\Delta}{\sqrt{n}}.$$

We note that the parameter overfitting term does not appear in the bound. Since the encoder is parameterized by  $\phi \in \mathbb{R}^{d_\phi}$ , the metric entropy is  $\mathcal{O}(d_\phi d_z \log(1/\delta))$  [71]. Setting  $\delta = \mathcal{O}(1/n)$  gives  $\text{gen}(n, \mathcal{D}) = \mathcal{O}(\sqrt{d_\phi d_z \log n/n})$ . This result suggests that regularizing the complexity of the encoder improves generalization, whereas the complexity of the decoder has limited influence on the generalization. See Appendix E.3 for the proof and further discussion.



## 5 IT analysis for data generation performance

Mbacke et al. [46] provided statistical guarantees for the generalization error and *data generation performance* of VAEs, albeit under the strong assumption of an *untrained* decoder. Building on their approach, we provide a theoretical guarantee for the data generation performance of VQ-VAEs from an IT analysis perspective when both the encoder and decoder are trained jointly.

We first briefly summarize the data generation process in VQ-VAEs. After training, new data is generated by sampling an index  $J$  from a prior distribution,  $\pi(J|\mathbf{e}, \phi)$ , often chosen as a uniform distribution [63], and using the decoder network  $g_\theta$  to reconstruct the corresponding latent representation  $e_J$  from the learned codebook  $\mathbf{e}$ . Thus, the prior imposed on the latent representation is defined as  $\pi(e = e_j|\mathbf{e}, \phi)$  for all  $j = 1, \dots, K$ , and the data distribution generated through this procedure can be expressed as  $\hat{\mu} := g_\theta \# \pi(e|\mathbf{e}, \phi)$ , where  $g_\theta \# \pi$  denotes the pushforward of the distribution  $\pi$  by the decoder network. See Appendix F for the formal definition.

The following is the result of our analysis on the data generation performance of VQ-VAEs.

**Theorem 5.** *Suppose that  $g_\theta$  is measurable for any  $\theta$ , and Assumption 1 holds. Then, for any data-independent prior  $\pi(J|\mathbf{e}, \phi)$  as defined in Eq. (6), we have  $\mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} W_2^2(\mathcal{D}, \hat{\mu}) \leq \frac{2\Delta}{\sqrt{n}} +$*

$$\mathbb{E}_{S \sim q(\mathbf{e}, \phi, \theta|S)} \mathbb{E} \left[ \frac{2}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, S_m)} l(S_m, g_\theta(e_{J_m})) + 4\Delta \sqrt{\frac{2}{n} \sum_{m=1}^n \text{KL}(q(J_m|\mathbf{e}, \phi, S_m) \parallel \pi(J|\mathbf{e}, \phi))} \right],$$

where  $W_2(\mathcal{D}, \hat{\mu})$  is the 2-Wasserstein distance between the data distribution  $\mathcal{D}$  and the generated-data distribution  $\hat{\mu}$ .

The complete proof can be seen in Appendix F. The results indicate that the quality of approximating  $\mathcal{D}$  by  $\hat{\mu}$  can be enhanced by minimizing the reconstruction loss and the KL regularization term on LVs, which aligns with common training strategies for VQ-VAEs. Furthermore, this bound holds for *any* prior that satisfies the conditions outlined in Theorem 2. Thus, designing a prior that reduces this bound could lead to improved data generation accuracy. One potential approach is to use the data-dependent prior defined in Eq. (9). However, since this prior was originally designed to yield a reasonable generalization error upper bound, its practicality for data generation tasks remains debatable. Although our analysis does not identify an optimal and practical prior design for minimizing this upper bound, we can expect that our findings will stimulate further discussions on prior designs that effectively improve the generalization performance and data generation capabilities of VQ-VAEs.

## 6 Conclusion and limitations

We conclude this paper by discussing the practical implications of our results and summarizing their limitations. For an additional comparison with existing work, see Appendix B.2. The key insights derived from Theorems 2, 3, 4, and 5 are that the generalization performance of VQ-VAEs is primarily governed by the complexity of the LVs induced by the encoder and the complexity of the encoder parameters  $(\mathbf{e}, \phi)$ , while the decoder complexity  $(\theta)$  plays a limited role. This suggests that it is crucial to employ regularization strategies for LVs, designed to effectively reduce the upper bounds derived in our analysis. The above results also substantiate the validity of many VQ-VAE training strategies that employ objective functions with KL regularization on LVs for improving generalization performance.

The limitation of our findings is that the upper bound presented in Theorem 3 is challenging to compute numerically, making it impractical as an evaluation metric for generalization performance at present (see Sections 3.2 and 4 for details). This difficulty stems from the CMI term in Eq. (10), where the  $\mathbf{T}$  is  $2n$ -dimensional dependent random variable. Consequently, the numerical evaluation methods commonly used in existing IT analyses cannot be directly applied. Developing an alternative bound that enabling numerical evaluation for VQ-VAEs is another essential avenue for future research. It should also be noted as a limitation that our analysis is currently justified only for VQ-VAEs, which are based on discrete LVs. Extending our analysis to models that handle continuous LVs is a crucial step toward a deeper understanding of the fundamental role of LVs.

## References

- [1] Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 2897–2905, 2018. doi: 10.1109/TPAMI.2017.2784440.
- [2] Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ITA.2018.8503149.
- [3] Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International conference on machine learning*, pp. 159–168. PMLR, 2018.
- [4] Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.
- [5] Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.
- [6] Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016.
- [7] Bartlett, P. L. and Maass, W. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pp. 1188–1192, 2003.
- [8] Bendavid, S., Cesabianchi, N., Haussler, D., and Long, P. Characterizations of learnability for classes of  $[0, \dots, n]$ -valued functions. *Journal of Computer and System Sciences*, 50(1): 74–86, 1995. ISSN 0022-0000. doi: <https://doi.org/10.1006/jcss.1995.1008>. URL <https://www.sciencedirect.com/science/article/pii/S0022000085710082>.
- [9] Blau, Y. and Michaeli, T. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.
- [10] Blum, A. and Langford, J. Pac-mdl bounds. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pp. 344–357. Springer, 2003.
- [11] Bolley, F. and Villani, C. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Sciences de Toulouse*, 14:331–352, 2005. URL <https://api.semanticscholar.org/CorpusID:18695658>.
- [12] Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347, 2021.
- [13] Chérif-Abdellatif, B.-E., Shi, Y., Doucet, A., and Guedj, B. On pac-bayesian reconstruction guarantees for vaes. In *International conference on artificial intelligence and statistics*, pp. 3066–3079. PMLR, 2022.
- [14] Clarke, B. S. and Barron, A. R. Jeffreys’ prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.
- [15] Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012.
- [16] Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 207–232. JMLR Workshop and Conference Proceedings, 2011.
- [17] Dong, Y., Gong, T., Chen, H., Yu, S., and Li, C. Rethinking information-theoretic generalization: Loss entropy induced pac bounds. In *The Twelfth International Conference on Learning Representations*, 2024.
- [18] Dubhashi, D. P. and Ranjan, D. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), 1996.

- [19] Epstein, B. and Meir, R. Generalization bounds for unsupervised and semi-supervised learning with autoencoders. *arXiv preprint arXiv:1902.01449*, 2019.
- [20] Futami, F. and Fujisawa, M. Time-independent information-theoretic generalization bounds for SGLD. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=Ks0RSFNxPO>.
- [21] Geiger, B. C. and Koch, T. On the information dimension of stochastic processes. *IEEE Transactions on Information Theory*, 65(10):6496–6518, 2019. doi: 10.1109/TIT.2019.2922186.
- [22] Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating information flow in deep neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2299–2308. PMLR, 09–15 Jun 2019.
- [23] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [24] Gottlieb, L.-A., Kontorovich, A., and Krauthgamer, R. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.
- [25] Gray, R. M. *Entropy and information theory*. Springer Science & Business Media, 2011.
- [26] Guermeur, Y. Lp-norm sauer–shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017. ISSN 0022-0000.
- [27] Guermeur, Y. Combinatorial and structural results for gamma-psi-dimensions. *arXiv preprint arXiv:1809.07310*, 2018.
- [28] Hafez-Kolahi, H., Kasaei, S., and Soleymani-Baghshah, M. Sample complexity of classification with compressed input. *Neurocomputing*, 415:286–294, 2020. ISSN 0925-2312.
- [29] Haghighifard, M., Rodríguez-Gálvez, B., Thobaben, R., Skoglund, M., Roy, D. M., and Dziugaite, G. K. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pp. 663–706. PMLR, 2023.
- [30] Harutyunyan, H., Raginsky, M., Steeg, G. V., and Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 24670–24682, 2021.
- [31] Haussler, D. and Oppner, M. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.
- [32] Hellström, F. and Durisi, G. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022.
- [33] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.
- [34] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.
- [35] Jang, E., Gu, S., and Poole, B. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=rkE3y85ee>.
- [36] Jin, Y. Upper bounds on the natarajan dimensions of some function classes. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1020–1025. IEEE, 2023.
- [37] Joag-Dev, K. and Proschan, F. Negative association of random variables with applications. *The Annals of Statistics*, pp. 286–295, 1983.

- [38] Kawaguchi, K., Deng, Z., Ji, X., and Huang, J. How does information bottleneck help deep learning? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16049–16096. PMLR, 23–29 Jul 2023.
- [39] Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [40] Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E*, 69:066138, 2004.
- [41] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4): 541–551, 1989.
- [42] Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. In *The Eighth International Conference on Learning Representations*, 2020.
- [43] Loftsgaarden, D. O. and Quesenberry, C. P. A Nonparametric Estimate of a Multivariate Density Function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.
- [44] Lyu, Y., Liu, X., Song, M., Wang, X., Peng, Y., Zeng, T., and Jing, L. Recognizable information bottleneck. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4028–4036, 2023.
- [45] Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=S1jE5L5gl>.
- [46] Mbacke, S. D., Clerc, F., and Germain, P. Statistical guarantees for variational autoencoders using pac-bayesian theory. *Advances in Neural Information Processing Systems*, 36, 2023.
- [47] McAllester, D. A. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.
- [48] Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pp. 605–638, 2018.
- [49] Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11015–11025, 2019.
- [50] Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pp. 3526–3545. PMLR, 2021.
- [51] Pensia, A., Jog, V., and Loh, P.-L. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550, 2018.
- [52] Rissanen, J. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 2006.
- [53] Ross, B. C. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2): 1–101, 02 2014.
- [54] Roy, A., Vaswani, A., Neelakantan, A., and Parmar, N. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*, 2018.
- [55] Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

- [56] Sefidgaran, M., Zaidi, A., and Krasnowski, P. Minimum description length and generalization guarantees for representation learning. *Advances in Neural Information Processing Systems*, 36, 2023.
- [57] Sefidgaran, M., Zaidi, A., and Krasnowski, P. Generalization guarantees for representation learning via data-dependent gaussian mixture priors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=fGdF8Bq1FV>.
- [58] Sefidgaran, M., Zaidi, A., and Krasnowski, P. Generalization guarantees for multi-view representation learning and application to regularization via gaussian product mixture prior. *arXiv preprint arXiv:2504.18455*, 2025.
- [59] Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, 2010. ISSN 0304-3975. Algorithmic Learning Theory (ALT 2008).
- [60] Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.
- [61] Sønderby, C. K., Poole, B., and Mnih, A. Continuous relaxation training of discrete latent variable image models. In *Beysian DeepLearning workshop, NIPS*, volume 201, 2017.
- [62] Steinke, T. and Zakynthinou, L. Reasoning About Generalization via Conditional Mutual Information. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 3437–3452, 2020.
- [63] Takida, Y., Shibuya, T., Liao, W., Lai, C.-H., Ohmura, J., Uesaka, T., Murata, N., Takahashi, S., Kumakura, T., and Mitsufuji, Y. SQ-VAE: Variational Bayes on discrete representation with self-annealed stochastic quantization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20987–21012. PMLR, 17–23 Jul 2022.
- [64] Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (itw)*, pp. 1–5. IEEE, 2015.
- [65] Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- [66] Tomczak, J. and Welling, M. Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pp. 1214–1223. PMLR, 2018.
- [67] Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.
- [68] Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [69] Vera, M., Piantanida, P., and Vega, L. R. The role of the information bottleneck in representation learning. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1580–1584, 2018. doi: 10.1109/ISIT.2018.8437679.
- [70] Vera, M., Rey Vega, L., and Piantanida, P. The role of mutual information in variational classifiers. *Machine Learning*, 112(9):3105–3150, 2023.
- [71] Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- [72] Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.



- 566 [73] Wang, H., Huang, Y., Gao, R., and Calmon, F. Analyzing the generalization capability of SGLD  
567 using properties of Gaussian channels. In *Advances in Neural Information Processing Systems*,  
568 volume 34, pp. 24222–24234, 2021.
- 569 [74] Wang, H., Gao, R., and Calmon, F. P. Generalization bounds for noisy iterative algorithms  
570 using properties of additive noise channels. *Journal of Machine Learning Research*, 24(26):  
571 1–43, 2023.
- 572 [75] Wang, Z. and Mao, Y. On the generalization of models trained with SGD: Information-theoretic  
573 bounds and implications. In *The Tenth International Conference on Learning Representations*,  
574 2022.
- 575 [76] Wang, Z. and Mao, Y. Tighter information-theoretic generalization bounds from supersamples.  
576 In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp.  
577 36111–36137, 2023.
- 578 [77] Wang, Z., Huang, S.-L., Kuruoglu, E. E., Sun, J., Chen, X., and Zheng, Y. PAC-bayes  
579 information bottleneck. In *International Conference on Learning Representations*, 2022.
- 580 [78] Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In  
581 *Proceedings of the 28th International Conference on International Conference on Machine*  
582 *Learning*, pp. 681–688, 2011.
- 583 [79] Williams, W., Ringer, S., Ash, T., MacLeod, D., Dougherty, J., and Hughes, J. Hierarchical  
584 quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535,  
585 2020.
- 586 [80] Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning  
587 algorithms. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2524–2533,  
588 2017.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims in the abstract and Section 1 match our theoretical and numerical claims in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Assumptions for each theorem are explicitly shown. Following each theorem, there is a discussion about the theorem's limitations and implications. Additionally, Section 6 includes a discussion on the limitations of the entire paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#) .

Justification: Complete proofs of all theorems are provided in Appendices [C](#) and [D](#). For the reader’s convenience, the exact location of each proof is explicitly indicated alongside the corresponding theorem in the main text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#) .

Justification: The experimental setup for reproducing our results is detailed in Appendix [G](#). We submitted our source codes through OpenReview.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] .

Justification: We only used the popular benchmark datasets (such as MNIST and CIFAR-10) that can be easily obtained. The experimental setup for reproducing our results is detailed in Appendix G. We submitted our source codes through OpenReview.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes] .

Justification: The experimental setup for reproducing our results is detailed in Appendix G. We submitted our source codes through OpenReview.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes] .

Justification: We reported the mean  $\pm$  std. of the generalization gap and our bound values for all experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We used NVIDIA GPUs with 32GB memory (NVIDIA DGX-1 with Tesla V100 and DGX-2) in our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes] .

Justification: We confirmed that our paper does not have issues concerning the NeurIPS Code of Ethics, although the primary emphasis of this paper is on theoretical analysis.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer:[Yes] .

Justification: Although the primary focus of this paper is theoretical analysis, discussions on the potential impacts of our research are presented in Sections 1 and 6.

Guidelines:



- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: The primary focus of this paper is theoretical analysis, and although it includes experiments, their purpose is to numerically validate the theory. Therefore, the concerns raised in the question do not apply.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes] .

