Latent Space Factorization in LoRA

Shashi Kumar[1,2] **Yacouba Kaloga**[1] **John Mitros**[1] **Petr Motlicek**[1,3] **Ina Kodrasi**[1]
[1]Idiap Research Institute, Switzerland
[2]EPFL, Switzerland [3]BUT, Czech Republic
`{shashi.kumar, yacouba.kaloga, petr.motlicek, ina.kodrasi}@idiap.ch`
`john.mitross@gmail.com`

Abstract

Low-rank adaptation (LoRA) is a widely used method for parameter-efficient finetuning. However, existing LoRA variants lack mechanisms to explicitly disambiguate task-relevant information within the learned low-rank subspace, potentially limiting downstream performance. We propose Factorized Variational Autoencoder LoRA (FVAE-LoRA), which leverages a VAE to learn two distinct latent spaces. Our novel Evidence Lower Bound formulation explicitly promotes factorization between the latent spaces, dedicating one latent space to task-salient features and the other to residual information. Extensive experiments on text, audio, and image tasks demonstrate that FVAE-LoRA consistently outperforms standard LoRA. Moreover, spurious correlation evaluations confirm that FVAE-LoRA better isolates task-relevant signals, leading to improved robustness under distribution shifts. Our code is publicly available at: `https://github.com/idiap/FVAE-LoRA`

1 Introduction

Foundation models have become ubiquitous across modalities such as vision [1, 2, 3, 4], audio [5, 6], and text [7, 8]. Recent state-of-the-art results are predominantly achieved by fine-tuning these large pre-trained models. Among various parameter-efficient fine-tuning (PEFT) strategies [9, 10, 11, 12], *Low-Rank Adaptation (LoRA)* [12] has emerged as a particularly efficient approach. In LoRA, the original weight matrices $\mathbf{W} \in \mathbb{R}^{k\times d}$ are kept frozen, and trainable low-rank matrices $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{k\times r}$ are introduced, with $r \ll min(d,k)$, such that the adapted weights become $\mathbf{W} + \mathbf{BA}$. This technique significantly reduces memory and computational requirements, while achieving performance comparable to full fine-tuning [12, 13].

Despite the remarkable performance shown by LoRA across a plethora of downstream tasks and modalities, we identify a potential limitation: the standard LoRA update mechanism lacks an explicit way to ensure that the learned low-rank subspace $\mathrm{Im}(\mathbf{A})$ primarily captures task-salient information. The projection $\mathbf{Ax}$ (where $\mathbf{x}$ is the input activation) is learned implicitly through gradient descent on the task objective. While effective, this process does not inherently guarantee that $\mathbf{A}$ isolates features crucial for the downstream task from potentially irrelevant or even detrimental information retained from pre-training. This lack of explicit control over the content of the low-rank update is pertinent. While the hypothesis that fine-tuning primarily involves low-rank updates provides a strong theoretical underpinning for LoRA [14], empirical evidence suggests nuances. Recent studies have shown that standard LoRA can still underperform full fine-tuning in certain scenarios [12, 15]. This suggests that simply constraining the update to be low-rank might not be sufficient; the task-relevant signal encoded within that low-rank adaptation is critical for achieving optimal downstream performance. Existing LoRA variants do not offer a principled mechanism to explicitly disentangle and prioritize task-relevant information within the learned update.

To address this limitation and enable explicit control over the information captured within the low-rank update, we propose **Factorized Variational Autoencoder LoRA (FVAE-LoRA)**. Our approach

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

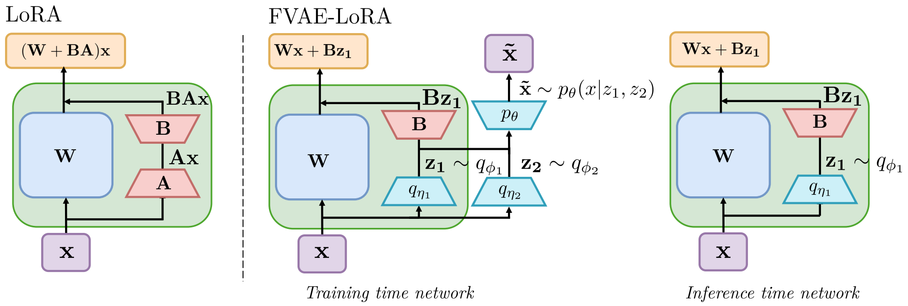

Figure 1: Comparison between LoRA and the proposed FVAE-LoRA. During training, FVAE-LoRA factorizes the latent space into two components, $\mathbf{z}_1$ and $\mathbf{z}_2$, where only the task-salient latent factor $\mathbf{z}_1$ is propagated downstream. At inference, only the encoder corresponding to $\mathbf{z}_1$ is required.

integrates a Variational Autoencoder (VAE) framework directly into the LoRA parameterization. Crucially, unlike standard VAEs, FVAE learns **two distinct latent spaces**, denoted by $\mathbf{z}_1$ and $\mathbf{z}_2$ (see Figure 1). We introduce a novel Evidence Lower Bound (ELBO) formulation specifically designed to promote **factorization** between these two spaces during training. This objective encourages the model to encode task-salient information, critical for downstream performance, primarily within the first latent space $\mathbf{z}_1$, while relegating residual information necessary for accurate reconstruction (as required by the FVAE objective) to the second latent space $\mathbf{z}_2$. During the forward pass for the downstream task, only samples drawn from the task-salient latent space $\mathbf{z}_1$ are utilized to generate the effective low-rank adaptation matrix $\mathbf{A}$. This mechanism allows FVAE-LoRA to explicitly select and leverage the most relevant learned features for the target task, while isolating potentially less useful or confounding variations within $\mathbf{z}_2$.

Our main contributions can be summarized as follows:

- **A Novel PEFT Method (FVAE-LoRA):** We propose FVAE-LoRA, integrating a VAE with factorized latent spaces ($\mathbf{z}_1$, $\mathbf{z}_2$) into the LoRA framework to explicitly disentangle task-salient information ($\mathbf{z}_1$) from residual information ($\mathbf{z}_2$).
- **Factorizing ELBO Formulation:** We introduce a novel ELBO objective specifically designed to enforce this factorization between the two latent spaces during training.
- **Strong Empirical Performance:** We demonstrate through extensive experiments on diverse image, text, and audio benchmarks that FVAE-LoRA consistently outperforms LoRA.
- **Empirical Validation of Robustness:** We empirically validate, using targeted spurious-correlation experiments, that the task-salient latent space $\mathbf{z}_1$ indeed captures task-critical information, leading to a robust performance even on challenging examples designed to mislead standard LoRA.

2 Related Work

We position FVAE-LoRA relative to PEFT methods, specifically LoRA variants, and techniques for latent space factorization in VAEs.

PEFT methods adapt large pre-trained models with minimal trainable parameters, overcoming the costs of full fine-tuning. Common approaches include inserting Adapter modules [9], optimizing continuous prompts or prefixes [10, 16], or tuning only bias terms [17]. LoRA [12] is a prominent PEFT technique that injects trainable low-rank matrices ($\Delta\mathbf{W} = \mathbf{BA}$, rank $r \ll min(d, k)$) into the model layers. Its efficiency and performance have led to wide adoption [14]. FVAE-LoRA builds upon the LoRA framework, aiming to enhance its effectiveness.