Justification: We provide citations or reference URLs for all of the code, data, and models used in our experiments (see Appendix G). We also declared the name of the licence is CC-BY 4.0 in our submission page of OpenReview.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The primary focus of this paper is theoretical analysis, and although it includes experiments, their purpose is to numerically validate the theory. Therefore, the concerns raised in the question do not apply.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: We do not utilize such services, so the concerns raised in the question are not applicable to us.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: The primary focus of this paper is theoretical analysis, and it has been confirmed that the concerns raised in the question are not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: This paper does not rely on LLMs for any theoretical analysis.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## 928 A Notation used in the main paper

929 We summarize the notation we used in the main part of our paper.

Category	Symbol	Meaning
Data and model	$n \in \mathbb{N}$	The sample size
	$\mathcal{X}, \mathcal{Z}$	A data and latent space
	$\mathcal{D}$	An unknown data generating distribution
	$\Delta \in \mathbb{R}^+$	A radius of a data space
	$X \in \mathcal{X} \subset \mathbb{R}^d$	A data
	$S = \{S_i\}_{i=1}^n \in \mathcal{X}^n$	A training dataset
	$\mathbf{e} = \{e_j\}_{j=1}^K \in \mathcal{Z}^K$	A codebook, where $K$ is the size of a codebook
	$\phi \in \Phi \subset \mathbb{R}^{d_\phi}$	An encoder parameter
	$\theta \in \Theta \subset \mathbb{R}^{d_\theta}$	A decoder parameter
	$W = \{\mathbf{e}, \phi, \theta\}$	A set of model parameters
	$f_\phi : \mathcal{X} \rightarrow \mathcal{Z}$	An encoder network
	$g_\theta : \mathcal{Z} \rightarrow \mathcal{X}$	A decoder network
	$q(J \mathbf{e}, \phi, X)$	A posterior distribution over $J$ given $\mathbf{e}, \phi, X$
	$\beta \in \mathbb{R}^+$	A temperature parameter used in a softmax
	$\mathcal{N}(\delta, \mathcal{F}, n)$	A $\delta$ -covering number with $n$ input for the encoder function class $\mathcal{F}$
Algorithm and loss functions	$\mathcal{A} : \mathcal{X}^n \rightarrow \mathcal{W}$	A randomized algorithm
	$q(\mathbf{e}, \phi, \theta S)$	A randomized algorithm given $S$
	$l : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$	a reconstruction loss function
	$l_0 : \mathcal{W} \times \mathcal{X} \rightarrow \mathbb{R}$	An expected loss function over $J$
	$\text{gen}(\mu, \mathcal{D})$	The expected generalization error based on a reconstruction loss
	$W_2(\mathcal{D}, \hat{\mu})$	The 2-Wasserstein distance between $\mathcal{D}$ and $\hat{\mu}$
Supersample setting	$\tilde{X} \in \mathcal{X}^{2n}$	A supersample used in the IT analysis
	$\tilde{X}_m$	The $m$ -th row of $\tilde{X}$
	$U = (U_1, \dots, U_n) \sim \text{Uniform}(\{0, 1\}^n)$	Random index used in the IT analysis
	$\tilde{X}_U := (\tilde{X}_{m, U_m})_{m=1}^n$	A training dataset in the supersample setting
	$\tilde{X}_{\bar{U}} := (\tilde{X}_{m, \bar{U}_m})_{m=1}^n$	A test dataset in the supersample setting, where $\bar{U}_m = 1 - U_m$
	$q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_U) := \prod_{m=1}^n q(J_m \mathbf{e}, \phi, \tilde{X}_{m, U_m})$	A joint distribution over index on the training dataset
	$q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\bar{U}}) := \prod_{m=1}^n q(J_m \mathbf{e}, \phi, \tilde{X}_{m, \bar{U}_m})$	A joint distribution over index on the test dataset
	$\mathbf{Q}_{\mathbf{J}, U} := q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_U)q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\bar{U}})$	A joint posterior distribution over $J$
	$\mathbf{Q}_{\mathbf{J}} := \mathbb{E}_U q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_U)q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\bar{U}})$	A data-dependent prior distribution over $J$
	$\pi(J \mathbf{e}, \phi)$	A data-independent prior distribution over $J$
Permutation symmetric setting	$\mathbf{T} = \{T_1, \dots, T_{2n}\} \sim P(\mathbf{T}) = 1/(2n)!$	A random permutation following a uniform distribution
	$\tilde{X}_{\mathbf{T}} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_{2n}})$	Randomly permuted supersamples
	$\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n})$	The test dataset
	$\tilde{X}_{\mathbf{T}_1} = (\tilde{X}_{T_{n+1}}, \dots, \tilde{X}_{T_{2n}})$	The training dataset
	$q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0}) = \prod_{m=1}^n q(J_m \mathbf{e}, \phi, \tilde{X}_{T_m})$	A joint distribution over index on the test dataset
	$q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1}) = \prod_{m=1}^n q(J_m \mathbf{e}, \phi, \tilde{X}_{T_{n+m}})$	A joint distribution over index on the training dataset
	$\mathbf{Q}_{\mathbf{J}, \mathbf{T}} = q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0})q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1})$	A joint posterior distribution over $J$
	$\mathbf{Q}_{\mathbf{J}} = \mathbb{E}_{\mathbf{T}} q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0})q(\mathbf{J} \mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1})$	A data-dependent prior distribution over $J$

## 930 B Additional discussion and related work

931 Here, we provide additional discussion and a comparison between our study and existing work.

### 932 B.1 Related work

933 Here we briefly introduce additional related existing work, especially about the IT analysis. In IT  
934 analysis [80], the generalization error is evaluated on the basis of the MI between learned parameters  
935 and training data. This approach is closely related to the PAC-Bayes theory and has been extended  
936 through supersample settings [62] to exploit the symmetry between test and training data. This setting  
937 has been applied to the study of generalization based on outputs of functions [30], losses [32, 76],  
938 and hypothesis entropy [17]. The relationship between IT analysis and the IB hypothesis has been  
939 discussed from numerical and algorithmic perspectives [77, 44]. More recently, Sefidgaran et al. [56]  
940 theoretically studied latent variable models using IT analysis, demonstrating that generalization can  
941 be characterized by the complexity of the encoder and latent variables without relying on decoder  
942 information. They also developed a theoretical link among IT analysis, the IB hypothesis, and MDL  
943 by using compression bounds [10].

## 944 B.2 Comparison with existing bounds

945 Here, we compare our bounds with those in existing work. Theorem 2 resembles the results of  
 946 Mbacke et al. [46] since both bounds include the empirical KL term in the upper bounds, and the  
 947 posterior distribution corresponds to the variational posterior distribution. The key difference is that  
 948 Mbacke et al. [46] assumed a fixed decoder, whereas our analysis incorporates the learning process  
 949 under the assumption of a discrete latent space and a squared reconstruction loss. Another distinction  
 950 is that their generalization bound does not become 0 as  $n \rightarrow \infty$  due to two reasons. One is the  
 951 presence of the empirical KL term, which we address in Theorem 3 using permutation symmetry. Our  
 952 technique can be regarded as developing the appropriate prior distribution in PAC-Bayes bound. The  
 953 second reason is the presence of the average distance  $\frac{1}{n} \sum_{m=1}^n \mathbb{E}_X \|X - S_m\|$  in the existing bound,  
 954 which is inherent to the data distribution and may not vanish as  $n \rightarrow \infty$ . Our use of the squared loss  
 955 in the analysis mitigates this problematic term, as detailed in Appendix D.1.

956 Our proof techniques are based on Sefidgaran et al. [56]. However, we could not directly apply  
 957 their methods, as the reconstruction loss reuses input data, unlike in classification settings. We  
 958 resolve this by combining the data regeneration technique used in the proof of Mbacke et al. [46].  
 959 Additionally, we introduced a new permutation symmetric setting, leading to a bound that controls  
 960 mutual information in Theorem 3. Our setting is closely related to the type-2 symmetry proposed  
 961 in Sefidgaran et al. [56], which involves random permutations selecting  $n$  indices from  $2n$  with  
 962 a uniform distribution  $1/\binom{2n}{n}$ , whereas our setting requires the consideration of the order of the  
 963 permutation index to evaluate the exponential moment (see Appendix E.1). Finally, we theoretically  
 964 studied the behavior of the CMI (Theorem 4) focusing on the complexity of the encoder, whereas  
 965 Sefidgaran et al. [56] provided the bounds based on the CMI without such discussion.

966 The existing analyses based on the IB hypothesis [69, 28, 38, 70] assumed that both the latent  
 967 variables and data are discrete, and their obtained bounds explicitly depend on the latent space size  
 968 or show exponential dependence on the MI. In contrast, we assume that only latent variables are  
 969 discrete and the resulting bound does not explicitly depend on the number of discrete states nor  
 970 exhibit exponential dependence on MI. Furthermore, our bound shows the dependency on  $d_z$  not  $K$ ,  
 971 which is the significant difference compared with existing bounds.

## 972 B.3 Discussion and comparison of our prior and posterior and existing work

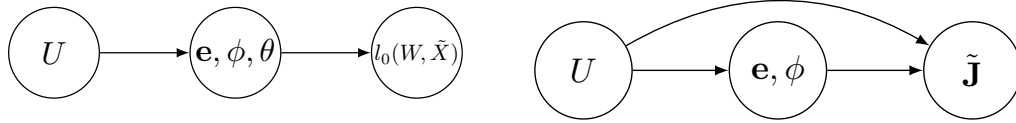


Figure 3: Graphical models illustrating the different dependency structures of the random variables considered in the basic IT analysis and in this study. The left figure represents the dependency structure in the basic IT analysis, which simply evaluates the loss function in supervised learning settings, whereas the right figure corresponds to our analysis in the unsupervised learning setting.

973 Here, we explain how the prior distribution is used in our proof and why two prior distributions  
 974 are introduced in our bound. First, the IT analysis with supersample reformulates generalization  
 975 analysis as the problem of estimating which samples were used for training when data is randomly  
 976 shuffled based on  $U$ . If this estimation is difficult, our model generalizes well. In the basic IT analysis  
 977 (Theorem 1), such difficulty is measured by the CMI between  $U$  and the loss function  $l_0$ .

978 Of course, such shuffling is not performed in actual algorithms; it is introduced only for theoretical  
 979 analysis using the Donsker-Varadhan inequality [25], where such shuffling is defined by the prior and  
 980 posterior distributions dependent on  $U$ .

981 In basic IT analyses (Theorem 1) for supervised learning, by shuffling with  $U$ , we observe how the  
 982 loss  $l_0(W, \tilde{X})$  changes. Here, the goal is to estimate  $U$  from the observed losses. As depicted in  
 983 Figure 3,  $U$  and  $l_0(W, \tilde{X})$  depend on all parameters, including the decoder, resulting in a bound that  
 984 depends on all parameters.



Our goal is to eliminate the dependency between the decoder and latent variables (LVs). To achieve this, we introduce a prior and posterior that establish the dependency as depicted in Figure 3. The key idea is that by introducing a new dependency between  $U$  and LVs, we can directly shuffle  $U$ , leading to a bound that isolates the role of LVs without involving the decoder. For additional discussions on the necessary conditions for the prior, see Appendix D.2.

Finally, we show the additional explanation of Figure 1. The figure illustrates the difference between the existing fCMI and our new CMI. The left figure illustrates the setting of existing fCMI where  $\tilde{J}$  follows the distribution in the setting of Eq. (5), see Appnxdix C.2 for the detail. Thus, in the existing fCMI,  $\tilde{J}$  and  $U$  are conditionally independent given  $\mathbf{e}$  and  $\phi$  and  $\tilde{X}$ . On the other hand, the right figure is our setting and there is an edge between  $U$  and  $\tilde{J}$  directly, and thus  $\tilde{J}$  and  $U$  are conditionally independent given  $\mathbf{e}$  and  $\phi$  and  $\tilde{X}$ , which results in the difference of existing fCMI and our CMI. See Appendix D.5 and Appendix D.3 for the additional discussion about the fCMI.

## C Proofs for Section 2 and additional discussion

### C.1 Proof of Theorem 1

This is just the consequence of the existing eCMI bound [32]. We can confirm this as follows;

Note that the generalization error can be expressed as the supsample

$$\begin{aligned} \text{gen}(n, \mathcal{D}) &= \left| \mathbb{E}_{S, X} \mathbb{E}_{q(\mathbf{e}, \phi, \theta | S)} \left( \mathbb{E}_{q(J | \mathbf{e}, \phi, X)} l(X, g_\theta(e_J)) - \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, S_m)} l(S_m, g_\theta(e_{J_m})) \right) \right| \\ &= \left| \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} \left( \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\tilde{J}_m | \mathbf{e}, \phi, X_m, \tilde{U}_m)} l(X_m, \tilde{U}_m, g_\theta(e_{J_m})) \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m)} l(X_m, U_m, g_\theta(e_{J_m})) \right) \right|. \end{aligned}$$

Given that the loss is bounded by  $[0, \Delta]$ , the integrated is a  $\Delta$ -sub-Gaussian random variable. Thus, from Hellström & Durisi [32], the generalization error bound that satisfies the  $\sigma^2$  sub Gaussianity is bounded as  $\sqrt{\frac{2\sigma^2}{n} I(l(\mathcal{A}(\tilde{X}_U), \tilde{X}); U | \tilde{X})}$ , we obtain the result. Finally  $I(l_0(W, \tilde{X}); U | \tilde{X}) \leq I(W; U | \tilde{X})$  holds by the data processing inequality.

### C.2 Proof for Eq. (5) and additional discussion

Here we prove Eq. (5). It is important to note that this upper bound is characterized by the CMI  $I(l_0(W, \tilde{X}); U | \tilde{X})$ . This CMI depends on the decoder and encoder information, distinguishing it from the results presented in our main Theorems 2 and 3, which do not require the decoder's information.

To clarify this distinction, let us introduce the necessary notation. Following the notation in Section 3.1, we define  $\tilde{Y} = g_\theta(e_{\tilde{J}})$ , where  $g_\theta(e_{\tilde{J}})$  implies applying  $g_\theta(\cdot)$  elementwise to  $e_{\tilde{J}}$ . Under these notations, we have the following relations:

$$I(l_0(W, \tilde{X}); U | \tilde{X}) \leq I(\tilde{Y}; U | \tilde{X}) \leq I(\theta; U | \tilde{X}) + I(e_{\tilde{J}}; U | \tilde{X}, \theta),$$

where the first inequality is obtained by the data processing inequality (DPI) and the second inequality is obtained by the chain rule of CMI and the DPI. This result demonstrates that the decoder information cannot be eliminated from the basic IT bound, which clarifies the fundamental difference compared to our result (Theorems 2 and 3). Moreover, since the decoder and encoder are learned simultaneously using the same training data, they are not independent. This makes it unclear how the latent variables and the encoder's capacity affect generalization, as it is difficult to eliminate the decoder's dependency on them.

### 1020 C.3 Additional discussion when $K = 1$

1021 Another limitation of the basic IT-bound arises when considering  $K = 1$  as a limiting setting. From  
 1022 the definition of the squared loss, the generalization error is given by:

$$\text{gen}(n, \mathcal{D}) \leq \sqrt{\text{Var}[X] \frac{\mathbb{E}\|g_\theta(e)\|^2}{n}} \leq \frac{\Delta}{\sqrt{n}}. \quad (11)$$

1023 The proof of this is described below. This upper bound is intuitive: for  $K = 1$ , the model effectively  
 1024 ignores the input data and embeds all samples into the same latent variable, which can be interpreted  
 1025 as a form of strong regularization. Consequently, the impact of overfitting due to training the decoder  
 1026 network is relatively limited, and the generalization error can be seen, in a sense, as being comparable  
 1027 to the inherent variability of the data itself.

1028 The above observations motivate us to develop a more sophisticated generalization bound that  
 1029 explicitly captures the role of representation.

1030 *Proof of Eq. (11).* Since  $K = 1$ , we express  $\mathbf{e} = \{e\}$ . By using the definition of the squared loss, we  
 1031 have

$$\text{gen}(n, \mathcal{D}) = \left| \mathbb{E}_S \mathbb{E}_{q(e, \phi, \theta|S)} \left( \mathbb{E}[X] - \frac{1}{n} \sum_{m=1}^n S_m \right) \cdot g_\theta(e) \right|,$$

1032 where we used the fact that the generated data always use  $e$  as a latent variable since  $\mathbf{e} = \{e\}$  when  
 1033  $K = 1$ . Then by using the Cauchy-Schwartz inequality, we have

$$\text{gen}(n, \mathcal{D}) \leq \sqrt{\text{Var}[X] \frac{\mathbb{E}\|g_\theta(e)\|^2}{n}} \leq \frac{\Delta}{\sqrt{n}},$$

1034 where we used the fact that the diameter of the instance space is bounded by  $\Delta$ . □

## 1035 D Proofs for Section 3

1036 In the proofs, we repeatedly use the following type of exponential moment inequality, which is often  
 1037 used in the proof of McDiarmid's inequality. A function  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  has the bounded differences  
 1038 property if for some nonnegative constants  $c_1, \dots, c_n$ , the following holds for all  $i$ :

$$\sup_{x_1, \dots, x_n, x'_i \in \mathcal{X}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

1039 Assuming  $X_1, \dots, X_n$  are independent random variables taking values in  $\mathcal{X}$ , we have the following  
 1040 lemma:

1041 **Lemma 4** (Used in the proof of McDiarmid's inequality). *Given a function  $f$  with the bounded*  
 1042 *differences property, for any  $t \in \mathbb{R}$ , we have:*

$$\mathbb{E} \left[ e^{t(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)])} \right] \leq e^{\frac{t^2}{8} \sum_{i=1}^n c_i^2}.$$

### 1043 D.1 Proof of Theorem 2

1044 We express  $q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}) = q(\tilde{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\tilde{U}}, \tilde{X}_U) = q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\tilde{U}})q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U)$ . Hereinafter, we  
 1045 simplify the notation by expressing  $\tilde{X}$  as  $X$ . For simplification in the proof, we omit the absolute  
 1046 operation for the generalization gap. The reverse bound can be proven in a similar manner. We first

1047 express the generalization error of the reconstruction loss using the supersample as follows

$$\begin{aligned}
& \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m | \mathbf{e}, \phi, X_m, \bar{U}_m) q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, \bar{U}_m, g_\theta(e_k)) \mathbb{1}_{k=\bar{J}_m} \\
& \quad - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m) q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, U_m, g_\theta(e_k)) \mathbb{1}_{k=J_m} \\
& = \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m | \mathbf{e}, \phi, X_m, \bar{U}_m) q(\mathbf{e}, \phi, \theta | X_U)} \|X_m, \bar{U}_m - g_\theta(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m} \\
& \quad - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m) q(\mathbf{e}, \phi, \theta | X_U)} \|X_m, U_m - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m}, \quad (12)
\end{aligned}$$

1048 where the first term corresponds to the test loss and the second term corresponds to the training loss.

1049 Recall the learning algorithm and posterior distribution:

$$\begin{aligned}
\mathbf{e}, \phi, \theta & \sim q(\mathbf{e}, \phi, \theta | X_U), \\
J_m & \sim q(\mathbf{J} | \mathbf{e}, \phi, S_m).
\end{aligned}$$

1050 Here  $\mathbf{e} = \{e_1, \dots, e_K\}$  is the codebook, and  $J$  and  $\mathbf{J} = \{J_1, \dots, j_n\}$  represents the index chosen  
1051 from the codebook.

1052 Conditioned on  $X$  and  $U$ , we then decompose Eq. (12) as follows

$$\begin{aligned}
& \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, \bar{U}_m, g_\theta(e_k)) \mathbb{E}_{q(\bar{J}_m | \mathbf{e}, \phi, X_m, \bar{U}_m)} \mathbb{1}_{k=\bar{J}_m} \\
& \quad - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, \bar{U}_m, g_\theta(e_k)) \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m)} \mathbb{1}_{k=J_m} \\
& \quad + \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, \bar{U}_m, g_\theta(e_k)) \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m)} \mathbb{1}_{k=J_m} \\
& \quad - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, U_m, g_\theta(e_k)) \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m)} \mathbb{1}_{k=J_m}. \quad (13)
\end{aligned}$$

1053 We will separately upper bound these terms.

#### 1054 D.1.1 Bounding first and second terms

1055 The decomposition of the generalization error, as shown in Eq. (13), allows us to bound the first and  
1056 second terms as follows.