LoRA Variations. Several methods have extended LoRA. AdaLoRA [18] adaptively allocates rank budgets. DoRA [15] decouples weight magnitude and direction, applying LoRA to the latter. LoRA+ [19] adjusts LoRA's optimization by using different learning rates for its two low-rank matrices. PiSSA [20] focuses on initializing the LoRA matrices in a way that better approximates full fine-tuning updates, typically setting $\mathbf{A} = \mathbf{0}$ and initializing $\mathbf{B}$ based on principal components of the gradient. RS-LoRA [21] aims to stabilize training and prevent rank collapse or excessive growth by incorporating regularization related to the singular values of the update matrix. Other contributions combine LoRA with quantization for further compression [22, 23]. These variants primarily modify the update's structure, optimization, or compression. Our work differs by focusing on the *semantic content* of the update. FVAE-LoRA introduces a VAE with factorized latent spaces ($\mathbf{z}_1, \mathbf{z}_2$) and a novel ELBO to explicitly separate task-salient information ($\mathbf{z}_1$) used for the update from residual information ($\mathbf{z}_2$), thereby controlling the information encoded in the low-rank adaptation.

Factorization and Disentanglement in VAEs. VAEs [24] learn latent representations by maximizing the ELBO. Significant research focuses on learning *disentangled* representations, where latent dimensions capture independent factors of data variation [25]. Techniques often modify the ELBO, such as β-VAE [26], FHVAE [27], FactorVAE [28], TCVAE [29], and DIP-VAE [30], or employ annealing strategies [31]. While we use a VAE and aim for factorization, our goal is distinct. We do not seek to disentangle underlying data factors. Instead, FVAE-LoRA employs a novel ELBO to factorize the latent space specifically for PEFT: separating information crucial for the *downstream task* ($\mathbf{z}_1$) from other sources of variation needed for reconstruction ($\mathbf{z}_2$). This task-conditional factorization within the LoRA update mechanism represents the core novelty of our VAE application.

3 Method

In this section, we present our low-rank adaptation approach based on a VAE. We begin by briefly reviewing the standard VAE framework, and then introduce our proposed Factorized VAE (FVAE). We highlight key properties of this formulation and conclude by describing how it enables the construction of an efficient low-rank adaptation model.

3.1 Variational Autoencoder Objective

Consider a dataset $X \in \mathbb{R}^{n \times d}$, where observations are generated according to the process $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$, with $\mathbf{z}$ a latent variable. The goal is to recover a latent representation $\mathbf{z}$ that explains the observations $\mathbf{x}$, which requires computing the posterior $p_\theta(\mathbf{z}|\mathbf{x})$. As this posterior is generally intractable, we introduce an approximate distribution $q_\phi(\mathbf{z}|\mathbf{x})$. It can be shown (see Appendix A.1) that the log-likelihood admits a lower bound, known as the *Evidence Lower Bound (ELBO)*:

$$\mathcal{L}_{\theta,\phi}^{\text{VAE}}(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right). \tag{1}$$

The ELBO is used as a tractable surrogate objective for maximizing $\log p_\theta(\mathbf{x})$. The first term encourages accurate reconstruction, while the second term regularizes the latent space by aligning the approximate posterior with the prior $p(\mathbf{z})$.

3.2 Factorized Variational Autoencoder Objective

The primary goal of FVAE is to factorize the information contained in $\mathbf{x}$ such that it is represented by two independent latent variables $\mathbf{z}_1$ and $\mathbf{z}_2$. This factorization is learned jointly with a downstream task loss applied specifically to $\mathbf{z}_1$, which guides the decomposition by encouraging $\mathbf{z}_1$ to capture task-relevant information, while $\mathbf{z}_2$ absorbs the remaining variability. Classical VAEs serve as a natural starting point to build such a model.

3.2.1 Preliminaries

The derivation of the classical VAE can be extended by assuming that $\mathbf{x}$ arises from a generative process involving two independent latent variables $\mathbf{z}_1$ and $\mathbf{z}_2$, with $p(\mathbf{z}_1, \mathbf{z}_2) = p_1(\mathbf{z}_1)\, p_2(\mathbf{z}_2)$. Additionally, we assume that the approximate posterior factorizes as $q_\phi(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x}) = q_{\phi_1}(\mathbf{z}_1|\mathbf{x})\, q_{\phi_2}(\mathbf{z}_2|\mathbf{x})$. Considering $\mathbf{z}_1 \sim q_{\phi_1}(\mathbf{z}_1|\mathbf{x})$ and $\mathbf{z}_2 \sim q_{\phi_2}(\mathbf{z}_2|\mathbf{x})$, the ELBO is given by

$$\mathcal{L}_{\theta,\phi}^{\text{VAE2LAT}}(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{z}_1,\mathbf{z}_2} \left[\log p_\theta(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2)\right] - D_{\text{KL}}\left(q_{\phi_1}(\mathbf{z}_1|\mathbf{x}) \,\|\, p_1(\mathbf{z}_1)\right) - D_{\text{KL}}\left(q_{\phi_2}(\mathbf{z}_2|\mathbf{x}) \,\|\, p_2(\mathbf{z}_2)\right). \tag{2}$$

This objective mirrors the standard VAE but extends it to the multi-latent setting. However, even though both the prior and the variational posterior are factorized, the model is not explicitly encouraged to selectively assign information to $\mathbf{z}_1$ or $\mathbf{z}_2$.

3.2.2 FVAE

To promote factorization, we introduce a regularization term that penalizes the similarity between $q_{\phi_2}(\mathbf{z}_2|\mathbf{x})$ and the uninformative prior $p_1(\mathbf{z}_1)$. Since $q_{\phi_1}(\mathbf{z}_1|\mathbf{x})$ is encouraged to align with p_1, this term prevents q_{ϕ_2} from encoding information in the same region of the latent space. Incorporating this into Equation (2), we obtain the objective

$$\max_{\theta,\phi_1,\phi_2} \; \mathcal{L}^{\text{VAE2LAT}}_{\theta,\phi}(\mathbf{x}) + \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}\left[\log \frac{q_{\phi_2}(\mathbf{z}_2|\mathbf{x})}{p_1(\mathbf{z}_1)}\right]. \tag{3}$$

To clarify the role of each component in Equation (3) and relate them to familiar VAE structures, we reorganize the objective using straightforward algebraic manipulations. In doing so, we isolate the standard reconstruction and KL divergence terms, and separate out the new cross-prior regularizer. Introducing scalar constants α, β and δ allows us to balance the influence of these components, yielding the structured objective

$$\mathcal{L}^{\text{FVAE}}_{\theta,\phi}(\mathbf{x}) = \alpha \mathop{\mathbb{E}}_{\mathbf{z}_1,\mathbf{z}_2}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2)\right] - \beta D_{\text{KL}}\left(q_{\phi_1}(\mathbf{z}_1|\mathbf{x}) \,\|\, p_1(\mathbf{z}_1)\right) + \delta \underbrace{\mathop{\mathbb{E}}_{\mathbf{z}_2,\mathbf{z}_1} \log \frac{p_2(\mathbf{z}_2)}{p_1(\mathbf{z}_1)}}_{\Gamma}. \tag{4}$$

The second term correspond to the D_{KL} in the β-VAE objective, ensuring that the main latent variable $\mathbf{z}_1$ captures the relevant information for reconstructing $\mathbf{x}$ while remaining close to its prior. The third term, Γ, acts as a repulsive regularizer, encouraging the second component $\mathbf{z}_2$ to decouple from $\mathbf{z}_1$. Note that, a priori, we could fix $\alpha = 1$ and use only β and δ to weight the contributions of all the terms. However, we prefer to use all three, as it will make the interpretation of each contribution clearer later on.