1057 We apply Donsker-Varadhan's inequality between the following two distributions:

$$\begin{aligned}
\mathbf{Q} & := P(U) q(\mathbf{e}, \phi, \theta | X_U) q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U) \\
\mathbf{P}_S & := P(U) q(\mathbf{e}, \phi, \theta | X_U) \mathbb{E}_{P(U')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'}), \quad (14)
\end{aligned}$$

1058 These correspond to the posterior and data-dependent prior distributions defined in Section 3.1.

1059 Then, for any  $\lambda \in \mathbb{R}^+$ , we have

$$\begin{aligned}
& \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_m, \bar{U}_m, g_\theta(e_k)) \left( \mathbb{E}_{q(\bar{J}_m | \mathbf{e}, \phi, X_m, \bar{U}_m)} \mathbb{1}_{k=\bar{J}_m} - \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_m, U_m)} \mathbb{1}_{k=J_m} \right) \\
& \leq \frac{1}{\lambda} \text{KL}(\mathbf{Q} | \mathbf{P}_S) + \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}_S} \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_m, \bar{U}_m, g_\theta(e_k)) (\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m}) \right).
\end{aligned}$$

1060 To simplify the notation, we express  $\bar{\mathbf{J}} = \mathbf{J}_0$ ,  $\bar{J}_m = J_{m,0}$ ,  $\mathbf{J} = \mathbf{J}_1$ , and  $J_m = J_{m,1}$ . Let  $U''$  be a  
 1061 random variable taking 0, 1 with a uniform distribution. Since  $\mathbf{P}_S$  is symmetric with respect to the  
 1062 permutation of  $\mathbf{J}_0$  and  $\mathbf{J}_1$ , we can bound the exponential moment as:

$$\begin{aligned}
 & \log \mathbb{E}_{P(U)q(\mathbf{e}, \phi, \theta | X_U)} \mathbb{E}_{P(U')} q(\mathbf{J}_0, \mathbf{J}_1 | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'}) \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{m, \bar{U}_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m,0}} - \mathbb{1}_{k=J_{m,1}}) \right) \\
 &= \log \mathbb{E}_{P(U)q(\mathbf{e}, \phi, \theta | X_U) P(U'')^n} \mathbb{E}_{P(U')} q(\mathbf{J}_0, \mathbf{J}_1 | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'}) P(U'')^N \\
 & \quad \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{m, \bar{U}_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m, \bar{U}''}} - \mathbb{1}_{k=J_{m, U''}}) \right) \\
 &= \log \mathbb{E}_{P(U)q(\mathbf{e}, \phi, \theta | X_U)} \mathbb{E}_{P(U')} q(\mathbf{J}_0, \mathbf{J}_1 | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'}) \mathbb{E}_{P(U'')^n} \\
 & \quad \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{m, \bar{U}_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m, U''}} - \mathbb{1}_{k=J_{m, U''}}) \right).
 \end{aligned}$$

1063 In the final line, we apply McDiarmid's inequality since  $U''^n$  are  $n$  i.i.d. random variables. To use  
 1064 McDiarmid's inequality in Lemma 4, we use the stability caused by replacing one of the elements of  $n$   
 1065 i.i.d. random variables. To estimate the coefficients of stability in Lemma 4, let  $U''^n = (U''_1, \dots, U''_N)$ ,  
 1066 then

$$\begin{aligned}
 & \sup_{\{U''_m\}_{m=1}^n, U''_{m'}} \left| \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{m, \bar{U}_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m, \bar{U}''}} - \mathbb{1}_{k=J_{m, U''}}) \right. \\
 & \quad - \frac{\lambda}{n} \sum_{k=1}^K \sum_{m \neq m'} l(X_{m, \bar{U}_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m, \bar{U}''}} - \mathbb{1}_{k=J_{m, U''}}) \\
 & \quad \left. - \frac{\lambda}{n} \sum_{k=1}^K l(X_{m', \bar{U}'_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m', \bar{U}''}} - \mathbb{1}_{k=J_{m', U''}}) \right| \\
 &= \sup_{\{U''_m\}_{m=1}^n, U''_{m'}} \left| \frac{\lambda}{n} \sum_{k=1}^K l(X_{m', \bar{U}'_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m', \bar{U}''}} - \mathbb{1}_{k=J_{m', U''}}) \right. \\
 & \quad \left. - \frac{\lambda}{n} \sum_{k=1}^K l(X_{m', \bar{U}'_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m', \bar{U}''}} - \mathbb{1}_{k=J_{m', U''}}) \right| \leq \frac{2\lambda\Delta}{n}.
 \end{aligned} \tag{15}$$

1067 Here, the maximum change caused by replacing one element of  $U''$  is  $2\lambda\Delta/n$ , thus, its log of the  
 1068 exponential moment is bounded by  $(2\lambda\Delta/n)^2/8 \times n = \lambda^2\Delta^2/2n$ . Thus from Lemma 4, we have

$$\begin{aligned}
 & \log \mathbb{E}_{P(U)q(\mathbf{e}, \phi, \theta | X_U)} \mathbb{E}_{P(U')} q(\mathbf{J}_0, \mathbf{J}_1 | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'}) \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{m, \bar{U}_m}, g_\theta(e_k)) (\mathbb{1}_{k=J_{m,0}} - \mathbb{1}_{k=J_{m,1}}) \right) \\
 & \leq \frac{\lambda^2\Delta^2}{2n}.
 \end{aligned}$$

1069 The first and second terms in Eq. (13) are upper bounded by

$$\frac{1}{\lambda} \mathbb{E}_X \text{KL}(\mathbf{Q} | \mathbf{P}_S) + \frac{\lambda\Delta^2}{2n}. \tag{16}$$

### 1070 D.1.2 Bounding third and fourth terms

1071 Next, we upper bound the third and fourth terms in Eq. (13);

$$\begin{aligned}
 & \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_{m, \bar{U}_m}, g_\theta(e_k)) \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{m, U_m})} \mathbb{1}_{k=J_m} \\
 & - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} l(X_{m, U_m}, g_\theta(e_k)) \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{m, U_m})} \mathbb{1}_{k=J_m}.
 \end{aligned} \tag{17}$$

1072 We simplify the notation by expressing  $\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})} \mathbb{1}_{k=J_m}$  as  $P_{k,m}$  and use the square loss:

$$\begin{aligned}
& \mathbb{E}_{X,U} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)} l(X_{m,\bar{U}_m}, g_\theta(e_k)) P_{k,m} - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)} l(X_{m,U_m}, g_\theta(e_k)) P_{k,m} \\
&= \mathbb{E}_{X,U} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)} (\|X_{m,\bar{U}_m}\|^2 - \|X_{m,U_m}\|^2) P_{k,m} \\
&+ \mathbb{E}_{X,U} \sum_{k=1}^K \frac{2}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)} (X_{m,\bar{U}_m} - X_{m,U_m}) \cdot g_\theta(e_k) P_{k,m} \\
&= \mathbb{E}_{X,U} \frac{1}{n} \sum_{m=1}^n (\|X_{m,\bar{U}_m}\|^2 - \|X_{m,U_m}\|^2) \mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)} \sum_{k=1}^K P_{k,m} \\
&+ \mathbb{E}_S \frac{2}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \sum_{k=1}^K g_\theta(e_k) P_{k,m} \\
&= \mathbb{E}_S \frac{2}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \sum_{k=1}^K g_\theta(e_k) P_{k,m}, \tag{18}
\end{aligned}$$

1073 where we express  $S = (X_{1,U_1}, \dots, X_{n,U_n}) = (S_1, \dots, S_n)$  as the training samples. In the last  
1074 inequality, we used  $\sum_{k=1}^K P_{k,m} = 1$  and  $\mathbb{E}_{X,U} \frac{1}{n} \sum_{m=1}^n (\|X_{m,\bar{U}_m}\|^2 - \|X_{m,U_m}\|^2) = 0$  since  $X$   
1075 and  $U$  are i.i.d.

1076 To evaluate the final line, we use the Donsker-Valadhan inequality between

$$\begin{aligned}
\mathbf{Q} &:= q(\mathbf{e}, \phi, \theta|S) \prod_{m=1}^n q(J_m|\mathbf{e}, \phi, S_m), \\
\mathbf{P}_S &:= q(\mathbf{e}, \phi, \theta|S) \prod_{m=1}^n \pi(J_m|\mathbf{e}, \phi),
\end{aligned}$$

1077 where  $\pi(J_m|\mathbf{e}, \phi)$  is the prior distribution, which never depends on the training data.

1078 Then we have

$$\begin{aligned}
& \mathbb{E}_S \frac{2}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \sum_{k=1}^K g_\theta(e_k) P_{k,m} \\
&\leq \mathbb{E}_S \frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}_S) + \mathbb{E}_S \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}_S} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (\mathbb{E}_X S - X_m) \cdot \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m} \right) \\
&\leq \mathbb{E}_S \frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}_S) \\
&+ \mathbb{E}_S \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}_S} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \sum_{k=1}^K g_\theta(e_k) (\mathbb{1}_{k=J_m} - P''_{k,m}) \right) \\
&+ \mathbb{E}_S \mathbb{E}_{\mathbf{P}_S} \frac{2}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \sum_{k=1}^K g_\theta(e_k) P''_{k,m}, \tag{19}
\end{aligned}$$



1079 where  $P''_{k,m} = \mathbb{E}_{q(J_m|\phi, \mathbf{e})} \mathbb{1}_{k=J_m}$ . Clearly, this does not depend on the index  $m$ , so we express  
 1080  $P''_{k,m} = P''_k$ . Then the last term becomes

$$\begin{aligned}
 \mathbb{E}_S \mathbb{E}_{\mathbf{P}_S} \frac{1}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \sum_{k=1}^K g_\theta(e_k) P''_k &\leq \mathbb{E}_S \mathbb{E}_{\mathbf{P}_S} \left\| \mathbb{E}_X X - \frac{1}{n} \sum_{m=1}^n S_m \right\| \left\| \sum_{k=1}^K g_\theta(e_k) P''_k \right\| \\
 &\leq \mathbb{E}_S \left\| \mathbb{E}_X X - \frac{1}{n} \sum_{m=1}^n S_m \right\| \sqrt{\Delta} \\
 &\leq \sqrt{\Delta \text{Var} \left( \frac{1}{n} \sum_{m=1}^n S_m \right)} \\
 &\leq \sqrt{\Delta \frac{\text{Var}(X)}{n}} \\
 &\leq \sqrt{\frac{\Delta}{4n}} \sqrt{\Delta} = \frac{\Delta}{2\sqrt{n}}, \tag{20}
 \end{aligned}$$

1081 where we used the fact that the variance of random variables with bounded in  $(a, b]$  is upper bounded  
 1082 by  $(b - a)^2/4n$  (the extension to the  $d$ -dimensional random variable is straightforward) and thus,  
 1083  $\text{Var}(X) \leq \Delta/4$ . Then the exponential moment term becomes

$$\begin{aligned}
 &\mathbb{E}_S \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}_S} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \sum_{k=1}^K g_\theta(e_k) (\mathbb{1}_{k=J_m} - P''_{k,m}) \right) \\
 &= \mathbb{E}_S \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}_S} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (\mathbb{E}_X X - S_m) \cdot \sum_{k=1}^K g_\theta(e_k) (\mathbb{1}_{k=J} - P''_k) \right).
 \end{aligned}$$

1084 Here we use the McDiarmid's inequality for  $n$  random variables  $\mathbf{J}$ . Then we estimate the stability  
 1085 coefficient similarly to Eq. (15), which is upper bounded by  $\lambda\Delta/n$ . Then from Lemma 4, the  
 1086 exponential moment is bounded by  $(2\lambda\Delta/n)^2/8 \times n = \lambda\Delta^2/2n$ . Thus, the second term is upper  
 1087 bounded by

$$\frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}_S) + \frac{\lambda\Delta^2}{2n} + \frac{\Delta}{\sqrt{n}}. \tag{21}$$

1088 By optimizing the first and second terms of Eqs. (16) and (21), we have

$$2\Delta \sqrt{\frac{(\mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi, \theta|X_U)} \text{KL}(\mathbf{Q}_1||\mathbf{Q}_2) + \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \text{KL}(\mathbf{Q}|\mathbf{P}_S))}{n}} + \frac{\Delta}{\sqrt{n}},$$

1089 where

$$\begin{aligned}
 \mathbf{Q}_1 &:= q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, X_{\bar{U}}, X_U) \\
 \mathbf{Q}_2 &:= \mathbb{E}_{P(U')} q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, X_{\bar{U}'}, X_{U'}), \\
 \mathbf{Q} &:= \prod_{m=1}^n q(J_m|\mathbf{e}, \phi, S_m), \\
 \mathbf{P}_S &:= \prod_{m=1}^n \pi(J_m|\mathbf{e}, \phi).
 \end{aligned}$$

## 1090 D.2 Necessarily conditions for the prior and the limitation of the existing supsample setting

1091 Here, we further discuss the necessary conditions for the prior distribution to derive a meaningful  
 1092 generalization bound. The proof strategy in Appendix D.1 clarifies this point: in the proof, we  
 1093 decompose the generalization bound in Eq. (13) and separately upper bound the first two terms and  
 1094 the latter two terms.

For the first and second terms, the analysis follows standard generalization error techniques. When using a prior and posterior distribution characterized by the shuffling of the supersample  $\tilde{X}$ , such as the index variable  $U$ , the shuffling must swap test and training data to enable generalization evaluation. By ensuring this swap, we can properly assess overfitting.

For the third and fourth terms, after applying the Donsker-Valadhan lemma, it is crucial to ensure that the probability  $P''_{k,m}$  does not depend on the sample index  $m$  to control the exponential moment in Eq. (19). This requires satisfying  $P''_{k,m} = P''_k$ , meaning that the probability of assigning the  $m$ -th data point to the  $k$ -th codebook must be independent of  $m$ . By definition, this condition holds when the distribution of the latent variables remains invariant after shuffling.

From these observations, we conclude that the prior used for shuffling must: (A) **Preserve the distribution of the LVs to eliminate interdependencies between LVs and the decoder**, and (B) **Swap test and training data points to evaluate overfitting**, as discussed in Section 3.1.

Using the supersample ensures condition (B). For condition (A), we employ the prior distribution  $\pi(J_m|\mathbf{e}, \phi)$ , which removes sample index dependency and guarantees  $P''_{k,m} = P''_k$ . Consequently, the empirical KL divergence in Theorem 2 arises from the third and fourth terms in Eq. (13), as detailed in Appendix D.1.2.

Based on these findings, we propose the following type of prior distribution:

$$\mathbf{P}_S := q(\mathbf{e}, \phi, \theta|S) \prod_{m=1}^n \sum_{m'=1}^n \frac{1}{N} q(J_m|\mathbf{e}, \phi, S_{m'}),$$

which provides an empirical approximation of the marginal distribution using available samples. Since this distribution does not explicitly depend on the sample index, we can bound the exponential moment similarly to the approach in Appendix D.1.2.

However, using the prior distribution in Eq. (14) to bound the third and fourth terms of Eq. (13) is not feasible. The issue is that applying the Donsker-Valadhan lemma with Eq. (14) to these terms does not yield a bound of order  $\mathcal{O}(1/\sqrt{n})$ , as achieved in Eq. (20). This limitation arises because the dependency on the sample index in Eq. (14) prevents us from leveraging the symmetry between the test and training datasets via the supersample index  $U$ . As a result, the prior distribution's symmetry cannot be exploited to simplify the bounds for these terms.

### D.3 Comparison with the fCMI

Here, we analyze the relationship between our CMI and existing forms of fCMI in more detail. As highlighted in the main paper, a key distinction is that our CMI is conditioned on all model parameters, whereas existing fCMI methods marginalize over these parameters.

To further explore this difference, we consider marginalizing over the encoder parameter,  $\phi$ . In the proof of Theorem 2, we perform this marginalization over  $\phi$  in Eq. (12) and obtain

$$\begin{aligned} & \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m|\mathbf{e}, \phi, X_m, \bar{U}_m)} q(\mathbf{e}, \phi, \theta|X_U) l(X_m, \bar{U}_m, g_\theta(e_k)) \mathbb{1}_{k=\bar{J}_m} \\ & - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_m, U_m)} q(\mathbf{e}, \phi, \theta|X_U) l(X_m, U_m, g_\theta(e_k)) \mathbb{1}_{k=J_m} \\ & = \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m|\theta, \mathbf{e}, X_m, \bar{U}_m)} q(\mathbf{e}, \theta|X_U) \|X_m, \bar{U}_m - g_\theta(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m} \\ & - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\theta, \mathbf{e}, X_m, U_m)} q(\mathbf{e}, \theta|X_U) \|X_m, U_m - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m}, \end{aligned}$$

and proceed with the proof in the same way. We apply the Donsker-Varadhan inequality between the following distributions, instead of Eq. (14):

$$\begin{aligned} \mathbf{Q} &:= P(U)P(U')q(\mathbf{e}, \theta|X_U)q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \theta, X_{\bar{U}}, X_U) \\ \mathbf{P} &:= P(U)q(\mathbf{e}, \theta|X_U)\mathbb{E}_{P(U')}q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \theta, X_{\bar{U}'}, X_{U'}). \end{aligned}$$

1129 This incorporates marginalization over  $\phi$  in Eq. (14), resulting in the following KL divergence in the  
 1130 upper bound:

$$\begin{aligned}\mathbb{E}_X \text{KL}(\mathbf{Q}|\mathbf{P}) &= \mathbb{E}_X \mathbb{E}_{P(U)q(\mathbf{e}, \phi|X_U)} \text{KL}(q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \theta, X_{\bar{U}}, X_U) | \mathbb{E}_{P(U')} q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \theta, X_{\bar{U}'}, X_{U'})) \\ &= I(\bar{\mathbf{J}}, \mathbf{J}; U|\mathbf{e}, \theta, X).\end{aligned}$$

1131 Unlike Theorem 2, this CMI explicitly involves the decoder parameter  $\theta$ . By marginalizing over  $\phi$ ,  
 1132 decoder information is integrated into the upper bound, making Theorem 2 distinct from existing  
 1133 fCMI bounds. In Appendix D.5, further discussion from the viewpoint of the difference of the  
 1134 graphical model between our CMI and existing fCMI is given.

#### 1135 D.4 Proof of Lemma 2

1136 We remark that the following relationship holds for  $m = 1 \dots, n$  by definition;

$$\begin{aligned}I(J_m; S_m|\mathbf{e}, \phi) &= \mathbb{E}_{q(\mathbf{e}, \phi)} \mathbb{E}_{q(S_m|\mathbf{e}, \phi)} \mathbb{E}_{q(J_m|\mathbf{e}, \phi, S_m)} \log \frac{q(J_m|\mathbf{e}, \phi, S_m)}{\mathbb{E}_{q(S_m|\mathbf{e}, \phi)} q(J_m|\mathbf{e}, \phi, S_m)} \\ &= \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi|S)} \mathbb{E}_{q(J_m|\mathbf{e}, \phi, S_m)} \log \frac{q(J_m|\mathbf{e}, \phi, S_m)}{\mathbb{E}_{q(S_m|\mathbf{e}, \phi)} q(J_m|\mathbf{e}, \phi, S_m)}.\end{aligned}\quad (22)$$

1137 Next, we show  $\mathbb{E}_{q(S_1|\mathbf{e}, \phi)} q(J_1|\mathbf{e}, \phi, S_1) = \dots = \mathbb{E}_{q(S_n|\mathbf{e}, \phi)} q(J_n|\mathbf{e}, \phi, S_n)$  holds under the given  
 1138 assumption. To prove this, it is suffice to show that  $q(S_1|\mathbf{e}, \phi) = \dots = q(S_n|\mathbf{e}, \phi)$  holds. Under the  
 1139 given assumption

$$q(\mathbf{e}, \phi|S_1) = \dots = q(\mathbf{e}, \phi|S_n)$$

1140 holds, see Li et al. [42] for the proof. Then for  $i \in [n]$ , we have

$$q(\mathbf{e}, \phi|S_i) p(S_i) = q(S_i|\mathbf{e}, \phi) p(\mathbf{e}, \phi)$$

1141 and since all training data points are drawn i.i.d form  $\mathcal{D}$ , we have

$$q(\mathbf{e}, \phi|S_i) \mathcal{D} = q(S_i|\mathbf{e}, \phi) p(\mathbf{e}, \phi).$$

1142 Then, for any  $j \neq i \in [n]$ , we also have

$$q(\mathbf{e}, \phi|S_j) \mathcal{D} = q(S_j|\mathbf{e}, \phi) p(\mathbf{e}, \phi)$$

1143 since  $q(\mathbf{e}, \phi|S_j) = q(\mathbf{e}, \phi|S_i)$ , we conclude that  $q(S_i|\mathbf{e}, \phi) = q(S_j|\mathbf{e}, \phi)$ . This implies  
 1144  $\mathbb{E}_{q(S_1|\mathbf{e}, \phi)} q(J_1|\mathbf{e}, \phi, S_1) = \dots = \mathbb{E}_{q(S_n|\mathbf{e}, \phi)} q(J_n|\mathbf{e}, \phi, S_n)$  holds under the given assumption. So we  
 1145 use the joint distribution these as  $\mathbf{P} = \prod_{m=1}^n \mathbb{E}_{q(S_m|\mathbf{e}, \phi)} q(J_m|\mathbf{e}, \phi, S_m)$ . From Eq. (22), we have

$$\mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi|S)} \text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P}) = I(J_m; S_m|\mathbf{e}, \phi).$$

1146 Finally, we show that above  $\mathbf{P}$  minimizes the  $\mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi|S)} \text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P})$ . We consider the prior  
 1147  $\mathbf{P}'$  that satisfies the assumption of the Theorem 7, that is, prepare some distributions that satisfies  
 1148  $q(J_1|\mathbf{e}, \phi) = \dots = q(J_n|\mathbf{e}, \phi)$  and define  $\mathbf{P}' := \prod_{m=1}^n \pi(J_m|\mathbf{e}, \phi)$

1149 By the definition, we have that

$$\mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi|S)} \text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P}') = I(J_m; S_m|\mathbf{e}, \phi) + \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi|S)} \text{KL}(\mathbf{P} \| \mathbf{P}').$$

1150 Thus, when using  $\mathbf{P}' = \mathbf{P}$  minimizes the empirical KL divergence.

1151 **D.5 Proof of Eq. (8)**

1152 Here we discuss how we can upper bound of the complexity term of the obtained bound. From the  
 1153 definition, we have the following relation;

$$\begin{aligned}
 & \mathbb{E}_{X,U} \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}}, U} \| \mathbf{Q}_{\tilde{\mathbf{J}}}) \\
 &= \mathbb{E}_{P(X)P(U)q(\mathbf{e}, \phi, \theta | X_U)q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U)} \log \frac{q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U)}{\mathbb{E}_{P(U')} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})} \\
 &= \mathbb{E}_{P(X)P(U)q(\mathbf{e}, \phi | X_U)q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U)} \log \frac{q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U)}{\mathbb{E}_{P(U')} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})} \\
 &= \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X, U)} \log \frac{q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U)}{\mathbb{E}_{P(U')} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})} \\
 &= \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X, U)} \log \frac{q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}}, X_U)}{\mathbb{E}_{P(U' | \mathbf{e}, \phi, X)} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})} \\
 &\quad + \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X, U)} \log \frac{\mathbb{E}_{P(U' | \mathbf{e}, \phi, X)} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})}{\mathbb{E}_{P(U')} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})} \\
 &= I(\tilde{\mathbf{J}}, \mathbf{J}; U | \mathbf{e}, \phi, X) + \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X, U)} \log \frac{\mathbb{E}_{P(U' | \mathbf{e}, \phi, X)} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})}{\mathbb{E}_{P(U')} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\bar{U}'}, X_{U'})} \\
 &\leq I(\tilde{\mathbf{J}}, \mathbf{J}; U | \mathbf{e}, \phi, X) + \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)} \log \frac{P(U' | \mathbf{e}, \phi, X)}{P(U')} \\
 &= I(\tilde{\mathbf{J}}, \mathbf{J}; U | \mathbf{e}, \phi, X) + \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)} \log \frac{P(\mathbf{e}, \phi | X, U')P(U' | X)}{\mathbb{E}_{P(U'' | X)} P(\mathbf{e}, \phi | X, U'')P(U')} \\
 &= I(\tilde{\mathbf{J}}, \mathbf{J}; U | \mathbf{e}, \phi, X) + \mathbb{E}_{P(X)P(\mathbf{e}, \phi | X)} \mathbb{E}_{P(U | \mathbf{e}, \phi, X)} \log \frac{P(\mathbf{e}, \phi | X, U')P(U')}{\mathbb{E}_{P(U'')} P(\mathbf{e}, \phi | X, U'')P(U')} \\
 &= I(\tilde{\mathbf{J}}, \mathbf{J}; U | \mathbf{e}, \phi, X) + I(\mathbf{e}, \phi; U | X),
 \end{aligned}$$

1154 where we used the data processing inequality of the KL divergence.

1155 **D.6 The role of  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$**

1156 The role of  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$  is clarified through the following upper bound:

$$\begin{aligned}
 I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X}) &\leq \sum_{m=1}^n I(e_J; \tilde{X}_m, \bar{U}_m | \mathbf{e}, \phi) \\
 &\quad + \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi | S)} \text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P}). \tag{23}
 \end{aligned}$$

1157 The first term represents the information retained by the LVs from the training data in the IB  
 1158 hypothesis, while the second term corresponds to the regularization based on the empirical KL  
 1159 divergence discussed earlier.

1160 Here we prove Eq. (23). We define  $\pi(\tilde{\mathbf{J}} | \mathbf{e}, \phi) = \prod_{m=1}^n \pi(\tilde{J}_m | \mathbf{e}, \phi)$ ,  $\pi(\mathbf{J} | \mathbf{e}, \phi) = \prod_{m=1}^n \pi(J_m | \mathbf{e}, \phi)$ ,  
 1161 and  $\pi(\tilde{\mathbf{J}} | \mathbf{e}, \phi) = \pi(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi) = \pi(\tilde{\mathbf{J}} | \mathbf{e}, \phi) \pi(\mathbf{J} | \mathbf{e}, \phi)$  where each  $\pi(\tilde{J}_m | \mathbf{e}, \phi)$  is the marginal distribu-  
 1162 tion of  $\pi(J_m | \mathbf{e}, \phi, X_m)$ .

1163 Then by the definition of the CMI, we have

$$\begin{aligned}
& I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X}) \\
&= \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_U)} \text{KL}(q(\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}) \| \mathbb{E}_{U'} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, \tilde{X}_{U'}, \tilde{X}_{U'})) \\
&\leq \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_U)} \text{KL}(q(\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}) \| \pi(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi)) \\
&= \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_U)} \text{KL}(q(\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}_U) \| \pi(\tilde{\mathbf{J}} | \mathbf{e}, \phi)) + \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_U)} \text{KL}(q(\mathbf{J} | \mathbf{e}, \phi, \tilde{X}_U) \| \pi(\mathbf{J} | \mathbf{e}, \phi)) \\
&= \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_U)} \sum_{m=1}^n \text{KL}(q(\tilde{J}_m | \mathbf{e}, \phi, \tilde{X}_{m, \tilde{U}_m}) \| \pi(\tilde{J}_m | \mathbf{e}, \phi)) \\
&\quad + \mathbb{E}_{\tilde{X}, U} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_U)} \sum_{m=1}^n \text{KL}(q(J_m | \mathbf{e}, \phi, \tilde{X}_{m, U_m}) \| \pi(J_m | \mathbf{e}, \phi)) \\
&= nI(J; X | \mathbf{e}, \phi) + \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi | S)} \frac{1}{n} \sum_{m=1}^n \text{KL}(q(J_m | \mathbf{e}, \phi, S_m) \| \pi(J_m | \mathbf{e}, \phi)) \\
&\leq nI(e_J; X | \mathbf{e}, \phi) + \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi | S)} \frac{1}{n} \sum_{m=1}^n \text{KL}(q(J_m | \mathbf{e}, \phi, S_m) \| \pi(J_m | \mathbf{e}, \phi)).
\end{aligned}$$

## 1164 D.7 Proof of Lemma 1 and 3 and additional discussion

1165 *Proof of Lemma 1.* From the definition of the CMI, we have

$$I(\tilde{\mathbf{J}}, \mathbf{J}; U | \mathbf{e}, \phi, X) = H[\tilde{\mathbf{J}} | \mathbf{e}, \phi, X] - H[\tilde{\mathbf{J}} | U, \mathbf{e}, \phi, X] \leq H[\tilde{\mathbf{J}} | \mathbf{e}, \phi, X] \leq H[\tilde{\mathbf{J}} | X].$$

1166 Here, we consider the case where  $f_\phi : \mathcal{X} \rightarrow [K]$  represents a deterministic encoder that maps  
 1167 input data to one of the  $K$  indices. This scenario can be viewed as a  $K$ -class classification problem,  
 1168 allowing us to directly apply the results from Harutyunyan et al. [30]. They demonstrated that the  
 1169 CMI for multi-class classification problems can be upper-bounded using the Natarajan dimension, a  
 1170 combinatorial measure that generalizes the VC dimension to the multiclass setting.

1171 Using this concept, we obtain the following characterization:

1172 When employing a deterministic encoder network  $f'_\phi : \mathcal{X} \rightarrow [K]$  that belongs to a class with finite  
 1173 Natarajan dimension  $d_K$  and assuming  $2n > d_K + 1$ , we derive the following bound:

$$I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X}) \leq d_K \log \left( \binom{K}{2} \frac{2en}{d_K} \right). \quad (24)$$

1174 The proof follows exactly as in Theorem 8 of Harutyunyan et al. [30].  $\square$

1175 Thus, by regularizing the capacity of the encoder model (via the Natarajan dimension), the CMI term  
 1176 scales as  $\mathcal{O}(\log n)$ , ensuring controlled generalization behavior. Examples of models that satisfy the  
 1177 finite Natarajan dimension are shown in Jin [36] and Daniely et al. [16]. Also, see Bendavid et al. [8],  
 1178 which shows that the VC dimension of the multiclass loss function characterizes the graph dimension,  
 1179 and the graph dimension upper bounds the Natarajan dimension.

1180 *Proof of Lemma 3.* Since we consider the setting of Lemma 2, we consider the case of  
 1181  $\text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P}) = \sum_m I(J_m; S_m | \mathbf{e}, \phi)$ . Following the above setting of  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$ , that is,  
 1182  $f'_{\mathbf{e}, \phi} : \mathcal{X} \rightarrow [K]$  satisfies the Natarajan dimension  $d_K > 1$ . Then for each  $m$ , we have

$$I(J_m; S_m | \mathbf{e}, \phi) = H[J_m | \mathbf{e}, \phi] - H[J_m | S_m, \mathbf{e}, \phi] = H[J_m | \mathbf{e}, \phi] \leq \log K \leq (d_K + 1) \log K.$$

1183 Thus  $\text{KL}(\mathbf{Q}_{\mathbf{J}, U} \| \mathbf{P})/n = \sum_m I(J_m; S_m | \mathbf{e}, \phi)/n \leq \log K = \mathcal{O}(1)$ .  $\square$

1184 The difference between  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$  and  $I(J_m; S_m | \mathbf{e}, \phi)$  lies in their conditioning. Since  
 1185  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$  is conditioned on all  $2n$  data points, it only depends on the combinatorial num-  
 1186 ber of distinct index values. In contrast,  $I(J_m; S_m | \mathbf{e}, \phi)$  does not condition on the input data, making  
 1187 regularization based solely on the Natarajan dimension insufficient to control complexity.

1188 For the discussion of the stochastic encoder, see Appendix E.4, where we consider the metric entropy  
 1189 of  $f_\phi(\cdot)$ , which leads to a similar discussion.



## 1190 D.7.1 Additional discussion for the Natarajan dimension

1191 Here, we briefly discuss the Natarajan dimension. First, it can be both upper and lower bounded by  
 1192 the graph dimension, another common combinatorial measure for multi-class classification problems  
 1193 (see Lemma 4 and Proposition 1 in Guermeur [27]).

1194 The Natarajan dimension can also be upper bounded by the  $\gamma$  fat-shattering dimension of each class.  
 1195 Specifically, given  $f'e, \phi : \mathcal{X} \rightarrow [K]$ , let the  $k$ -th element of its output be denoted as  $f'e, \phi'^{(k)}$  for  
 1196  $k = 1, \dots, K$ . If each  $f'e, \phi'^{(k)}$  has a finite  $\gamma$ -shattering dimension, then the Natarajan dimension of  
 1197  $f'e, \phi$  can be bounded by the sum of the  $\gamma$ -shattering dimensions of its components, multiplied by a  
 1198 constant coefficient (see Lemma 10 in Guermeur [27]).

1199 Examples of fat-shattering dimension evaluations can be found in Bartlett & Maass [7], which  
 1200 analyzes neural network models, and Gottlieb et al. [24], which examines the fat-shattering dimension  
 1201 of Lipschitz function classes. If our encoder network satisfies these properties, its covering number  
 1202 can be appropriately bounded.

## 1203 D.8 Discussion about the overfitting term

Here, we discuss how the overfitting terms relate to different algorithms. First, from the data  
 processing inequality [15], we obtain

$$I(\mathbf{e}, \phi; U|\tilde{X}) \leq I(\mathbf{e}, \phi; S),$$

1204 where we express  $\tilde{X}_U$  as the training dataset  $S$ . Since this expression does not include conditioning,  
 1205 we refer to it as the parameter MI. Several existing studies have analyzed parameter MI under  
 1206 commonly used algorithms.

1207 Pensia et al. [51] first established the relationship between noisy iterative algorithms and parameter  
 1208 MI. Subsequently, Wang et al. [73] and Wang et al. [74] investigated the parameter MI of the SGLD  
 1209 algorithm from the perspective of noisy iterative algorithms, while Futami & Fujisawa [20] analyzed  
 1210 it in the continuous-time limit. Neu et al. [50] was the first to examine parameter MI in SGD, with  
 1211 Wang & Mao [75] later improving its dependency on the step size. Furthermore, Haghifam et al. [29]  
 1212 provided formal limitations in the context of stochastic convex optimization.

In addition to these, in the Bayesian setting, where we assume that the training dataset is conditionally  
 i.i.d (see Clarke & Barron [14] for the formal settings), Clarke & Barron [14] (see also Rissanen  
 [52], Haussler & Opper [31]) clarified that the mutual information between learned parameter and  
 training dataset is described as follows: if  $w$  takes a value in a  $d$ -dimensional compact subset of  $\mathbb{R}^d$   
 and  $p(y|x; w)$  is smooth in  $w$ , then as  $n \rightarrow \infty$ , we have

$$I(W; S) = \frac{d}{2} \log \frac{n}{2\pi e} + h(W) + \mathbb{E} \log \det J + o(1),$$

1213 where  $h(W)$  is the differential entropy of  $W$ , and  $J$  is the Fisher information matrix of  $p(Y|X; W)$ .

1214 Steinke & Zakynthinou [62] clarified that the CMI is upper bounded by the the stability. For example,  
 1215 if the training algorithm satisfies  $\sqrt{2\epsilon}$ -differentially private (DP) algorithm, then CMI is upper-  
 1216 bounded by  $\epsilon n$ . So this  $\epsilon$  is controlled by the DP algorithm. The Gibbs algorithm equipped with  
 1217  $[0, 1]$  bounded loss function, satisfies  $\mathcal{O}(1/n)$ -DP, thus its CMI is controlled adequately. Steinke &  
 1218 Zakynthinou [62] also clarified that if the algorithm is  $\delta$  stable in total variation distance, then CMI is  
 1219 upper bounded by  $\delta n$ . Li et al. [42] studied the total variation stability for the SGD, and Mou et al.  
 1220 [48] studied such stability of the SGLD algorithm and its relation to the PAC-Bayesian bound. [49]  
 1221 investigated the CMI of SGLD as the noisy iterative algorithm.