3.3 Mechanism of the Γ modulator

Γ introduces an indirect interaction between the two encoders by modulating their alignment with their respective priors. Rather than enforcing separation through a direct divergence between posteriors, it shifts their latent support via prior-based regularization. To analyze the effect of Γ, we first rewrite it as the sum of a mismatch term and a discrepancy term, i.e.,

$$\Gamma = \underbrace{\mathbb{E}_{\mathbf{z}_2 \sim q_{\phi_2}}\left[\log p_2 - \log p_1\right]}_{\text{mismatch: } \Lambda} + \underbrace{\left[\mathbb{E}_{\mathbf{z}_2 \sim q_{\phi_2}} \log p_1 - \mathbb{E}_{\mathbf{z}_1 \sim q_{\phi_1}} \log p_1\right]}_{\text{discrepancy: } \Delta}, \tag{5}$$

where the mismatch term can be further equivalently expressed as a difference of KLs, i.e., $\Lambda = D_{KL}(q_{\phi_2}||p_1) - D_{KL}(q_{\phi_2}||p_2)$. This decomposition reveals a meaningful structure, as outlined in the following.

Maximizing the mismatch encourages q_{ϕ_2} to align with its prior p_2. This mirrors the behavior expected in a two-variable standard VAE, where each encoder is regularized toward its respective prior. As a result, we retain effective control over the behavior of q_{ϕ_2}, providing a structural safeguard against degenerate or unconstrained posterior collapse. In contrast, it disincentivizes q_{ϕ_2} from aligning with the prior p_1. Consequently, q_{ϕ_2} is encouraged to preserve or discover features and structures that are distinct from, and not merely reflections of, the assumptions embedded within p_1. The mismatch term also highlights that the two priors p_1 and p_2 should be different, but still partially overlapping. If they are identical, some terms will simply cancel out, and if they are too far apart, the separation becomes trivial, resulting in no fruitful competition for occupying the latent space. Since the priors are usually Gaussian with variance 1, this competition is parameterized by $|\mu_1 - \mu_2|$; the effect of the mismatch is null when this parameter is null.

In addition, we can demonstrate (see Appendix B.1) that the discrepancy Δ is bounded by a term depending on the 2-Wasserstein distance, provided that the Hessian $\|\nabla^2 \log p_1\| \leq L$ is bounded. In practice, p_1 is typically a standard normal $\mathcal{N}(0, I)$, and q_{ϕ_1} is a diagonal Gaussian. Under these assumptions, the bound becomes:

$$\Delta \leq \frac{L}{2}\mathcal{W}_2^2(q_{\phi_1}, q_{\phi_2}) + \sqrt{\sum_j \mu_j^2 + \sigma_j^2} \cdot \mathcal{W}_2(q_{\phi_1}, q_{\phi_2}),$$

where μ_j and σ_j^2 are the parameters of q_{ϕ_1}. Since q_{ϕ_1} is typically optimized to approximate p_1, the square-root term remains bounded in most settings. Both terms in the bound grow with $\mathcal{W}_2(q_{\phi_1}, q_{\phi_2})$, making Δ an effective surrogate for inducing Wasserstein repulsion. In particular, maximizing Δ increases $\mathcal{W}_2(q_{\phi_1}, q_{\phi_2})$, driving the two encoders apart in a geometrically meaningful way.

3.4 FVAE-LoRA

Building upon the FVAE framework, we leverage its ability to split the latent space to gain finer control over the representation, ultimately achieving better performance. To accomplish this, we proceed as illustrated in Figure 1.

For each targeted linear layer, we train an FVAE simultaneously with the downstream task, aiming to replace the $\mathbf{A}$ matrices used in classical LoRA (see the left side of the figure). During training, the input to the target layer is fed into the FVAE to compute a reconstruction loss based on that input. In parallel, the latent embedding $\mathbf{z}_1$ produced by the encoder q_{ϕ_1} is passed through a learned matrix $\mathbf{B}$ and added to the output of the frozen base weights $\mathbf{W}$. This yields the output $\mathbf{Wx} + \mathbf{Bz}_1$. At inference time, only q_{ϕ_1} is used to produce the output, either by sampling from it or by taking the mean of the distribution. Note that while we propose using FVAE with LoRA, the method is generic in the sense that it can be applied to give latent space control to any explicit LoRA method. In summary, the loss to be optimized in the proposed FVAE-LoRA approach is given by

$$\min_{\phi,\theta} \mathcal{L}_{\text{downstream-task}} - \boldsymbol{\lambda} \sum_{l \in \text{layer}} \mathcal{L}_{\theta,\phi}^{\text{FVAE}}(\mathbf{x}_l), \tag{6}$$

with $\boldsymbol{\lambda}$ being the hyper-parameter vector of weights assigned to the FVAE loss in each layer.

In practice: Both q_{ϕ_1} and q_{ϕ_2} are parameterized as diagonal Gaussian distributions, with their means and variances learned by neural networks. The reconstruction term $p_\theta(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2)$ is also parameterized by a neural network. The prior $p_1 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard normal distribution, while p_2 is empirically chosen to be centered at $\mathbf{1.5}$. The intuition is to give the two priors distinct non-overlapping "location" in the latent space to initialize and encourage separation. By setting μ_1 at 0 and μ_2 at 1.5, we provide a clear signal for the repulsive regularizer to push the posteriors apart. See additional insights in Appendices E and F.

4 Experimental Results

Motivation. The objective of the experimental evaluation is two fold. First, we aim to comprehensively evaluate FVAE-LoRA by comparing its performance against standard LoRA and its relevant variants across diverse image, text, and audio tasks. The specific selection of relevant variants for each domain is detailed within the respective modality subsections, guided by the aim to provide the most insightful and relevant benchmarks for each specific context. Second, we seek to empirically validate that FVAE-LoRA learns more robust representations by preferentially encoding task-salient information in $\mathbf{z}_1$.

Overall Setup. To ensure fair comparisons of parameter efficiency for the core adaptation mechanism, the LoRA rank r is set to 16 for all LoRA-based methods throughout our experiments. This rank also corresponds to the dimensionality of the task-salient latent space $\mathbf{z}_1$ in FVAE-LoRA. All LoRA-based baselines as well as FVAE-LoRA are applied to the query and key matrices within the transformer models. Detailed hyperparameter settings for FVAE-LoRA, including the balancing coefficients α, β and δ, learning rates, and specific VAE architectural choices for each task, are provided in Appendix D. We also provide a practical guide for selecting the key factorization hyperparameters, β and δ, in Appendix G.

4.1 Efficacy of FVAE-LoRA for Various Downstream Tasks

4.1.1 Image Tasks

Datasets. We evaluate FVAE-LoRA on six diverse image classification datasets: DTD [32], EuroSAT [33], GTSRB [34], RESISC45 [35], SUN397 [36], and SVHN [37]. These datasets span various image types, domains, and complexities.

Implementation Details. The pre-trained Vision Transformer (ViT-B/16) [3] serves as the backbone model for all image classification tasks. We compare FVAE-LoRA against full fine-tuning (Full FT) and several LoRA variants, i.e., standard LoRA [12], PiSSA [20], rsLoRA [21], DoRA [15], and OLoRA [38]. This broad selection of LoRA variants represents established PEFT methods for image classification. The evaluation metric is top-1 accuracy. Detailed hyperparameters can be found in D.1.

Table 1: Fine-tuning results of ViT-B/16 on image classification tasks. We fine-tune ViT-B/16 using full fine-tuning and LoRA variants across DTD, EuroSAT, GTSRB, RESISC45, SUN397, and SVHN datasets. **Bold** indicates the highest performance, while underline represents the second-highest performance.