## 1222 E Proofs for Section 4

### 1223 E.1 Proof of Theorem 3

1224 We define  $\mathbf{T} = \{\mathbf{T}_0, \mathbf{T}_1\}$ , where  $\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n})$  serves as the test dataset and  $\tilde{X}_{\mathbf{T}_1} =$   
 1225  $(\tilde{X}_{T_{n+1}}, \dots, \tilde{X}_{T_{2n}})$  serves as the training dataset. We further express  $\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n}) =$   
 1226  $(\tilde{X}_{T_{0,1}}, \dots, \tilde{X}_{T_{0,n}})$  and  $\tilde{X}_{\mathbf{T}_1} = (\tilde{X}_{T_{1,1}}, \dots, \tilde{X}_{T_{1,n}})$ . To emphasize the dependence of the

dataset on  $\mathbf{T}$ , we write the posterior distribution as  $q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}}) = q(\tilde{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}}) =$   
 $q(\tilde{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0}, \tilde{X}_{\mathbf{T}_1}) = q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0})q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1}).$   
 Hereinafter, we express  $\tilde{X}$  as  $X$  to simplify the notation. Under the permutation symmetric settings,  
 the generalization error can be expressed as

$$\begin{aligned} & \mathbb{E}_{S, X} \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \left( \mathbb{E}_{q(J|\mathbf{e}, \phi, X)} l(X, g_{\theta}(e_J)) - \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, S_m)} l(S_m, g_{\theta}(e_{J_m})) \right) \\ &= \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m|\mathbf{e}, \phi, X_{\mathbf{T}_0, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} l((X_{\mathbf{T}_0, m}, g_{\theta}(e_k))) \mathbb{1}_{k=\bar{J}_m} \\ &- \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_{\mathbf{T}_1, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} l(X_{\mathbf{T}_1, m}, g_{\theta}(e_k)) \mathbb{1}_{k=J_m} \\ &= \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m|\mathbf{e}, \phi, X_{\mathbf{T}_0, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_0, m} - g_{\theta}(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m} \\ &- \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_{\mathbf{T}_1, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_1, m} - g_{\theta}(e_k)\|^2 \mathbb{1}_{k=J_m}. \end{aligned}$$

We then decompose the loss as follows

$$\begin{aligned} & \text{gen}(n, \mathcal{D}) \\ &= \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\bar{J}_m|\mathbf{e}, \phi, X_{\mathbf{T}_0, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_0, m} - g_{\theta}(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m} \\ &- \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_{\mathbf{T}_1, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_0, m} - g_{\theta}(e_k)\|^2 \mathbb{1}_{k=J_m} \\ &+ \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_{\mathbf{T}_1, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_0, m} - g_{\theta}(e_k)\|^2 \mathbb{1}_{k=J_m} \\ &- \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_{\mathbf{T}_1, m})q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_1, m} - g_{\theta}(e_k)\|^2 \mathbb{1}_{k=J_m}. \end{aligned} \tag{25}$$

First, we upper bound the first two terms by applying the Donsker-Varadhan inequality. Consider the  
 joint distribution and the prior distribution, defined as follows:

$$\begin{aligned} \mathbf{Q} &:= P(\mathbf{T})q(\mathbf{e}, \theta, \phi|X_{\mathbf{T}_1})q(\tilde{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, X_{\mathbf{T}}), \\ \mathbf{P} &:= P(\mathbf{T})q(\mathbf{e}, \theta, \phi|X_{\mathbf{T}_1}) \mathbb{E}_{P(\mathbf{T}')} q(\tilde{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, X_{\mathbf{T}'}). \end{aligned} \tag{26}$$

This corresponds to the posterior and data-dependent prior distributions defined in Section 4.1.

Then we then obtain

$$\begin{aligned} & \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_0, m} - g_{\theta}(e_k)\|^2 \left( \mathbb{E}_{q(\bar{J}_m|\mathbf{e}, \phi, X_{\mathbf{T}_1, m})} \mathbb{1}_{k=\bar{J}_m} - \mathbb{E}_{q(J_m|\mathbf{e}, \phi, X_{\mathbf{T}_0, m})} \mathbb{1}_{k=J_m} \right) \\ &\leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n \|X_{\mathbf{T}_0, m} - g_{\theta}(e_k)\|^2 (\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m}) \right). \end{aligned} \tag{27}$$

1236 Note that  $\mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'})$  is symmetric with respect to the permutation of  $\mathbf{T}$ . Thus, we have

$$\begin{aligned}
& \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m}) \right) \\
&= \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) P(\mathbf{T}'') \\
&\quad \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}) \right) \\
&= \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \\
&\quad \mathbb{E}_{P(\mathbf{T}'')} \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}) \right).
\end{aligned}$$

1237 To simplify the notation, we define  $\mathbf{T}'' = \{\mathbf{T}''_0, \mathbf{T}''_1\} = \{\mathbf{T}''_{0,1}, \dots, \mathbf{T}''_{0,n}, \mathbf{T}''_{1,1}, \dots, \mathbf{T}''_{1,n}\}$ . Note  
1238 that  $\mathbf{T}''_{j,m}$  for  $m = 1, \dots, n$  and  $j = 0, 1$  are not independent of each other due to the permutation  
1239 that generates them. Therefore, we cannot directly apply standard concentration inequalities, as is  
1240 possible in the existing supersample setting.

1241 To address this, we use the results from Joag-Dev & Proschan [37], which concern the negative  
1242 association of permutation variables. From Theorem 2.11 in Joag-Dev & Proschan [37], the distri-  
1243 bution  $P(\mathbf{T})$  satisfies negative association. Additionally, as discussed in Section 3.3 of Joag-Dev &  
1244 Proschan [37] and further in Proposition 4 and 5 of Dubhashi & Ranjan [18], we have that

$$\begin{aligned}
& \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \\
& \quad \mathbb{E}_{P(\mathbf{T}'')} \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}) \right) \\
& \leq \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \\
& \quad \mathbb{E}_{\prod_{m=1}^n \prod_{j=0,1} P(\mathbf{T}''_{j,m})} \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}) \right),
\end{aligned}$$

1245 where  $P(\mathbf{T}''_{j,m})$  is the marginal distribution, implying that  $\mathbf{T}''_{j,m}$  are now  $2n$  independent random  
1246 variables. Intuitively, the results in Joag-Dev & Proschan [37] indicate that the elements of the  
1247 permutation index, which follow the permutation distribution, are negatively correlated. As a result,  
1248 the expectation of the marginal distribution is larger than that of the joint distribution.

1249 Since  $\{\mathbf{T}''_{j,m}\}$  are independent, we can apply McDiarmid's inequality, which leads to the results in

$$\begin{aligned}
& \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \\
& \quad \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m}) \right) \\
& \leq \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \\
& \quad \mathbb{E}_{\prod_{m=1}^n \prod_{j=0,1} P(\mathbf{T}''_{j,m})} \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) (\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}) \right) \\
& \leq \frac{\lambda^2 \Delta^2}{n}. \tag{28}
\end{aligned}$$

1250 This is derived similarly to Eq. (15). Note that there are  $2n$  variables so the calculation of the upper  
1251 bound is  $(\Delta\lambda/n)^2/8 \times 2n = \lambda^2 \Delta^2/4n$ .

1252 Next, we focus on the third and fourth terms in Eq. (25). Similarly to Eq. (18), we have

$$\begin{aligned}
& \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{\mathbf{T}_{1,m}}) q(\mathbf{e}, \phi, \theta | X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
& - \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{\mathbf{T}_{1,m}}) q(\mathbf{e}, \phi, \theta | X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
& = \mathbb{E}_{X, \mathbf{T}} \frac{2}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{\mathbf{T}_{1,m}}) q(\mathbf{e}, \phi, \theta | X_{\mathbf{T}_1})} \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m} \\
& \leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q} | \mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m} \right) \\
& \leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q} | \mathbf{P}) \\
& + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{P(\mathbf{T}) q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \mathbb{E} \prod_{m=1}^n \prod_{j=0,1} P(\mathbf{T}''_{j,m}) \\
& \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m} \right). \tag{29}
\end{aligned}$$

1253 We first evaluate the expectation of the exponential moment;

$$\Omega := \mathbb{E}_{P(\mathbf{T}) q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \frac{2}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m}. \tag{30}$$

1254 Let us now focus on the expectation  $\mathbb{E}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'})$ . Due to the permutation symmetry,

1255  $\mathbb{E}_{P(\mathbf{T}')} \mathbb{E}_{q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'})} \sum_{k=1}^K \mathbb{1}_{k=J_m}$  is the same for all  $m$ .

1256 For instance, when  $n = 2$ , the possible permutations of  $\mathbf{T}$  are  $\mathbf{T} =$   
 1257  $(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), \dots$ , resulting in 24 distinct patterns and thus

$$\begin{aligned}
P_{k,1} &= \mathbb{E}_{P(\mathbf{T}')} \mathbb{E}_{q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'})} \mathbb{1}_{k=\bar{J}_1} = \mathbb{E}_{\frac{1}{4} q(J_1 | \mathbf{e}, \phi, X_1) + \frac{1}{4} q(J_1 | \mathbf{e}, \phi, X_2) + \frac{1}{4} q(J_1 | \mathbf{e}, \phi, X_3) + \frac{1}{4} q(J_1 | \mathbf{e}, \phi, X_4)} \mathbb{1}_{k=J_1} \\
P_{k,2} &= \mathbb{E}_{P(\mathbf{T}')} \mathbb{E}_{q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'})} \mathbb{1}_{k=\bar{J}_2} = \mathbb{E}_{\frac{1}{4} q(J_2 | \mathbf{e}, \phi, X_1) + \frac{1}{4} q(J_2 | \mathbf{e}, \phi, X_2) + \frac{1}{4} q(J_2 | \mathbf{e}, \phi, X_3) + \frac{1}{4} q(J_2 | \mathbf{e}, \phi, X_4)} \mathbb{1}_{k=J_2} \\
&\vdots
\end{aligned}$$

Thus, all  $P_{k,m}$  does not depend on the index  $m$ . So we express  $\mathbb{E}_{P(\mathbf{T}')} q(\mathbf{J}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \sum_{k=1}^K \mathbb{1}_{k=J_m}$  as  $P_k$ . Then Eq. (30) can be written as

$$\begin{aligned}
& \mathbb{E}_X \mathbb{E}_{P(\mathbf{T})} q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \left( \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{0,m}} \right) \cdot \sum_{k=1}^K g_\theta(e_k) P_k \\
&= \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_X \left( \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{0,m}} \right) \cdot q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \sum_{k=1}^K g_\theta(e_k) P_k \\
&= \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} \mathbb{E}_{X_{\mathbf{T}_0}} \left( \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{0,m}} \right) \cdot q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \sum_{k=1}^K g_\theta(e_k) P_k \\
&= \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} \left( \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \mathbb{E}_{X_{\mathbf{T}_0}} \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{0,m}} \right) \cdot q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \sum_{k=1}^K g_\theta(e_k) P_k \\
&= \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} \left( \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \mathbb{E}_X X \right) \cdot q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \sum_{k=1}^K g_\theta(e_k) P_k \\
&\leq \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \left\| \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \mathbb{E}_X X \right\| \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \left\| \sum_{k=1}^K g_\theta(e_k) P_k \right\| \\
&\leq \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \left\| \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \mathbb{E}_X X \right\| \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \left\| \sum_{k=1}^K g_\theta(e_k) P_k \right\| \\
&\leq \mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} \left\| \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \mathbb{E}_X X \right\| \sqrt{\Delta}.
\end{aligned}$$

We bound the above exactly same ways as Eq. (20), that is, we can upper bound the above by the variance of bounded random variable and thus, we have

$$\mathbb{E}_{P(\mathbf{T})} \mathbb{E}_{X_{\mathbf{T}_1}} \left\| \frac{1}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \mathbb{E}_X X \right\| \leq \sqrt{\frac{\Delta}{4n}}.$$

Thus, we have

$$\Omega = \mathbb{E}_X \mathbb{E}_{P(\mathbf{T})} q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \left( \frac{2}{n} \sum_{m=1}^n X_{\mathbf{T}_{1,m}} - \frac{2}{n} \sum_{m=1}^n X_{\mathbf{T}_{0,m}} \right) \cdot \sum_{k=1}^K g_\theta(e_k) P_k \leq \frac{\Delta}{\sqrt{n}},$$

Let us back to the evaluation of the exponential moment in Eq. (29), we will evaluate the following

$$\mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q} | \mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m} - \lambda \Omega \right) + \Omega. \quad (31)$$

We then evaluate this similarly to Eq. (28), which uses the negative association of the permutation distribution and McDiarmid's inequality. The the exponential moment is upper bounded by  $(2\Delta\lambda/n)^2/8 \times 2n = \lambda^2 \Delta^2/n$  We then obtain

$$\begin{aligned}
& \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{\mathbf{T}_{1,m}}) q(\mathbf{e}, \phi, \theta | X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
& - \mathbb{E}_{X, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \phi, X_{\mathbf{T}_{0,m}}) q(\mathbf{e}, \phi, \theta | X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
& \leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q} | \mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp \left( \frac{2\lambda}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \sum_{k=1}^K g_\theta(e_k) \mathbb{1}_{k=J_m} - \lambda \Omega \right) + \Omega \\
& \leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q} | \mathbf{P}) + \frac{\lambda \Delta^2}{n} + \frac{\Delta}{\sqrt{n}}. \quad (32)
\end{aligned}$$

1267 In conclusion, from Eqs. (28) and (32) we have

$$\text{gen}(n, \mathcal{D}) \leq \mathbb{E}_X \frac{2}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}) + \frac{5\lambda\Delta^2}{4n} + \frac{\Delta}{\sqrt{n}},$$

1268 and optimizing the  $\lambda$ , we have

$$\text{gen}(n, \mathcal{D}) \leq 2\Delta \sqrt{\frac{5\mathbb{E}_X \text{KL}(\mathbf{Q}|\mathbf{P})}{2n}} + \frac{\Delta}{\sqrt{n}}.$$

1269 We can slightly improve the coefficient of the first term in the above bound as follows. The above  
 1270 proof follows the approach in Appendix D.1. We separately apply the Donsker-Valadhan lemma for  
 1271 the first two terms and latter two terms in Eq. (25). However, since the posterior and prior distributions  
 1272 used for the Donsker-Valadhan lemma are the same as shown in Eq. (26), we only need to use the  
 1273 Donsker-Valadhan lemma once. This leads to an improved coefficient.

1274 Specifically, the proof goes as follows; combining Eqs. (27) and (31), we have simultaneously treat  
 1275 all terms in Eq. (25). By Donsker-Valadhan lemma, we have

$$\begin{aligned} & \text{gen}(n, \mathcal{D}) \\ & \leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \\ & \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_{\theta}(e_k)) (\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m}) + \frac{2\lambda}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \sum_{k=1}^K g_{\theta}(e_k) \mathbb{1}_{k=J_m} - \lambda\Omega \right) + \Omega. \end{aligned}$$

1276 From the negative association property, the exponential moment term can be upper-bounded as

$$\begin{aligned} & \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1})} \mathbb{E}_{P(\mathbf{T}')} q(\tilde{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}) \mathbb{E} \prod_{j=1}^n \prod_{j=0,1} P(\mathbf{T}_{j,m}'') \\ & \exp \left( \frac{\lambda}{n} \sum_{k=1}^K \sum_{m=1}^n l(X_{\mathbf{T}_{0,m}}, g_{\theta}(e_k)) (\mathbb{1}_{k=J_{\mathbf{T}_{0,m}}''} - \mathbb{1}_{k=J_{\mathbf{T}_{1,m}}''}) + \frac{2\lambda}{n} \sum_{m=1}^n (X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}) \cdot \sum_{k=1}^K g_{\theta}(e_k) \mathbb{1}_{k=J_{\mathbf{T}_{1,m}}''} - \lambda\Omega \right), \end{aligned}$$

1277 Since  $\{\mathbf{T}_{j,m}''\}$  are independent, we can apply McDiarmid's inequality. The the exponential moment  
 1278 is upper bounded by  $((1+2)\Delta\lambda/n)^2/8 \times 2n = 9\lambda^2\Delta^2/4n$ . Thus, we have

$$\text{gen}(n, \mathcal{D}) \leq \mathbb{E}_X \frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}) + 9\lambda^2\Delta^2/4n + \frac{\Delta}{\sqrt{n}}.$$

1279 By optimizing  $\lambda$ , we have

$$\text{gen}(n, \mathcal{D}) \leq 3\Delta \sqrt{\frac{\mathbb{E}_X \text{KL}(\mathbf{Q}|\mathbf{P})}{n}} + \frac{\Delta}{\sqrt{n}}.$$

## 1280 E.2 Proof of Eq. (10) and discussion about the deterministic encoder

1281 First, we can show

$$\mathbb{E}_{\tilde{X}, \mathbf{T}} \mathbb{E}_{q(\mathbf{e}, \phi | \tilde{X}_{\mathbf{T}_1})} \text{KL}(\mathbf{Q}_{\tilde{\mathbf{J}}, \mathbf{T}} \| \mathbf{Q}_{\tilde{\mathbf{J}}}) \leq I(\mathbf{e}, \phi; \mathbf{T} | \tilde{X}) + I(\tilde{\mathbf{J}}; \mathbf{T} | \mathbf{e}, \phi, \tilde{X}).$$

1282 exactly same way as Appendix D.6.

1283 By the definition of the CMI, the CMI is expressed as the difference of entropy and conditional  
 1284 entropy. Since  $\tilde{J}$  is discrete, the entropy is always larger than 0. Thus, we have

$$I(\tilde{\mathbf{J}}; \mathbf{T} | \mathbf{e}, \phi, \tilde{X}) \leq H[\tilde{\mathbf{J}} | \mathbf{e}, \phi, \tilde{X}] \leq H[\tilde{\mathbf{J}} | \tilde{X}].$$

1285 where  $H$  is the Shannon entropy. Note that the entropy is bounded by the growth function, i.e., the  
 1286 maximum number of different ways in which a dataset of size  $2n$  can be classified in  $K$ . And such  
 1287 quantity is bounded in the proof of Theorem 8 of Harutyunyan et al. [30], thus

$$I(\tilde{\mathbf{J}}; \mathbf{T} | \mathbf{e}, \phi, \tilde{X}) \leq d_K \log \left( \binom{K}{2} \frac{2en}{d_K} \right).$$

1288 holds similarly to Eq. (24).