Method	**Params (%)**	**DTD**	**EuroSAT**	**GTSRB**	**RESISC45**	**SUN397**	**SVHN**	**Average**
Full FT	-	78.12±0.59	**98.30±0.47**	**98.85±0.14**	**94.35±0.54**	69.34±0.59	**97.34±0.03**	89.38
LoRA	0.7240	74.65±1.08	97.28±0.36	96.95±0.56	90.11±0.53	71.11±0.07	94.22±0.14	87.39
PiSSA	0.7240	74.22±1.69	97.33±0.31	96.95±1.28	89.82±0.37	69.09±0.19	94.83±0.73	87.04
rsLoRA	0.7240	72.23±1.00	97.48±0.21	96.63±0.83	88.04±0.30	67.69±0.32	93.81±0.77	85.98
DoRA	0.7451	75.74±1.91	97.28±0.80	97.27±0.44	91.72±1.17	71.53±0.25	96.41±0.72	88.32
OLoRA	0.7240	72.23±0.40	96.62±1.06	97.08±0.46	88.94±0.45	69.64±0.43	94.86±0.30	86.63
FVAE-LoRA	0.7311	**78.19±0.68**	97.78±0.15	97.98±0.56	93.57±0.22	**73.14±0.21**	96.55±0.05	**89.53**

Results. The effectiveness of FVAE-LoRA for image classification is shown in Table 1. FVAE-LoRA achieves an average accuracy of 89.53% across six diverse datasets, outperforming LoRA and surpassing variants such as DoRA, all within a comparable inference-time parameter budget.

Notably, FVAE-LoRA's average performance slightly surpasses that of full fine-tuning (89.38%). This result suggests that the structured latent factorization inherent to FVAE-LoRA can guide the model towards learning highly effective adaptations. By explicitly encouraging the disentanglement of task-salient information within $\mathbf{z}_1$, FVAE-LoRA might be more adept at focusing the ViT backbone on critical visual features for the downstream task, potentially mitigating the risk of overfitting to spurious correlations or less generalizable patterns that can sometimes affect full fine-tuning on these datasets. On challenging datasets such as DTD (characterized by fine-grained textures) and SUN397 (complex scenes), FVAE-LoRA particularly excels, achieving the highest scores and outperforming full fine-tuning. For instance, on SUN397, FVAE-LoRA demonstrates a clear advantage, indicative of its capacity to distill critical visual cues for complex recognition tasks. While full fine-tuning outperforms all LoRA variants on datasets like EuroSAT and GTSRB, FVAE-LoRA consistently stands as the leading or a highly competitive PEFT method, often closing the gap significantly (e.g., achieving 97.78% on EuroSAT, closely trailing Full FT's 98.30%).

The presented results show that FVAE-LoRA is able to learn highly effective low-rank updates through a principled approach to information selection.

4.1.2 Text Tasks

Datasets. For natural language tasks, we use two benchmark categories:

1. **Commonsense Reasoning**: Training is done on a predefined corpus [39][1] of query-answer pairs, and the evaluation set includes seven sub-tasks: PIQA [40] (physical commonsense), SIQA [41] (social interaction understanding), ARC-c and ARC-e [42] (science question answering), OBQA [43] (multi-hop reasoning over facts), HellaSwag [44] (commonsense natural language inference)), and WinoGrande [45] (fill-in-the-blank).
2. **GLUE Benchmark**: A subset of the GLUE [46] is used, comprising SST2 (sentiment analysis), CoLA (linguistic acceptability), QNLI (question-answering NLI), MRPC (paraphrase detection), RTE (textual entailment), STSB (semantic textual similarity), and WNLI (coreference resolution).

Implementation Details. We employ Llama-3-8B [8] for the commonsense reasoning tasks and roberta-base [47] for the GLUE benchmark tasks. For commonsense reasoning tasks, we compare against Prompt Tuning [11], P-Tuning [10, 48], standard LoRA, and HiRA [13]. For completeness,

[1] `https://github.com/AGI-Edgerunners/LLM-Adapters/tree/main/dataset`

we also present the performance of ChatGPT taken from [15]. Considering the computational cost of LLM fine-tuning, our LoRA-based comparisons focus on standard LoRA and HiRA, as HiRA has recently demonstrated strong performance, offering a relevant and challenging benchmark in this setting. For roberta-base on GLUE, comparisons are made against Full FT and standard LoRA. This allows for a direct assessment of FVAE-LoRA's parameter efficiency relative to the crucial full fine-tuning upper bound and the widely adopted LoRA baseline. Evaluation uses accuracy for commonsense tasks following [13] and standard GLUE metrics (Matthews Correlation for CoLA, Pearson Correlation for STSB, Accuracy for the rest). Detailed hyperparameters can be found in D.2.

Table 2: Accuracy comparison among various PEFT methods on commonsense reasoning datasets for Llama-3-8B. **Bold** indicates the best performance, while underline represents the second-best performance. ChatGPT performance values are taken from [15], whereas Prompt Tuning and P-Tuning from [13].

Model	Method	Params (%)	PIQA	SIQA	ARC-c	ARC-e	OBQA	HellaSwag	WinoGrande	Average
ChatGPT	-	-	85.40	68.50	79.90	89.80	74.80	78.50	66.10	77.57
Llama-3-8B	Prompt Tuning	0.0010	45.05	36.13	31.57	32.74	29.20	14.01	50.12	34.12
	P-Tuning	0.6240	11.64	8.19	7.42	8.63	9.60	1.77	37.65	12.13
	LoRA	0.0848	80.74	75.59	67.58	82.11	75.20	85.73	77.82	77.82
	HiRA	0.0848	88.63	80.40	**81.66**	**93.56**	**87.20**	94.48	85.87	87.40
	FVAE-LoRA	0.0850	**88.96**	**81.58**	81.06	92.72	86.20	**95.30**	**88.95**	**87.82**

Results on Commonsense Reasoning using Llama-3-8B model. Table 2 reports the performance of FVAE-LoRA and baselines across seven commonsense reasoning benchmarks using the LLaMA-3-8B model. Our approach achieves the highest average accuracy of 87.82%, outperforming both the strong HiRA baseline (87.40%) and LoRA (77.82%) under comparable inference-time parameter budgets. These results indicate that FVAE-LoRA's strategy of factorizing latent information is particularly beneficial for complex reasoning tasks in LLMs. By explicitly guiding the $\mathbf{z}_1$ latent space to capture task-salient semantic and contextual cues necessary for reasoning, FVAE-LoRA enables Llama-3-8B to make more accurate inferences.

Notably, FVAE-LoRA demonstrates strong individual performances on tasks like HellaSwag (95.30%) and WinoGrande (88.95%), which require nuanced understanding of everyday situations and disambiguation. This suggests that the information isolated in $\mathbf{z}_1$ is indeed critical for these types of reasoning, allowing the LLM to leverage its capabilities more effectively than with less structured adaptation techniques. The ability to improve upon already powerful models like Llama-3-8B with such parameter efficiency highlights the potential of FVAE-LoRA for targeted capability enhancement in large language models.

Table 3: Results of fine-tuning roberta-base using full fine-tuning and LoRA on a subset of the GLUE datasets. **Bold** indicates the best results, while underline represents the second-best results.

Method	Params (%)	SST2	CoLA	QNLI	MRPC	RTE	STSB	WNLI	Average
Full FT	-	**94.77±0.25**	**62.43±1.16**	**91.97±0.05**	89.40±0.70	79.53±1.31	**90.30±0.29**	56.30±0.00	80.67
LoRA	0.4710	93.97±0.47	59.60±1.02	91.87±0.19	88.73±0.61	77.87±0.74	88.90±0.45	57.73±2.03	79.81
FVAE-LoRA	0.4759	94.10±0.43	60.37±0.49	91.63±0.34	**89.53±0.97**	**79.90±0.85**	88.60±0.16	**64.33±2.38**	**81.21**

Results on GLUE benchmark. Table 3 presents the performance of FVAE-LoRA when adapting the roberta-base model on a subset of the GLUE benchmark. Our method achieves the highest average score (81.21), outperforming both full fine-tuning (80.67) and standard LoRA (79.81). Notably, FVAE-LoRA shows particular strength on tasks like MRPC and WNLI. This strong performance on roberta-base demonstrates that the benefits of FVAE-LoRA's explicit latent factorization are not confined to large-scale models like Llama-3-8B (as seen in commonsense reasoning tasks), but also translate effectively to smaller, yet widely utilized encoder models. The ability to enhance these more moderately-sized architectures suggests that FVAE-LoRA's principled approach to focusing adaptations via $\mathbf{z}_1$ on task-critical linguistic features is robust across different model scales.

4.1.3 Audio Tasks

Datasets. We conduct automatic speech recognition (ASR) on the TIMIT acoustic-phonetic corpus [49] for phoneme recognition.

Implementation Details. The pre-trained Wav2Vec2-Large model [5] serves as the backbone. Fine-tuning utilizes the Connectionist Temporal Classification loss [50]. We compare against Full FT and standard LoRA. Performance is measured by Phoneme Error Rate (PER). Detailed hyperparameters are provided in Appendix D.3.

Results. As shown in Table 4, FVAE-LoRA achieves a PER of 8.09 on TIMIT, outperforming standard LoRA and approaching the performance of full fine-tuning (7.48), demonstrating its effectiveness for ASR.

Table 4: Fine-tuning results of Wav2Vec2-Large on the TIMIT speech recognition task using CTC loss. **Bold** indicates the best PER (↓), underline the second-best.

Method	Params (%)	**TIMIT** PER (↓)
Full FT	-	**7.48**
LoRA	0.4961	9.38
FVAE-LoRA	0.4999	<u>8.09</u>

4.2 Probing Latent Factorization via Spurious Correlation

To empirically validate our hypothesis that FVAE-LoRA learns more robust representations by preferentially encoding task-salient information in $\mathbf{z}_1$, we conduct experiments using datasets with controlled spurious correlations. Spurious correlations occur when input features are statistically associated with target labels without a true causal link [51, 52, 53], potentially misleading models and hindering generalization, especially on out-of-distribution or minority-group data. Our aim is to assess whether FVAE-LoRA's disentanglement mechanism renders it more robust to such misleading cues compared to standard LoRA.

Experimental Design. We leverage datasets where spurious attributes (e.g., background scene) are intentionally correlated with the true class labels (e.g., object category) during training. For example, a “landbird" might predominantly appear against a "land" background, and a "waterbird" against "water". Effective factorization should enable the model to learn the true object category via $\mathbf{z}_1$, irrespective of the potentially misleading background. Figure 2 illustrates this concept, distinguishing between an input image ($\mathbf{x}$), its core features ($\mathbf{x}_{\text{core}}$), and its spurious features ($\mathbf{x}_{\text{spurious}}$).

Datasets. Following prior works [54, 51, 52, 53, 55, 56], we consider three standard benchmarks to introduce spurious correlations: *Waterbirds* [56], where bird type (landbird vs. waterbird) is correlated with background (land vs. water); *CelebA* [56], where a target attribute (e.g., blonde hair) might be correlated with another attribute (e.g., being female); and *Animals* [57], a larger-scale dataset derived from ImageNet [58] with four animal classes spuriously correlated with background types (e.g., waterbirds with water, small dogs with indoor scenes). These datasets are structured into groups based on combinations of true labels and spurious attributes, with varying majority-to-minority group ratios between training and test splits (details in Appendix C.1 and Table 7).

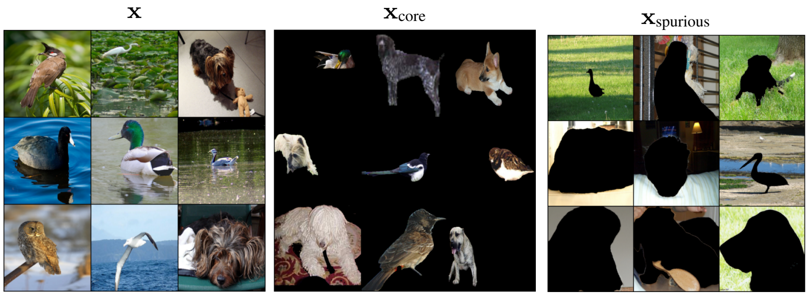

Figure 2: Random samples drawn from the train split of the Animals dataset, illustrating an original image ($\mathbf{x}$), its core object features ($\mathbf{x}_{\text{core}}$), and its spurious background features ($\mathbf{x}_{\text{spurious}}$).

Implementation Details and Evaluation Metrics. We adapt the ViT-B/16 [3] backbone using LoRA and our proposed FVAE-LoRA. Following common practice in literature [59, 60, 61], performance is evaluated using three key metrics:

- **Worst-Group Accuracy (WG):** Accuracy on the test subgroup where the model performs poorest, indicating robustness to spurious correlations and performance on minority groups.

- **Average Accuracy (AVG):** Standard overall accuracy on the test set.
- **Accuracy Disparity:** The absolute difference $|\text{WG} - \text{AVG}|$, quantifying the performance variation across groups. A smaller disparity suggests more uniform and equitable performance.

Table 5: Fine-tuning results of ViT-B/16 on spurious correlation benchmarks. We compare LoRA with FVAE-LoRA on ANIMALS (8 groups, 4 classes), WATERBIRDS (4 groups, 2 classes), and CELEBA (4 groups, 2 classes) datasets.

Method	**Params (%)**	**ANIMALS**		**WATERBIRDS**		**CELEBA**		**Disparity**
		WG	**AVG**	**WG**	**AVG**	**WG**	**AVG**	**| WG - AVG |**
LoRA	0.7240	54.79±8.08	88.20±1.17	75.49±0.9	90.39±0.78	40.00±3.54	**96.09±0.02**	34.8
FVAE-LoRA	0.7311	**62.0±4.83**	**89.55±0.96**	**75.85±3.72**	**90.99±0.51**	**43.33±6.68**	95.77±0.18	31.71

Results. Table 5 summarizes the performance of standard LoRA, and FVAE-LoRA on the spurious correlation benchmarks. Across all datasets, FVAE-LoRA consistently achieves higher WG and lower Accuracy Disparity compared to LoRA, while maintaining competitive AVG. These findings strongly suggest that FVAE-LoRA is less susceptible to being misled by spurious features present in the training data. We attribute this enhanced robustness to the explicit factorization encouraged by our novel ELBO. By compelling $\mathbf{z}_1$ to capture genuinely task-relevant, causal features and relegating other variations, FVAE-LoRA learns a more robust adaptation. This leads to improved generalization, particularly on minority groups where spurious cues are often unreliable or reversed, thereby validating the intended robust learning mechanism of our proposed method.

Table 6: Ablation study comparing our FVAE-LoRA when fine-tuning ViT-B/16 to a two-latent-variable VAE (VAE2LAT, as defined in Eq. (2)), and β-VAE2LAT, the β-VAE version of VAELAT (where all the DKL terms are multiplied by 10). Results are presented on DTD, EuroSAT, GTSRB, RESISC45, SUN397, and SVHN. **Bold** indicates the highest results, while underlined indicates the second-highest.