1289 Thus, by regularizing the capacity of the encoder model (via the Natarajan dimension), the CMI term  
 1290  $I(\tilde{\mathbf{J}}; \mathbf{T} | \mathbf{e}, \phi, \tilde{X})/n$  scales as  $\mathcal{O}(\log n)$ . See Appendix D.7 for the additional discussion.



### 1291 E.3 Proof of Theorem 4

1292 To prove the theorem, we prove a more general result than Theorem 4, and then we apply that result  
1293 to the specific setting of Theorem 4. Therefore, we first derive such a general result.

#### 1294 E.3.1 Discretization in encoder function

1295 Here, we present the results for a general stochastic encoder. For fixed  $\phi$  and  $\mathbf{e}$ , assume that for  
1296 all  $\mathbf{x} \in \tilde{X}$ , for any  $j \in [K]$ , and for a fixed  $\delta \in \mathbb{R}^+$ , the following holds:  $q(J = j|\mathbf{e}, f_\phi(x)) \leq$   
1297  $e^{h(\delta)} q(J = j|\mathbf{e}, \hat{f}(x))$  with  $h : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ .

1298 **Theorem 6.** Assume that there exists a positive constant  $\Delta_z$  such that  $\sup_{z, z' \in \mathcal{Z}} \|z - z'\| < \Delta_z$ .  
1299 Then, when using Eq. (2) and under the same setting as Theorem 3, for any  $\delta \in (0, 1]$ , we have

$$\text{gen}(n, \mathcal{D}) \leq 2\Delta\sqrt{nh(\delta)} + 3\Delta\sqrt{\frac{2\log\mathcal{N}(\delta, \mathcal{F}, 2n)}{n}} + \frac{\Delta}{\sqrt{n}}.$$

1300 We can show that Eq. (2) satisfies  $h(\delta) = 8\beta\Delta_z\delta$ , see Appendix E.3.3 for this proof. Thus by  
1301 substituting this into the above Theorem, we obtain Theorem 4.

1302 *Proof.* When analyzing the contribution of the encoder model to generalization, it is often necessary  
1303 to discretize the function or parameters of the encoder to control the CMI using the metric entropy of  
1304 the model. To achieve this, we consider a  $\delta$ -cover  $\hat{f}$  of the function  $f$ . In this derivation, we examine  
1305 both the supersample and permutation-invariant settings, highlighting that the supersample setting  
1306 fails to establish a uniform convergence bound.

1307 First, we begin with the supersample setting. Given a supersample  $\tilde{X}$ , we recall the definition of  
1308 the indices. In this theorem, we focus on the distribution of the index defined by the codebook  $\mathbf{e}$   
1309 and  $z \in \mathcal{Z}$ , where  $z$  represents the output of the encoder  $f_\phi(\cdot)$ . Thus, we express it as  $q(J|\mathbf{e}, z)$ .  
1310 Moreover, in this section, we use the notation  $q(\mathbf{e}, \phi, \theta|\tilde{X}, U) = q(\mathbf{e}, \phi, \theta|\tilde{X}_U)$ . The joint distribution  
1311 is then given by:

$$\begin{aligned}\mathbf{Q}' &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta|\tilde{X}, U)q(\tilde{\mathbf{J}}|\mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}}|\mathbf{f}, U)p(\mathbf{f}|\phi, \tilde{X}), \\ \mathbf{Q}'_\delta &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta|\tilde{X}, U)q(\tilde{\mathbf{J}}|\mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}}|\hat{\mathbf{f}}, U)p(\hat{\mathbf{f}}|\mathbf{f})p(\mathbf{f}|\phi, \tilde{X}),\end{aligned}$$

1312 where  $q(\tilde{\mathbf{J}}|\mathbf{e}, \tilde{\mathbf{f}})$  represents the elementwise application of  $q(J|\mathbf{e}, \cdot)$  to  $\tilde{\mathbf{f}} \in \mathcal{Z}^{2n}$ . And  $p(\mathbf{f}|\phi, \tilde{X})$  is  
1313 the elementwise application of  $p(\mathbf{f}|\phi, \cdot)$  to  $\tilde{X}$ , which simply computes the encoder output for each  
1314 sample in  $\tilde{X}$ .

1315 Then, in  $p(\hat{\mathbf{f}}|\mathbf{f})$ , the discretization process is performed using the  $\delta$ -cover (thus, it is represented by  
1316 the Dirac mass). We express this as  $p(\hat{\mathbf{f}}|\mathbf{f}) = \delta(\hat{\mathbf{f}}, \hat{\mathbf{f}}^\phi)$ , where  $\hat{\mathbf{f}}^\phi$  is the selected point from the  $\delta$ -cover.  
1317 Then, for  $p(\tilde{\mathbf{f}}|\hat{\mathbf{f}}, U)$ , we randomly shuffle  $\hat{\mathbf{f}} \in \mathbb{R}^{2n}$  with  $U$ , formally defining  $\hat{\mathbf{f}}_{\tilde{U}} := (\mathbf{f}_U, \mathbf{f}_{\tilde{U}})$ . Thus,  
1318 we write  $p(\tilde{\mathbf{f}}|\hat{\mathbf{f}}, U) = \delta(\tilde{\mathbf{f}}, \hat{\mathbf{f}}_{\tilde{U}})$ . Similarly, we define  $p(\tilde{\mathbf{f}}|\mathbf{f}, U) = \delta(\tilde{\mathbf{f}}, \mathbf{f}_{\tilde{U}})$ .

1319 This definition differs slightly from the posterior distribution in Eq. (14), where we first shuffle  $\tilde{X}$   
1320 with  $U$  before passing it through the encoder. This simple modification allows us to derive the bound  
1321 based on metric entropy. When evaluating the generalization error bound, we are only concerned  
1322 with  $\tilde{J}$ . By integrating out  $\tilde{\mathbf{f}}$ ,  $\phi$ , and  $\hat{\mathbf{f}}$ , we focus on the following posterior distributions:

$$\begin{aligned}\mathbf{Q} &:= P(\tilde{X})P(U)q(\mathbf{e}, \mathbf{f}, \theta|\tilde{X}, U)p(\tilde{\mathbf{J}}|\mathbf{e}, \mathbf{f}_{\tilde{U}}), \\ \mathbf{Q}_\delta &:= P(\tilde{X})P(U)q(\mathbf{e}, \mathbf{f}, \theta|\tilde{X}, U)p(\tilde{\mathbf{J}}|\mathbf{e}, \hat{\mathbf{f}}_{\tilde{U}}^\phi).\end{aligned}$$

1323 To prove this lemma, we first replace the output of the encoder with that obtained using the  $\delta$ -cover  
 1324 of the encoder network. First note that the generalization error can be written as

$$\begin{aligned} \text{gen}(n, \mathcal{D}) &= \mathbb{E}_{p(\tilde{X})P(U)} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\tilde{J}_m | \mathbf{e}, \mathbf{f}_\phi(X_{m, \tilde{U}_m}))} q(\mathbf{e}, \phi, \theta | X_U) l(X_{m, \tilde{U}_m}, g_\theta(e_k)) \mathbb{1}_{k=\tilde{J}_m} \\ &\quad - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \mathbf{f}_\phi(X_{m, U_m}))} q(\mathbf{e}, \phi, \theta | X_U) l((X_{m, U_m}, g_\theta(e_k)) \mathbb{1}_{k=J_m}) \\ &= \mathbb{E}_{p(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | X, U)p(\tilde{\mathbf{J}} | \mathbf{e}, \mathbf{f}_{\tilde{U}})} \left[ \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n l(X_{m, \tilde{U}_m}, g_\theta(e_k)) \mathbb{1}_{k=\tilde{J}_m} - l((X_{m, U_m}, g_\theta(e_k)) \mathbb{1}_{k=J_m}) \right]. \end{aligned}$$

1325 We also define the generalization under the delta cover of original function, conditioned o

$$\begin{aligned} \text{gen}(n, \mathcal{D}, \delta) &:= \mathbb{E}_{p(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | X, U)} \left[ \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\tilde{J}_m | \mathbf{e}, \hat{\mathbf{f}}(X_{m, \tilde{U}_m}))} l(X_{m, \tilde{U}_m}, g_\theta(e_k)) \mathbb{1}_{k=\tilde{J}_m} \right. \\ &\quad \left. - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m | \mathbf{e}, \hat{\mathbf{f}}(X_{m, U_m}))} l((X_{m, U_m}, g_\theta(e_k)) \mathbb{1}_{k=J_m}) \right] \\ &= \mathbb{E}_{p(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | X, U)p(\tilde{\mathbf{J}} | \mathbf{e}, \hat{\mathbf{f}}_{\tilde{U}}^\phi)} \left[ \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n l(X_{m, \tilde{U}_m}, g_\theta(e_k)) \mathbb{1}_{k=\tilde{J}_m} - l((X_{m, U_m}, g_\theta(e_k)) \mathbb{1}_{k=J_m}) \right]. \end{aligned}$$

1326 For the latter purpose, we define

$$\Delta_L := \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n l(X_{m, \tilde{U}_m}, g_\theta(e_k)) \mathbb{1}_{k=\tilde{J}_m} - \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n l((X_{m, U_m}, g_\theta(e_k)) \mathbb{1}_{k=J_m}).$$

1327 To evaluate these gap, we apply the Donsker-Valadhan lemma between the two distributions  $\mathbf{Q}_J$  and  
 1328  $\mathbf{Q}_{\delta, J}$ .

$$\begin{aligned} \text{gen}(n, \mathcal{D}) & \leq \text{gen}(n, \mathcal{D}, \delta) + \mathbb{E}_{U, \tilde{X}} \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} \frac{1}{\lambda} \text{KL}(\mathbf{Q} \| \mathbf{Q}_\delta) + \mathbb{E}_{U, \tilde{X}} \mathbb{E}_{q(\mathbf{e}, \phi, \theta | X_U)} \frac{1}{\lambda} \log \mathbb{E}_{p(\tilde{\mathbf{J}} | \mathbf{e}, \hat{\mathbf{f}}_{\tilde{U}}^\phi)} \exp \left( \lambda \Delta_L - \mathbb{E}_{p(\tilde{\mathbf{J}} | \mathbf{e}, \hat{\mathbf{f}}_{\tilde{U}}^\phi)} \lambda \Delta_L \right) \\ & \leq \text{gen}(n, \mathcal{D}, \delta) + \frac{2nh(\delta)}{\lambda} + \frac{\lambda \Delta^2}{2}, \end{aligned} \tag{33}$$

1329 where we evaluated the KL divergence as

$$\text{KL}(\mathbf{Q} \| \mathbf{Q}_\delta) = \mathbb{E}_{\mathbf{Q}} \log \frac{\mathbf{Q}}{\mathbf{Q}_\delta} \leq 2nK \log e^{h(\delta)} = 2nh(\delta).$$

1330 The inequality is owing to the proper that for all  $\mathbf{x} \in \tilde{X}$ , for any  $j \in [K]$ , and for a fixed  $\delta \in \mathbb{R}^+$ ,  
 1331  $q(J = j | \mathbf{e}, f_\phi(x)) \leq e^{h(\delta)} q(J = j | \mathbf{e}, \hat{f}(x))$  holds by assumption. We also evaluated the exponential  
 1332 moment term by using the fact that  $-\lambda \Delta \leq \lambda l(X, g_\theta(e_J)) - \frac{\lambda}{n} \sum_{m=1}^n l(S_m, g_\theta(e_{J_m})) \leq \lambda \Delta$  to  
 1333 upper bound the exponential moment.

1334 This implies that the first term corresponds to the generalization bound when using the  $\delta$ -cover of the  
 1335 encoder network. We can bound this term similarly to Theorem 2,

$$\text{gen}(n, \mathcal{D}, \delta) \leq 2\Delta \sqrt{\frac{\text{KL}(\mathbf{Q}'_\delta \| \mathbf{P}'_\delta) + \text{KL}(\mathbf{Q}'_\delta \| \mathbf{P})}{n}} + \frac{\Delta}{\sqrt{n}},$$

1336 where we consider the following posterior and data-dependent, and data-independent prior distribu-  
 1337 tions:

$$\begin{aligned} \mathbf{Q}'_\delta &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | \tilde{X}, U)q(\tilde{\mathbf{J}} | \mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}} | \hat{\mathbf{f}}, U)p(\hat{\mathbf{f}} | \mathbf{f})p(\mathbf{f} | \phi, \tilde{X}), \\ \mathbf{P}'_\delta &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | \tilde{X}, U)p(\tilde{\mathbf{J}} | \mathbf{e}, \tilde{\mathbf{f}})\mathbb{E}_{U'} p(\tilde{\mathbf{f}} | \hat{\mathbf{f}}, U')p(\hat{\mathbf{f}} | \mathbf{f})p(\mathbf{f} | \phi, \tilde{X}), \\ \mathbf{P} &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | \tilde{X}, U)q(\tilde{\mathbf{J}} | \mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}} | \hat{\mathbf{f}}, U)\pi(\hat{\mathbf{f}})p(\mathbf{f} | \phi, \tilde{X}), \end{aligned}$$

1338 where  $\pi(\hat{\mathbf{f}})$  is the data independent prior distribution over the  $\delta$ -covering, such as the uniform  
 1339 distribution.

1340 Combining these, we have

$$\text{gen}(n, \mathcal{D}) \leq 2\Delta\sqrt{nh(\delta)} + 2\Delta\sqrt{\frac{\text{KL}(\mathbf{Q}'_\delta\|\mathbf{P}'_\delta) + \text{KL}(\mathbf{Q}'_\delta\|\mathbf{P})}{n}} + \frac{\Delta}{\sqrt{n}}.$$

1341 As for the CMI term, we have

$$\text{KL}(\mathbf{Q}'_\delta\|\mathbf{P}'_\delta) \leq 2\log\mathcal{N}(\delta, \mathcal{F}, 2n). \quad (34)$$

1342 The proof of Eq. (34) is shown in below and this term can be bounded  $\mathcal{O}(\log n)$  under moderate  
 1343 assumptions.

1344 However, the second term  $\text{KL}(\mathbf{Q}'_\delta\|\mathbf{P})$ , which corresponds to the empirical KL term, cannot be small  
 1345 as discussed in Theorem 2. That is, under the settings of Lemma 2, the empirical KL behaves  $\mathcal{O}(1)$ ,  
 1346 which is undesirable behavior.

1347 So we consider using the permutation symmetric setting. We can proceed the discretization almost  
 1348 the same in the above super sample setting. Under this distribution, the generalization gap can again  
 1349 upper bounded similar to Eq. (33). Then from Theorem 3, we have

$$\begin{aligned} \text{gen}(n, \mathcal{D}) &\leq \mathbb{E}_{\tilde{X}, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(\tilde{J}_m|\mathbf{e}, \hat{f}(X_{\mathbf{T}_{0,m}}))q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=\tilde{J}_m} \\ &\quad - \mathbb{E}_{\tilde{X}, \mathbf{T}} \sum_{k=1}^K \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \hat{f}(X_{\mathbf{T}_{1,m}}))q(\mathbf{e}, \phi, \theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} + 2\Delta\sqrt{nh(\delta)} \\ &\leq 3\Delta\sqrt{\frac{\text{KL}(\mathbf{Q}'_\delta\|\mathbf{P}'_\delta)}{n}} + \frac{\Delta}{\sqrt{n}} + 2\Delta\sqrt{nh(\delta)}, \end{aligned}$$

1350 where

$$\begin{aligned} \mathbf{Q}'_\delta &:= P(\tilde{X})P(\mathbf{T})q(\mathbf{e}, \phi, \theta|\tilde{X}, \mathbf{T})q(\tilde{\mathbf{J}}|\mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}}|\hat{\mathbf{f}}, \mathbf{T})p(\hat{\mathbf{f}}|\mathbf{f})p(\mathbf{f}|\phi, \tilde{X}), \\ \mathbf{P}'_\delta &:= P(\tilde{X})P(\mathbf{T})q(\mathbf{e}, \phi, \theta|\tilde{X}, \mathbf{T})p(\tilde{\mathbf{J}}|\mathbf{e}, \tilde{\mathbf{f}})\mathbb{E}_{\mathbf{T}'}p(\tilde{\mathbf{f}}|\hat{\mathbf{f}}, \mathbf{T}')p(\hat{\mathbf{f}}|\mathbf{f})p(\mathbf{f}|\phi, \tilde{X}), \end{aligned}$$

1351 We can show that

$$\text{KL}(\mathbf{Q}'_\delta\|\mathbf{P}'_\delta) \leq 2\log\mathcal{N}(\delta, \mathcal{F}, 2n). \quad (35)$$

1352 see Appendix E.3.2 for the proof. We can analyze the behavior of the upper bound of Eq. (35) in  
 1353 Appendix E.4.

1354 Thus, we have

$$\text{gen}(n, \mathcal{D}) \leq 3\Delta\sqrt{\frac{2\log\mathcal{N}(\delta, \mathcal{F}, 2n)}{n}} + \frac{\Delta}{\sqrt{n}} + 2\Delta\sqrt{nh(\delta)}.$$

1355

□

### 1356 E.3.2 Proof of Eq. (34)

1357 We consider the following posterior and data-dependent prior distributions

$$\begin{aligned} \mathbf{Q} &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta|\tilde{X}, U)p(\tilde{\mathbf{J}}|\mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}}|\mathbf{f}, U)p(\mathbf{f}|\phi, \tilde{X}) \\ \mathbf{P}_S &:= P(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta|\tilde{X}, U)p(\tilde{\mathbf{J}}|\mathbf{e}, \mathbf{f})\mathbb{E}_{p(U')}p(\tilde{\mathbf{f}}|\mathbf{f}, U')p(\mathbf{f}|\phi, \tilde{X}) \end{aligned}$$

1358 When using the Donsker-Valadhan inequality, all calculation remains the same except for the KL  
 1359 divergence term as described below

$$\begin{aligned}
\mathbb{E}_{\mathbf{Q}} \log \frac{\mathbf{Q}}{\mathbf{P}_S} &= \mathbb{E}_{p(\tilde{X})P(U)q(\mathbf{e}, \phi, \theta | \tilde{X}, U)p(\tilde{\mathbf{J}} | \mathbf{e}, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}} | \mathbf{f}, U)p(\mathbf{f} | \phi, \tilde{X})} \log \frac{p(\tilde{\mathbf{f}} | \mathbf{f}, U)}{\mathbb{E}_{p(U')} p(\tilde{\mathbf{f}} | \mathbf{f}, U')} \\
&= \mathbb{E}_{p(\tilde{X})P(U)q(\mathbf{f} | X, U)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{p(\tilde{\mathbf{f}} | \mathbf{f}, U)}{\mathbb{E}_{p(U')} p(\tilde{\mathbf{f}} | \mathbf{f}, U')} \\
&= \mathbb{E}_{p(\tilde{X})P(\mathbf{f} | X)} \mathbb{E}_{P(U | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{p(\tilde{\mathbf{f}} | \mathbf{f}, U)}{\mathbb{E}_{p(U')} p(\tilde{\mathbf{f}} | \mathbf{f}, U')} \\
&= \mathbb{E}_{p(\tilde{X})P(\mathbf{f} | X)} \mathbb{E}_{P(U | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{p(\tilde{\mathbf{f}} | \mathbf{f}, U)}{\mathbb{E}_{p(U' | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U')} \\
&\quad + \mathbb{E}_{p(\tilde{X})P(\mathbf{f} | X)} \mathbb{E}_{P(U | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{\mathbb{E}_{p(U' | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U')}}{\mathbb{E}_{p(U')} p(\tilde{\mathbf{f}} | \mathbf{f}, U')} \\
&= I(\tilde{\mathbf{f}}; U | \mathbf{f}, X) + \mathbb{E}_{p(\tilde{X})P(\mathbf{f} | X)} \mathbb{E}_{P(U | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{\mathbb{E}_{p(U' | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U')}}{\mathbb{E}_{p(U')} p(\tilde{\mathbf{f}} | \mathbf{f}, U')} \\
&\leq I(\tilde{\mathbf{f}}; U | \mathbf{f}, X) + \mathbb{E}_{p(\tilde{X})P(\mathbf{f} | X)} \mathbb{E}_{P(U | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{p(U' | \mathbf{f}, X)}{p(U')} \\
&= I(\tilde{\mathbf{f}}; U | \mathbf{f}, X) + \mathbb{E}_{p(\tilde{X})P(\mathbf{f} | X)} \mathbb{E}_{P(U | \mathbf{f}, X)p(\tilde{\mathbf{f}} | \mathbf{f}, U)} \log \frac{p(U' | X)p(\mathbf{f} | U', X)}{\mathbb{E}_{p(U' | X)p(\mathbf{f} | U', X)p(U')} \\
&= I(\tilde{\mathbf{f}}; U | \mathbf{f}, X) + I(\mathbf{f}; U | X)
\end{aligned}$$

1360 We can derive the similar arguments for  $\mathbf{Q}'_\delta$  and  $\mathbf{P}'_\delta$ , and we have

$$\text{KL}(\mathbf{Q}'_\delta \| \mathbf{P}'_\delta) \leq I(\tilde{\mathbf{f}}; U | \hat{\mathbf{f}}, X) + I(\hat{\mathbf{f}}; U | X)$$

1361 Note that we consider the CMI for the discrete variable, it is upper bounded by the entropy [15], and  
 1362 we have

$$I(\tilde{\mathbf{f}}; U | \hat{\mathbf{f}}, X) \leq H[\tilde{\mathbf{f}} | \hat{\mathbf{f}}, X] - H[\tilde{\mathbf{f}} | U, \mathbf{f}, X] \leq H[\tilde{\mathbf{f}} | X] \leq \log \mathcal{N}(\delta, \mathcal{F}, 2n).$$

1363 and

$$I(\hat{\mathbf{f}}; U | X) \leq H[\hat{\mathbf{f}} | X] - H[\hat{\mathbf{f}} | U, X] \leq H[\hat{\mathbf{f}} | X] \leq \log \mathcal{N}(\delta, \mathcal{F}, 2n).$$

1364 The first inequality follows from the fact that MI is defined as the difference between the entropy  
 1365 and the conditional entropy, and the entropy of discrete variables is always non-negative. The  
 1366 second inequality arises because  $\tilde{\mathbf{J}}, \mathbf{J}$  are outputs of a function evaluated at  $2n$  points. Thus, we  
 1367 considered the covering number at  $2n$  points, defined as  $\mathcal{N}(\delta, \mathcal{F}, n) := \sup_{x^{2n} \in \mathcal{X}^{2n}} \mathcal{N}(\delta, \mathcal{F}, x^{2n})$ .  
 1368 Since the entropy is bounded above by the logarithm of the maximum cardinality, we obtain the  
 1369 second inequality.

### 1370 E.3.3 Behavior of Eq. (2)

1371 Finally, we show that Eq. (2) satisfies  $h(\delta) = 8\beta\Delta_z\delta$  because

$$\begin{aligned}
&\frac{q(J = j | \mathbf{e}, f_\phi(x))}{q(J = j | \mathbf{e}, \hat{f}(x))} \\
&= \frac{e^{-\beta\|f_\phi(x) - e_j\|^2}}{e^{-\beta\|\hat{f}(x) - e_j\|^2}} \times \frac{\sum_{k=1}^K e^{-\beta\|\hat{f}(x) - e_k\|^2}}{\sum_{k=1}^K e^{-\beta\|f_\phi(x) - e_k\|^2}} \\
&= e^{-\beta\|f_\phi(x) - e_j\|^2 + \beta\|\hat{f}(x) - e_j\|^2} \times \frac{\sum_{k=1}^K e^{\beta\|f_\phi(x) - e_k\|^2}}{\sum_{k=1}^K e^{\beta\|\hat{f}(x) - e_k\|^2}} \\
&\leq e^{\beta(\hat{f}(x) - f_\phi(x)) \cdot (\hat{f}(x) + f_\phi(x)) - 2\beta e_j \cdot (\hat{f}(x) - f_\phi(x))} \times \sup_{k \in [K]} e^{-\beta\|\hat{f}(x) - e_k\|^2 + \beta\|f_\phi(x) - e_k\|^2} \\
&\leq e^{4\beta\Delta_z\delta} \times e^{4\beta\Delta_z\delta}.
\end{aligned}$$

#### 1372 E.4 Discussion about the metric entropy for regularized model

1373 Here we discuss the upper bound of metric entropy in our setting. Since the latent variable lies in  
1374  $\mathbb{R}^{d_z}$ , the encoder network operates as  $f_\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_z}$ , making it a multivariate function.

1375 Let us define a function class  $\mathcal{F}_i : \mathcal{X} \rightarrow \mathbb{R}$  for  $i = 1 \dots, d_z$  and define  $\mathcal{F}_0 = \prod_{i=1}^{d_z} \mathcal{F}_i$ . Then by  
1376 definition,  $\mathcal{F} \subset \mathcal{F}_0$  holds. We define the covering number for each  $\mathcal{F}_i$ ; Given  $x^n := (x_1, \dots, x_n) \in$   
1377  $\mathcal{X}^n$ , define the pseudo-metric  $d'_n$  on  $\mathcal{F}_i$  as  $d'_n(f, g) := \max_{i \in [n]} |f(x_i) - g(x_i)|$  for  $f, g \in \mathcal{F}_i$ . The  
1378  $\delta$ -covering number of  $\mathcal{F}_i$  with respect to  $d'_n$  is denoted as  $\mathcal{N}(\delta, \mathcal{F}_i, x^n)$ , and we define  $\mathcal{N}(\delta, \mathcal{F}_i, n) :=$   
1379  $\sup_{x^n \in \mathcal{X}^n} \mathcal{N}(\delta, \mathcal{F}_i, x^n)$ . Then by definition, the cardinality of  $\mathcal{F}$  is smaller than  $\mathcal{F}_0$ , so we have

$$\mathcal{N}(\delta, \mathcal{F}, n) \leq \prod_{i=1}^{d_z} \mathcal{N}(\delta, \mathcal{F}_i, n).$$

1380 We can see a similar argument in Lemma 1 in Guermur [26], which considers more general settings.

1381 For simplicity, we assume that  $\mathcal{F}' = \mathcal{F}_1 = \dots = \mathcal{F}_{d_z}$  holds. Then, we can rewrite Theorem 4 as  
1382 follows

$$\text{gen}(n, \mathcal{D}) \leq 4\Delta \sqrt{2n\beta\Delta_z\delta} + 3\Delta \sqrt{\frac{2d_z \log \mathcal{N}(\delta, \mathcal{F}', 2n)}{n}} + \frac{\Delta}{\sqrt{n}}.$$

1383 For example, assume that the encoder function, which has  $d_\phi$  dimensional parameters, shows  $L_0$ -  
1384 Lipschitz continuity ( $L_0 > 0$ ) with respect to parameter, then we can obtain  $\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \delta) \asymp$   
1385  $d_\phi \log \frac{L_0}{\delta}$  [72]. Thus, by setting  $\delta = \mathcal{O}(1/(n))$ , we have that

$$\text{gen}(n, \mathcal{D}) = \mathcal{O} \left( \sqrt{\frac{d_\phi d_z \log(n)}{n}} \right)$$

1386 Instead of using the assumption of parametric function class, the metric entropy can be bounded  
1387 by the fat-shattering dimension of each function, as discussed in Lemma 3.5 of Alon et al. [5].  
1388 Examples of fat-shattering dimension evaluations can be found, for instance, in Bartlett & Maass [7],  
1389 which discusses neural network models, and Gottlieb et al. [24], which addresses the fat-shattering  
1390 dimension of Lipschitz function classes. If our encoder network adheres to these properties, we can  
1391 bound its covering number accordingly.

1392 As discussed in Appendix D.7.1, when we use the deterministic decoder, we can use the Natarajan  
1393 dimension to quantify the complexity of the LVs and such Natarajan dimension can be bounded  
1394 by the fat-shattering dimension. Thus, it is essential to bound the fat-shattering dimension in both  
1395 deterministic and stochastic settings.

#### 1396 F Proof of Theorem 5

1397 Before the proof, we define the Wasserstein distance. Given a metric  $d(\cdot, \cdot)$  and probability distribu-  
1398 tions  $p$  and  $q$  on  $\mathcal{X}$ , let  $\Pi(p, q)$  denote the set of all couplings of  $p$  and  $q$ . The 2-Wasserstein distance  
1399 is defined as:

$$W_2(p, q) = \sqrt{\inf_{\rho \in \Pi} \int_{\mathcal{X} \times \mathcal{X}} d(x, x')^2 d\rho(x, x')}.$$

1400 In this work, we use the Euclidean metric  $\|\cdot\|$  as  $d(\cdot, \cdot)$ .

1401 Next, we define the pushforward. Let  $\pi$  represent a distribution on  $\mathcal{Z}$ , and let us assume that for any  
1402  $\theta \in \Theta$ , the decoder  $g_\theta(\cdot) : \mathcal{Z} \rightarrow \mathcal{X}$  is measurable. The pushforward of the distribution  $\pi$  by the  
1403 decoder, denoted as  $g_\theta \# \pi$ , defines a distribution on  $\mathcal{X}$  as  $g_\theta \# \pi(A) = \pi(g_\theta^{-1}(A))$  for any measurable  
1404 set  $A \subseteq \mathcal{X}$ .

1405 *Proof.* Conditioned on the encoder parameter, codebook, and input  $X$ , selecting the index  $J$  corre-  
1406 sponds to selecting the latent representation  $e_J$ . Since the posterior over the index is  $q(J|\mathbf{e}, \phi, X)$ , we  
1407 express the posterior imposed on the latent representation as  $q(e = e_j|\mathbf{e}, \phi, X)$  for all  $j = 1, \dots, K$ .

Using this notation, we first define the distribution obtained by the training dataset as follows; conditioned on  $\mathbf{e}, \phi, S$ , we have

$$\hat{\mu}_S = \frac{1}{n} \sum_{m=1}^n g_\theta \# q(e|\mathbf{e}, \phi, S_m).$$

From the triangle inequality, we have

$$W_2(\mathcal{D}, \hat{\mu}) \leq W_2(\mathcal{D}, \hat{\mu}_S) + W_2(\hat{\mu}_S, \hat{\mu}). \quad (36)$$

We then have

$$W_2^2(\mathcal{D}, \hat{\mu}) \leq 2W_2^2(\mathcal{D}, \hat{\mu}_S) + 2W_2^2(\hat{\mu}_S, \hat{\mu}).$$

The first term of Eq. (36) is bounded as follows;

$$\begin{aligned} \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} W_2^2(\mathcal{D}, \hat{\mu}_S) &\leq \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \mathbb{E}_X \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(e|\mathbf{e}, \phi, S_m)} \|X - g_\theta(e)\|^2 \\ &= \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \mathbb{E}_X \sum_{k=1}^K \|X - g_\theta(e_k)\|^2 \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, S_m)} \mathbb{1}_{k=J_m} \end{aligned} \quad (37)$$

The first inequality is obtained by the definition of the Wasserstein distance.

This term corresponds to the first term of Eq. (17), where  $X$  corresponds to the test data  $X_m, \tilde{U}_m$ . Therefore, Eq. (37) can be upper-bounded by applying Eq. (21), which serves as the upper bound for Eq. (17).

$$\begin{aligned} \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} W_2^2(\mathcal{D}, \hat{\mu}_S) \\ \leq \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(J_m|\mathbf{e}, \phi, S_m)} \|S_m - g_\theta(e_{J_m})\|^2 + \frac{1}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}) + \frac{\lambda \Delta^2}{2n} + \frac{\Delta}{\sqrt{n}}, \end{aligned} \quad (38)$$

where

$$\mathbf{Q} := q(\mathbf{e}, \phi, \theta|S) \prod_{m=1}^n q(J_m|\mathbf{e}, \phi, S_m), \quad \mathbf{P} := q(\mathbf{e}, \phi, \theta|S) \prod_{m=1}^n \pi(J_m|\mathbf{e}, \phi).$$

Next, the second term of Eq. (36) is bounded as follows; we use the weighted CKP inequality [11]. From the particular case 2.5. in Bolley & Villani [11], we directly have

$$\begin{aligned} \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} W_2^2(\hat{\mu}_S, \hat{\mu}) &\leq \Delta \sqrt{2\text{KL}(\hat{\mu}_S \|\hat{\mu})} \leq \Delta \sqrt{2 \frac{1}{n} \sum_{m=1}^n \text{KL}(g_\theta \# q(e|\mathbf{e}, \phi, S_m) \| g_\theta \# \pi(e|\mathbf{e}, \phi))} \\ &\leq \Delta \sqrt{2 \frac{1}{n} \sum_{m=1}^n \text{KL}(q(J_m|\mathbf{e}, \phi, S_m) \|\pi(J_m|\mathbf{e}, \phi))} \end{aligned} \quad (39)$$

Combining Eqs. (38) and (39), we have

$$\begin{aligned} \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} W_2^2(\mathcal{D}, \hat{\mu}) &\leq 2\mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \frac{1}{n} \sum_{m=1}^n \mathbb{E}_{q(e_{(m)}|\mathbf{e}, \phi, S_m)} \|S_m - g_\theta(e_{(m)})\|^2 \\ &\quad + \frac{2}{\lambda} \text{KL}(\mathbf{Q}|\mathbf{P}) + \frac{\lambda \Delta^2}{n} + \frac{2\Delta}{\sqrt{n}} + 2\Delta \sqrt{2 \frac{1}{n} \sum_{m=1}^n \text{KL}(q(J_m|\mathbf{e}, \phi, S_m) \|\pi(J_m|\mathbf{e}, \phi))}. \end{aligned}$$

Then by optimizing  $\lambda$ , we have

$$\begin{aligned} \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} W_2^2(\mathcal{D}, \hat{\mu}) \\ \leq \mathbb{E}_S \mathbb{E}_{q(\mathbf{e}, \phi, \theta|S)} \frac{2}{n} \sum_{m=1}^n \mathbb{E}_{q(e_{(m)}|\mathbf{e}, \phi, S_m)} \|S_m - g_\theta(e_{(m)})\|^2 + 4\Delta \sqrt{2 \frac{1}{n} \sum_{m=1}^n \text{KL}(q(J_m|\mathbf{e}, \phi, S_m) \|\pi(J_m|\mathbf{e}, \phi))} + \frac{2\Delta}{\sqrt{n}}. \end{aligned}$$

1422

□



## G Experimental settings and additional experimental results

Our experiments were based on the Gaussian stochastically quantized VAE (SQ-VAE) model proposed by Takida et al. [63], and were conducted by adapting the code from their GitHub<sup>1</sup> to suit our experimental configurations. Therefore, we first introduce the basics of (Gaussian) SQ-VAE in Sections G.1 and G.2 and finally explain our experimental settings in Section G.3.

### G.1 Overview of SQ-VAE

The SQ-VAE is a generative model that, similar to VQ-VAE, employs a learnable codebook  $\mathbf{e} = \{e_k\}_{k=1}^K \in \mathcal{Z}^K$ . The objective of SQ-VAE is to learn the *stochastic decoder*  $x \sim p_\theta(x|Z_q)$  using latent variables  $Z_q$  to generate samples belonging to the data distribution  $p_{\text{data}}(x)$ , where  $p_\theta(x|Z_q) = \mathcal{N}(g_\theta(Z_q), \sigma^2 \mathbf{I})$ ,  $\mathcal{N}(m, \sigma \mathbf{I})$  is the Gaussian distribution with mean and equal variance parameter  $\{m, \sigma^2 \mathbf{I}\}$ ,  $\sigma^2 \in \mathbb{R}_+$ , and  $\mathbf{I}$  is the identity matrix. Here,  $Z_q$  is sampled from a prior distribution  $P(Z_q)$  over the discrete latent space  $\mathbf{e}^{dz}$ .