Method	**DTD**	**EuroSAT**	**GTSRB**	**RESISC45**	**SUN397**	**SVHN**	**Average**
FVAE-LoRA (Proposed)	**78.19±0.68**	**97.78±0.15**	**97.98±0.56**	**93.57±0.22**	**73.14±0.21**	**96.55±0.05**	**89.53**
β-VAE2LAT (8) (where $\beta = 10$)	77.16±0.43	96.86±0.15	95.75±0.46	89.46±0.13	72.91±0.32	91.58±0.69	87.29
VAE2LAT (3)	75.96±0.80	96.64±0.59	94.38±0.94	88.42±0.32	71.68±0.54	91.50±0.58	86.43

4.3 Ablation Studies

To demonstrate the relevance of introducing the regularization term in Equation ((3)), we replicate our image results using the two-variable VAE model (2) and its equivalent for β-VAE with two latent variables (where the two KL divergences have been multiplied by 10; see Equation (A.3)). The results can be seen in Table 6. The baseline model performs the worst across all datasets. The β-VAE with two latent variables shows some improvement, but it is still outperformed by our proposed method.

5 Conclusions

We introduced Factorized Variational Autoencoder LoRA (FVAE-LoRA), a novel PEFT method designed to explicitly disentangle task-salient information within the LoRA framework. By employing a VAE with two latent spaces, $\mathbf{z}_1$ (task-salient) and $\mathbf{z}_2$ (residual), and a specialized ELBO, FVAE-LoRA ensures that the adaptive updates are primarily driven by task-critical features learned in $\mathbf{z}_1$. Our comprehensive evaluations on diverse text, audio, and image benchmarks demonstrated that FVAE-LoRA consistently surpasses standard LoRA in performance. Crucially, experiments on datasets with spurious correlations empirically confirmed that FVAE-LoRA's factorization leads to more robust representations, as evidenced by improved worst-group accuracy. FVAE-LoRA highlights the potential of latent space factorization for enhancing parameter-efficient fine-tuning.

6 Acknowledgments

Shashi Kumar was partially supported by the EU Horizon 2020 project ELOQUENCE (grant number 101070558). Yacouba Kaloga was partially supported by Swiss National Science Foundation project no CRSII5_202228 on Characterisation of motor speech disorders and processes.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

[3] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. International Conference on Machine Learning*, pages 28492–28518, Honolulu, USA, July 2023.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

[10] X Liu, Y Zheng, Z Du, M Ding, Y Qian, Z Yang, and J Tang. Gpt understands, too. arxiv preprint arxiv: 210310385. 2021.

[11] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[12] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[13] Qiushi Huang, Tom Ko, Zhan Zhuang, Lilian Tang, and Yu Zhang. Hira: Parameter-efficient hadamard high-rank adaptation for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[15] Shih-yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *ICML*, 2024.

[16] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021.

[17] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022.

[18] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

[19] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models, 2024.

[20] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*, 2024.

[21] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

[22] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[23] Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. Qa-lora: Quantization-aware low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717*, 2023.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[25] F Locatello, S Bauer, M Lucic, G Rätsch, S Gelly, B Schölkopf, and O Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. arxiv preprint arxiv: 1811.12359, 2018.

[26] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[27] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in neural information processing systems*, 30, 2017.

[28] Hyunjik Kim and Andriy Mnih. Disentangling by factorising, 2019.

[29] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2019.

[30] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations, 2018.

[31] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018.

[32] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[33] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[34] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *IJCNN*, 2013.

[35] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[36] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NeurIPS workshop*, 2011.

[38] Kerim Büyükakyüz. Olora: Orthonormal low-rank adaptation of large language models. *arXiv preprint arXiv:2406.01775*, 2024.

[39] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*, 2023.

[40] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[41] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[42] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

[43] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[44] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[45] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[46] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[48] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[49] John S Garofolo, Lori F Lamel, William M Fisher, David S Pallett, Nancy L Dahlgren, Victor Zue, and Jonathan G Fiscus. Timit acoustic-phonetic continuous speech corpus. *(No Title)*, 1993.

[50] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. International Conference on Machine learning*, pages 369–376, Pittsburgh, USA, June 2006.

[51] Guanwen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Feature learning in the presence of spurious correlations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 41658–41697, 2024.

[52] Gautam Sreekumar and Vishnu Naresh Boddeti. Spurious correlations and where to find them. *arXiv preprint arXiv:2308.11043*, 2023.

[53] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020.

[54] Justin Cui, Ruochen Wang, Yuanhao Xiong, and Cho-Jui Hsieh. Ameliorate spurious correlations in dataset condensation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 9696–9721, 2024.

[55] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9109–9119. PMLR, 13–18 Jul 2020.

[56] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[57] Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Challenges and opportunities in improving worst-group generalization in presence of spurious features, 2025.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[59] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[60] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[61] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

A Variational Auto-Encoder

A.1 VAE Objective Derivation

We derive the Evidence Lower Bound (ELBO) by starting from the marginal log-likelihood:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right)\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log\frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} + \log\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\,\|\,p(\mathbf{z})\right) + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\,\|\,p_\theta(\mathbf{z}|\mathbf{x})\right).
\end{aligned}$$

The last term is always non-negative, which justifies interpreting the remaining two terms as a lower bound, i.e.,

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}^{\mathrm{VAE}}_{\theta,\phi}(\mathbf{x}),$$

with

$$\mathcal{L}^{\mathrm{VAE}}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\,\|\,p(\mathbf{z})\right).$$

A.2 VAE2LAT: VAE with 2 Latent Variables Objective Derivation

We simply start from the ELBO previously derived with two variables, i.e.,

$$\mathcal{L}^{\mathrm{VAE2LAT}}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2)\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x})\,\|\,p(\mathbf{z}_1,\mathbf{z}_2)\right).$$

Applying the independence assumption, we obtain

$$\mathcal{L}^{\mathrm{VAE2LAT}}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{\substack{\mathbf{z}_1\sim q_{\phi_1}(\mathbf{z}_1|\mathbf{x})\\ \mathbf{z}_2\sim q_{\phi_2}(\mathbf{z}_2|\mathbf{x})}}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2)\right] - D_{\mathrm{KL}}\left(q_{\phi_1}(\mathbf{z}_1|\mathbf{x})\,\|\,p_1(\mathbf{z}_1)\right) - D_{\mathrm{KL}}\left(q_{\phi_2}(\mathbf{z}_2|\mathbf{x})\,\|\,p_2(\mathbf{z}_2)\right). \tag{7}$$

A.3 β-VAE2LAT: A β-VAE with 2 latents variables

The loss of β-VAE2LAT, i.e., a straightforward extension of β-VAE to two latent variables is given by studies 4.3.

$$\mathcal{L}^{\boldsymbol{\beta}-\mathrm{VAE2LAT}}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{\substack{\mathbf{z}_1\sim q_{\phi_1}(\mathbf{z}_1|\mathbf{x})\\ \mathbf{z}_2\sim q_{\phi_2}(\mathbf{z}_2|\mathbf{x})}}\left[\log p_\theta(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2)\right] - \beta D_{\mathrm{KL}}\left(q_{\phi_1}(\mathbf{z}_1|\mathbf{x})\,\|\,p_1(\mathbf{z}_1)\right) - \beta D_{\mathrm{KL}}\left(q_{\phi_2}(\mathbf{z}_2|\mathbf{x})\,\|\,p_2(\mathbf{z}_2)\right). \tag{8}$$

This formulation is used in the ablation studies in Section 4.3.