**In the main training process** of SQ-VAE, we assume  $P(Z_q)$  to be an i.i.d. uniform distribution, identical to VQ-VAE, meaning each codebook element is selected with equal probability ( $P(z_{q,i} = b_k) = 1/K$  for  $k \in [K]$ ). Subsequently, a **second training stage** is conducted to learn  $P(Z_q)$ . Since computing the posterior  $p_\theta(Z_q|x)$  exactly is intractable, we utilize an approximate posterior distribution  $q_\phi(Z_q|x)$  instead.

At the encoding process, directly mapping from  $x$  to the discrete  $Z_q$  is challenging due to the discrete nature of  $Z_q$ . To overcome this issue, Takida et al. [63] proposed to construct a stochastic encoder by introducing the following two processes:

- **Stochastic Dequantization Process:** The transformation function from  $Z_q$  to the auxiliary continuous variable,  $Z$ , denoted as  $p_\psi(Z|Z_q)$ , where  $\psi$  is its parameters.
- **Stochastic Quantization Process:** The transformation from  $Z$  to  $Z_q$  is given by  $\hat{P}_\phi(Z_q|Z) \propto p_\phi(Z|Z_q)P(Z_q)$  obtained via Bayes' theorem, which is represented as the categorical distribution  $q(J|\mathbf{e}, \phi, x)$  through the softmax function as in Eq. (2).

We can obtain  $\hat{Z}_q$  from a deterministic encoder  $f_\phi(x)$ , where we expect that  $\hat{Z}_q$  is close to  $Z_q$ . Therefore, we can similarly define the dequantization process of  $\hat{Z}_q$  as  $Z|\hat{Z}_q \sim p_\psi(Z|\hat{Z}_q)$ . By combining this process with the stochastic quantization process, we can establish the following *stochastic encoding* process from  $x$  to  $Z_q$ :  $\mathbb{E}_{q_\omega(Z|x)}[\hat{P}_\phi(Z_q|Z)]$ , where  $\omega := \{\phi, \psi\}$  and  $q_\omega(Z|x) := p_\psi(Z|f_\phi(x))$ .

According to these facts, we can derive the following evidence lower bound (ELBO) for SQ-VAE:

$$\begin{aligned} -\mathcal{L}_{\text{SQ}}(x; \theta, \omega, \mathbf{e}) \\ := \underbrace{\mathbb{E}_{q_\omega(Z|x), \hat{P}_\phi(Z_q|Z)} \left[ \log \frac{p_\theta(z|Z_q)p_\phi(Z|Z_q)}{q_\omega(Z|x)} \right]}_{=\text{KL}(\mathbf{Q} \parallel \mathbf{P})} + \mathbb{E}_{q_\omega(Z|x)} H(\hat{P}_\phi(Z_q|Z)) + (\text{Const.}), \end{aligned}$$

where  $H(\hat{P}_\phi(Z_q|Z))$  is the entropy of  $\hat{P}_\phi(Z_q|Z)$ .

From the above, the optimization problem of SQ-VAE is minimizing  $\mathbb{E}_{p_{\text{data}}(x)}[\mathcal{L}_{\text{SQ}}(x; \theta, \omega, \mathbf{e})]$  w.r.t.  $\{\theta, \omega, \mathbf{e}\}$ . This approach eliminates the need for heuristic techniques traditionally required, such as stop-gradient, exponential moving average (EMA), and codebook reset [79].

Moreover, the categorical posterior distribution  $\hat{P}_\phi(Z_q|Z) = q(J|\mathbf{e}, \phi, x)$  can be approximated using the Gumbel–Softmax relaxation [35, 45], where the Gumbel–Softmax function is defined as, for all  $k$  ( $1 \leq k \leq K$ ),

$$\frac{\exp(-\beta \|f_\phi(x) - e_k\|^2 + G_k)/\tau)}{\sum_{j=1}^K \exp(-\beta \|f_\phi(x) - e_j\|^2 + G_j)/\tau)},$$

<sup>1</sup><https://github.com/sony/sqvae/tree/main/vision>

Table 1: Experimental settings on MNIST.

Experimental setup for MNIST experiments	
Model	Gaussian stochastically quantized VAE (SQ-VAE) [63]
Network architecture	ConvResNets with three convolutional layers, two transpose convolutional layers, and one ResBlocks.
The size of a codebook ( $K$ ) and the dimension of the latent space $d_z$	$K = \{16, 32, 64, 128\}$ ; $d_z = 64$
Optimizer	Adam with 0.001 initial learning rate
Batch size	32
Num. of training/validation samples	[250, 1000, 2000, 4000]
Num. of epochs	200
Num. of samples for CMI estimation	3
Num. of samplings for $U$	5

Table 2: Experimental settings on CIFAR10.

Experimental setup for CIFAR10 experiments	
Model	Gaussian stochastically quantized VAE (SQ-VAE) [63]
Network architecture	ConvResNets with three convolutional layers, two transpose convolutional layers, and one ResBlocks.
The size of a codebook ( $K$ ) and the dimension of the latent space $d_z$	$K = \{16, 32, 64, 128\}$ ; $d_z = 64$
Optimizer	Adam with 0.001 initial learning rate
Batch size	32
Num. of training/validation samples	[1000, 5000, 10000, 20000]
Num. of epochs	200
Num. of samples for CMI estimation	3
Num. of samplings for $U$	5

where  $G_k$  is an i.i.d. sample from the Gumbel distribution and  $\tau$  is the temperature parameter that is deferent from  $\beta$  in Eq. (2). This allows the application of the reparameterization trick from VAEs during backpropagation, enabling efficient gradient computation and model training.

## G.2 Gaussian SQ-VAE

Gaussian SQ-VAE assumes that the dequantization process  $p_\psi(Z|Z_q)$  follows a Gaussian distribution. In this paper, we set the following Gaussian distribution:  $p_\psi(Z_i|Z_q) = \mathcal{N}(Z_{q,i}, \sigma_\psi^2 \mathbf{I})$ , where  $\sigma_\psi^2 \in \mathbb{R}_+$ . Then, the stochastic decoder and the stochastic dequantization process in SQ-VAE can be written as  $p_\theta(x|Z_q) = \mathcal{N}(g_\theta(Z_q), \sigma^2 \mathbf{I})$  and  $p_\psi(Z_i|\hat{Z}_q) = \mathcal{N}(\hat{Z}_{q,i}, \sigma_\psi^2 \mathbf{I})$ .

## G.3 Details of experimental settings

**Dataset:** We used the MNIST dataset [41], which is  $28 \times 28$  gray scale images with 10 classes. We prepared the subset dataset with  $\{1000, 2000, 4000, 8000\}$  samples from the default training dataset (60000 samples). Then, we split it as the training and the validation datasets following the supsample setting as in Section 2.3.

**Model architecture and training procedure:** We adopted the ConvResNets with the architecture provided by Google DeepMind<sup>2</sup>. We summarize the details of this model in Table 1.

Regarding the training procedure, we adopted the settings in Takida et al. [63] as follows. We used the Adam optimizer with 0.001 initial learning rate. The learning rate was halved every 3 epochs if the validation loss is not improving. We trained the model 200 epochs with 32 mini-batch size. As for the annealing schedule for the temperature parameter of the Gumbel-softmax sampling, we set  $\tau = \exp(10^{-5} \cdot t)$  as in Jang et al. [35], where  $t$  is the global training step size.

**GPU environment:** We used NVIDIA GPUs with 32GB memory (NVIDIA DGX-1 with Tesla V100 and DGX-2) in our experiments.

**Mutual information estimation:** To estimate the mutual information  $I(\tilde{\mathbf{J}}; U | \mathbf{e}, \phi, \tilde{X})$  in Eq. (7), we developed a plug-in estimator for it, which is computed using estimators for the probability density of  $\tilde{\mathbf{J}}$  and  $\tilde{X}$ , as well as their joint probability density, employing  $k$ -nearest-neighbor-based density estimation [43]. The estimation strategy is incorporated into the `sklearn.feature_selection.mutual_info_classif` function<sup>3</sup>. We set  $k = 3$  following the default setting of this function and Kraskov et al. [40], Ross [53].

<sup>2</sup>[https://github.com/deepmind/sonnet/blob/v2/examples/vqvae\\_example.ipynb](https://github.com/deepmind/sonnet/blob/v2/examples/vqvae_example.ipynb)

<sup>3</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.mutual\\_info\\_classif.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html)

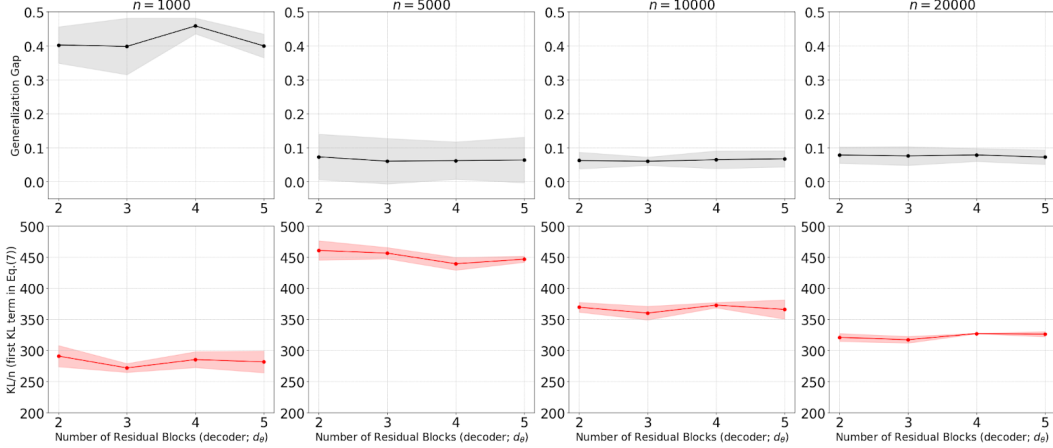


Figure 4: The behavior of the generalization gap and the empirical KL term ( $\text{KL}(\mathbf{Q}_{\mathbf{J},U} \parallel \mathbf{P})/n$ ) on the CIFAR-10 dataset when increasing the number of residual blocks to enlarge the decoder dimension  $d_\theta$  ( $K = 128$ ,  $d_z = 64$ ).

#### G.4 Additional experimental results

Here, we summarize our additional experimental results. We first show the results on CIFAR-10 in Figure 4. This results also support our implication even on the more complex dataset than MNIST that increasing the complexity of  $g_\theta$  has a limited effect on the generalization performance because adding a single ResBlock—introducing approximately 74, 000 additional parameters—has a negligible effect on the generalization gap.

We further numerically evaluated the bound in Theorem 4 using estimation techniques from prior work in IT analysis [30] under the MNIST setting.

As shown in Figure 5, the first KL term does not decrease monotonically with  $n$ , consistent with Lemmas 2 and 3, due to the use of a data-independent prior. In contrast, the second KL term decreases steadily with  $n$ , supporting our claim (in Section 3.2) that the data-dependent prior effectively captures generalization behavior in VQ-VAE. We also varied the latent dimension  $K$  under fixed  $n$ , and observed that both KL terms increased with larger  $K$ , further confirming our theoretical predictions.

Furthermore, we conducted additional numerical experiments to evaluate how the complexity of the decoder network ( $g_\theta$ ) and the dimension of the latent space ( $d_z$ ) influence the generalization gap, using SQ-VAE, which utilizes a stochastic mechanism, and VQ-VAE, which employs a deterministic mechanism. The results are shown in Figure 6.

For SQ-VAE experiments, the setups are almost identical to those of Table 1, except that we used a batch size of 64 and set  $K = 256$ . The size of the training data is specified in the figure. In the most left panel, we fixed  $d_z$ , increased the number of decoder ResBlocks across four commonly used datasets, and evaluated how generalization gaps are affected. We observed that the number of decoder ResBlocks does not significantly impact the generalization gaps. In the second panel from the left, we fixed the number of ResBlocks and increased the latent dimension  $d_z$ . We found a tendency for the generalization gap to increase as  $d_z$  increases. These results are consistent with Theorem 4, which suggests that when using a stochastic mechanism for LV, then the upper bound of the generalization gap is independent of the decoder complexity, while it depends on  $d_z$ .

In addition to SQ-VAE, we conducted a similar experiment using VQ-VAE, which uses a deterministic mechanism, following the settings of the image experiments in Van Den Oord et al. [68]. The differences between our experimental setting and Van Den Oord et al. [68] are that we used  $K = 258$  and the training epoch is set to 20.

When using the deterministic mechanism shown in the rightmost panel, we observed that increasing  $d_z$  tends to increase the generalization gap, consistent with the experimental findings under the SQ-VAE setting. The effect of the decoder is shown in the second panel from the right. Our

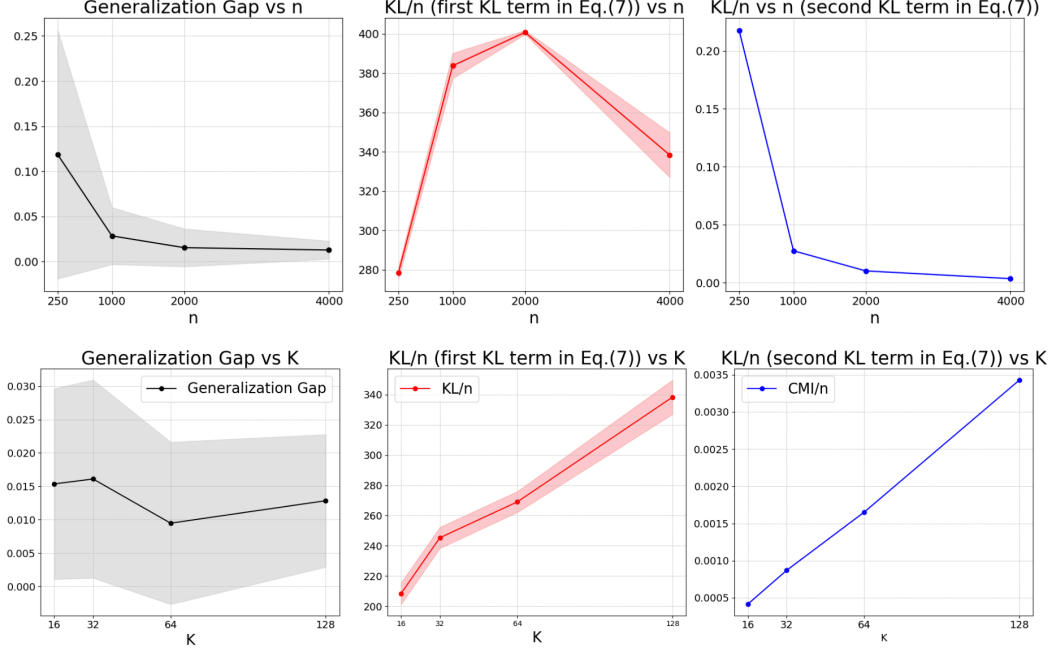


Figure 5: The behavior of the generalization gap and the empirical KL term ( $KL(\mathbf{Q}_{J,U}||\mathbf{P})/n$ ) on the CIFAR-10 dataset when increasing the number of residual blocks to enlarge the decoder dimension  $d_\theta$  ( $K = 128, d_z = 64$ ).

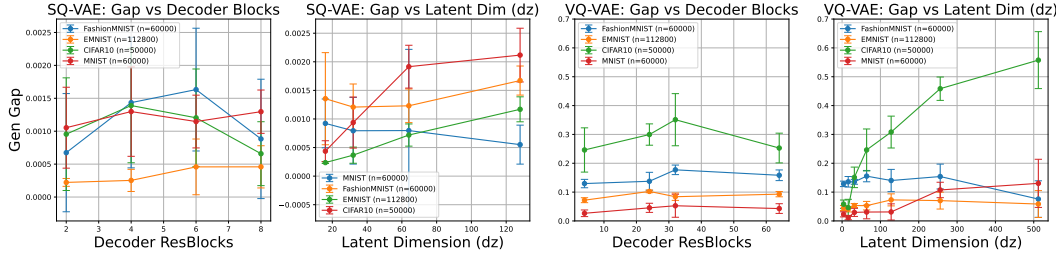


Figure 6: The behavior of the generalization gap when increasing the number of residual blocks of the decoder network and the latent dimension  $d_z$  in SQ-VAE and VQ-VAE models

numerical experiments confirm that the influence of decoder complexity is limited, as predicted by our theoretical results (Theorems 2 and 3). Specifically, we significantly increased the decoder size beyond the commonly used range, yet its impact on the generalization gap remained limited.

However, we also observed that increasing decoder complexity within a moderate range (e.g., from 2 to 6 ResBlocks) tends to increase the generalization gap. This does not contradict our theory. Indeed, Theorems 2 and 3 state that the upper bound on the generalization gap becomes independent of the decoder. This implies that, although increasing decoder complexity substantially does not further affect the generalization gap beyond a certain point, some increase within a moderate range can still worsen generalization as long as it remains under the upper bound. We believe that our numerical results correspond precisely to this regime.

We further conjecture that this behavior can be attributed to the fact that Theorems 2 and 3 do not depend solely on architectural properties. The upper bounds in these theorems depend on the learning algorithm  $q(w|S)$ , and increasing decoder complexity can influence the learned distributions  $e$  and  $\phi$ , which are defined as the marginal distributions under  $q(w|S)$ . This provides a plausible explanation

1536 for why decoder architecture may still affect the generalization gap in the moderate-complexity  
1537 regime.

1538 In contrast, the experimental results for SQ-VAE are explained by Theorem 4, which is independent  
1539 of the learning algorithm  $q(w|S)$  and fully eliminates the influence of the decoder. This explains the  
1540 consistent lack of decoder impact observed in the SQ-VAE experiments. Overall, the experimental  
1541 findings from both SQ-VAE and VQ-VAE suggest that the influence of decoder complexity depends  
1542 on whether the latent variable mechanism is stochastic or deterministic. A more refined theoretical  
1543 analysis of how stochastic mechanisms affect the generalization gap is an important direction for  
1544 future work.