B FVAE

B.1 Bounding the Discrepancy Term Δ via the 2-Wasserstein Distance

We bound the discrepancy

$$\Delta \;=\; \mathbb{E}_{\mathbf{z}\sim q_{\phi_2}}[\log p_1(\mathbf{z})] \;-\; \mathbb{E}_{\mathbf{z}\sim q_{\phi_1}}[\log p_1(\mathbf{z})]\,,$$

assuming only that the log-prior $f(z) = \log p_1(z)$ is C^2 with a globally bounded Hessian:

$$\|\nabla^2 f(z)\|_{\mathrm{op}} \;\leq\; L \quad \forall z \in \mathbb{R}^d. \tag{H}$$

Step 1 – Second-order Taylor control. For any two points z_1, z_2, Taylor's formula with integral remainder gives

$$f(z_2)-f(z_1) \;=\; \langle \nabla f(z_1),\, z_2-z_1\rangle + (z_2-z_1)^{\top}\bigg(\int_0^1 (1-s)\,\nabla^2 f\big(z_1+s\,(z_2-z_1)\big)ds\bigg)\,(z_2-z_1).$$

Bounding the remainder using assumption (H) gives the point-wise inequality

$$\big|f(z_2)-f(z_1)-\langle \nabla f(z_1), z_2-z_1\rangle\big| \;\le\; \frac{L}{2}\,\|z_2-z_1\|^2. \tag{A}$$

Step 2 – Integrate over a coupling. Let $\gamma \in \Pi(q_{\phi_1}, q_{\phi_2})$ be *any* coupling of the two distributions, and write $(\mathbf{z}_1, \mathbf{z}_2) \sim \gamma$, $d = \mathbf{z}_2 - \mathbf{z}_1$. Taking expectations in (A), applying the triangle inequality, and then Cauchy–Schwarz to the linear term,

$$\big|\Delta\big| \;\le\; \frac{L}{2}\,\mathbb{E}_\gamma\|d\|^2 \;+\; \underbrace{\big|\mathbb{E}_\gamma\langle \nabla f(\mathbf{z}_1), d\rangle\big|}_{\text{“linear term expectation”}} \;\le\; \frac{L}{2}\,\mathbb{E}_\gamma\|d\|^2 \;+\; \sqrt{\mathbb{E}_{q_{\phi_1}}\|\nabla f\|^2}\,\sqrt{\mathbb{E}_\gamma\|d\|^2}.$$

Now minimise the rightmost expression over γ. Since the function $g(x) = \frac{L}{2}x^2 + Cx$ (for $C = \sqrt{\mathbb{E}_{q_{\phi_1}}\|\nabla f\|^2} \ge 0$) is non-decreasing for $x = \sqrt{\mathbb{E}_\gamma\|d\|^2} \ge 0$, the infimum is attained when $\mathbb{E}_\gamma\|d\|^2$ is minimized. The infimum of $\mathbb{E}_\gamma\|d\|^2$ is $\mathcal{W}_2^2(q_{\phi_1}, q_{\phi_2})$. Hence :

$$\big|\Delta\big| \;\le\; \frac{L}{2}\,\mathcal{W}_2^2(q_{\phi_1}, q_{\phi_2}) \;+\; \sqrt{\mathbb{E}_{q_{\phi_1}}\big\|\nabla \log p_1(\mathbf{z})\big\|^2}\;\mathcal{W}_2(q_{\phi_1}, q_{\phi_2}).$$

Step 3 – Specialization to Gaussian case. Assume $p_1 = \mathcal{N}(0, I)$ and $q_{\phi_1} = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$. Then the gradient becomes $\nabla \log p_1(\mathbf{z}) = -\mathbf{z}$, and the expectation simplifies as:

$$\mathbb{E}_{q_{\phi_1}}\|\nabla \log p_1(\mathbf{z})\|^2 = \mathbb{E}_{q_{\phi_1}}\big[\|\mathbf{z}\|^2\big] = \sum_j \mu_j^2 + \sigma_j^2.$$

Therefore, the bound becomes:

$$\big|\Delta\big| \;\le\; \frac{1}{2}\,\mathcal{W}_2^2(q_{\phi_1}, q_{\phi_2}) \;+\; \sqrt{\sum_j \mu_j^2 + \sigma_j^2}\cdot \mathcal{W}_2(q_{\phi_1}, q_{\phi_2}).$$

Since q_{ϕ_1} is trained to approximate p_1, the square-root term is typically bounded in practice. Hence, both terms contribute to increasing $\mathcal{W}_2(q_{\phi_1}, q_{\phi_2})$, and the discrepancy Δ serves as an effective Wasserstein repulsion.

C Dataset Details

C.1 Spurious Correlation Experiments

Table 7: Statistics of the datasets used in the spurious experiment.

Dataset	SpuCoAnimals	Waterbirds	CelebA
# Classes	4	2	2
# Groups	8	4	4
Train	42000	4795	162770
Validation	2100	1199	19867
Test	4000	5794	19962
Class Ratio	25:25:25:25	76.8:23.2	85:15

D Hyperparameters

This section details the hyperparameters used for the experiments presented in the main paper. For all LoRA-based methods, including FVAE-LoRA, the LoRA rank (r) was set to 16, and LoRA was applied to the query and key matrices of the attention layers. The latent dimension of $\mathbf{z}_1$ in FVAE-LoRA corresponds to this LoRA rank.

D.1 Image Experiments

The following hyperparameters were used for fine-tuning ViT-B/16 on DTD, EuroSAT, GTSRB, RESISC45, SUN397, and SVHN datasets.

Table 8: Hyperparameters for Image Classification tasks using ViT-B/16.

Parameter	Value / Setting
General Training Parameters	
Optimizer	AdamW
Learning Rate	5×10^{-3}
LR Scheduler	Linear
Warmup Ratio	0.1
Batch Size	32
Number of Epochs	30
Weight Decay	0.01
Seeds	1, 2, 42
LoRA Parameters	
LoRA Rank (r)	16
LoRA Dropout	0.1
FVAE-LoRA Specific Parameters	
Latent Dim. $\mathbf{z}_1$	16 (same as LoRA rank)
Latent Dim. $\mathbf{z}_2$	16
FVAE $q_{\phi_i}(\mathbf{z}_i \vert \mathbf{x})$ Enc. Arch.	$\mathbf{x} \xrightarrow{\text{Linear}} \text{dim}(\mathbf{z}_i) \xrightarrow{\text{ReLU}} \text{HiddenState}_{\mathbf{z}_i} \xrightarrow{\text{Linear}} (\boldsymbol{\mu}_{\mathbf{z}_i}, \log \boldsymbol{\sigma}^2_{\mathbf{z}_i})$
FVAE $p_\theta(\mathbf{x} \vert \mathbf{z}_1, \mathbf{z}_2)$ Dec. Arch.	Concat($\mathbf{z}_1, \mathbf{z}_2$) $\xrightarrow{\text{Linear}} H_D = 128 \xrightarrow{\text{ReLU}}$ Linear $\rightarrow \hat{\mathbf{x}}$ (Input Dim)
Prior $p_1(\mathbf{z}_1)$	$\mathcal{N}(0, I)$
Prior $p_2(\mathbf{z}_2)$	$\mathcal{N}(1.5, I)$
$\boldsymbol{\lambda}$ (Eq. 6)	1×10^{-3}
ELBO Coeff. α (Reconstr.)	1
ELBO Coeff. β (KL $q_1 \vert\vert p_1$)	1 or 10
ELBO Coeff. δ	1

D.2 Text Experiments

Table 9: Hyperparameters for Commonsense Reasoning using Llama-3-8B.

Parameter	Value / Setting
General Training Parameters	
Optimizer	AdamW
Learning Rate	1×10^{-3} (LoRA, HiRA), 3×10^{-4} (FVAE-LoRA)
LR Scheduler	Linear
Warmup Steps	100
Batch Size	8
Gradient Accumulation Steps	4
Number of Epochs	3
Weight Decay	0.0
Seed	42
LoRA Parameters	
LoRA Rank (r)	16
LoRA Dropout	0.1
Target Modules	q_proj, k_proj
FVAE-LoRA Specific Parameters	
Latent Dim. $\mathbf{z}_1$	16
Latent Dim. $\mathbf{z}_2$	16
FVAE $q_{\phi_i}(\mathbf{z}_i\|\mathbf{x})$ Enc. Arch.	$\mathbf{x} \xrightarrow{\text{Linear}} \text{dim}(\mathbf{z}_i) \xrightarrow{\text{ReLU}} \text{HiddenState}_{\mathbf{z}_i} \xrightarrow{\text{Linear}} (\boldsymbol{\mu}_{\mathbf{z}_i}, \log \boldsymbol{\sigma}^2_{\mathbf{z}_i})$
FVAE $p_\theta(\mathbf{x}\|\mathbf{z}_1, \mathbf{z}_2)$ Dec. Arch.	$\text{Concat}(\mathbf{z}_1, \mathbf{z}_2) \xrightarrow{\text{Linear}} H_D = 128 \xrightarrow{\text{ReLU}} \text{Linear} \rightarrow \hat{\mathbf{x}}$ (Input Dim)
Prior $p_1(\mathbf{z}_1)$	$\mathcal{N}(0, I)$
Prior $p_2(\mathbf{z}_2)$	$\mathcal{N}(1.5, I)$
$\boldsymbol{\lambda}$ (Eq. 6)	1×10^{-4}
ELBO Coeff. α (Reconstr.)	1
ELBO Coeff. β (KL $q_1\|\|p_1$)	1 or 10
ELBO Coeff. δ	1

Table 10: Hyperparameters for GLUE Benchmark tasks using RoBERTa-base.

Parameter	Value / Setting
General Training Parameters	
Optimizer	AdamW
Learning Rate	3×10^{-4}
LR Scheduler	Linear
Warmup Ratio	0.06
Batch Size	32
Number of Epochs	30
Seed	1, 2, 42
LoRA Parameters	
LoRA Rank (r)	16
LoRA Dropout	0.1
FVAE-LoRA Specific Parameters	
Latent Dim. $\mathbf{z}_1$	16
Latent Dim. $\mathbf{z}_2$	16
FVAE $q_{\phi_i}(\mathbf{z}_i\|\mathbf{x})$ Enc. Arch.	$\mathbf{x} \xrightarrow{\text{Linear}} \text{dim}(\mathbf{z}_i) \xrightarrow{\text{ReLU}} \text{HiddenState}_{\mathbf{z}_i} \xrightarrow{\text{Linear}} (\boldsymbol{\mu}_{\mathbf{z}_i}, \log \boldsymbol{\sigma}^2_{\mathbf{z}_i})$
FVAE $p_\theta(\mathbf{x}\|\mathbf{z}_1, \mathbf{z}_2)$ Dec. Arch.	$\text{Concat}(\mathbf{z}_1, \mathbf{z}_2) \xrightarrow{\text{Linear}} H_D = 128 \xrightarrow{\text{ReLU}} \text{Linear} \rightarrow \hat{\mathbf{x}}$ (Input Dim)
Prior $p_1(\mathbf{z}_1)$	$\mathcal{N}(0, I)$
Prior $p_2(\mathbf{z}_2)$	$\mathcal{N}(1.5, I)$
$\boldsymbol{\lambda}$ (Eq. 6)	1×10^{-3} or 1×10^{-4}
ELBO Coeff. α (Reconstr.)	0.1 or 1
ELBO Coeff. β (KL $q_1\|\|p_1$)	1 or 10
ELBO Coeff. δ	1

D.3 Audio Experiments

The following hyperparameters were used for fine-tuning Wav2Vec2-Large on the TIMIT dataset.

Table 11: Hyperparameters for ASR on TIMIT using Wav2Vec2-Large.

Parameter	Value / Setting
General Training Parameters	
Optimizer	AdamW
Learning Rate	5×10^{-5} (Full FT), 5×10^{-4} (LoRA and FVAE-LoRA)
LR Scheduler	Linear
Warmup Steps	500
Batch Size	32
Number of Epochs	30
Weight Decay	0.005
CTC Loss Reduction	Sum
LoRA Parameters (Standard LoRA & FVAE-LoRA's LoRA part)	
LoRA Rank (r)	16
LoRA Dropout	0.1
FVAE-LoRA Specific Parameters	
Latent Dim. $\mathbf{z}_1$	16
Latent Dim. $\mathbf{z}_2$	16
FVAE $q_{\phi_i}(\mathbf{z}_i\|\mathbf{x})$ Enc. Arch.	$\mathbf{x} \xrightarrow{\text{Linear}} \text{dim}(\mathbf{z}_i) \xrightarrow{\text{ReLU}} \text{HiddenState}_{\mathbf{z}_i} \xrightarrow{\text{Linear}} (\boldsymbol{\mu}_{\mathbf{z}_i}, \log \boldsymbol{\sigma}^2_{\mathbf{z}_i})$
FVAE $p_\theta(\mathbf{x}\|\mathbf{z}_1, \mathbf{z}_2)$ Dec. Arch.	$\text{Concat}(\mathbf{z}_1, \mathbf{z}_2) \xrightarrow{\text{Linear}} H_D = 128 \xrightarrow{\text{ReLU}} \text{Linear} \rightarrow \hat{\mathbf{x}}$ (Input Dim)
Prior $p_1(\mathbf{z}_1)$	$\mathcal{N}(0, I)$
Prior $p_2(\mathbf{z}_2)$	$\mathcal{N}(1.5, I)$
$\boldsymbol{\lambda}$ (Eq. 6)	1×10^{-3}
ELBO Coeff. α (Reconstr.)	1
ELBO Coeff. β (KL $q_1\|\|p_1$)	1
ELBO Coeff. δ	1

D.4 Spurious Correlation Experiments

These experiments (Waterbirds, CelebA, Animals) used ViT-B/16 as the backbone. Base training and LoRA parameters are similar to those in Section D.1, with specific FVAE-LoRA coefficients tuned for robustness.

Table 12: Key FVAE-LoRA Hyperparameters for Spurious Correlation tasks (ViT-B/16).

Parameter	Value / Setting
General & LoRA Parameters	
See Table 8 for most general and LoRA parameters.	
Batch Size	128
Number of Epochs	30
FVAE-LoRA Specific Parameters	
Latent Dim. $\mathbf{z}_1$	16
Latent Dim. $\mathbf{z}_2$	16
FVAE Architecture	Similar to Table 8
Prior $p_1(\mathbf{z}_1)$	$\mathcal{N}(0, I)$
Prior $p_2(\mathbf{z}_2)$	$\mathcal{N}(1.5, I)$
$\boldsymbol{\lambda}$ (Eq. 6)	1×10^{-3} or 1×10^{-4}
ELBO Coeff. α (Reconstr.)	0.1 or 1
ELBO Coeff. β (KL $q_1\|\|p_1$)	1 or 10
ELBO Coeff. δ	1

E Early Attempts at Latent Space Factorization

The most straightforward way to enforce repulsion between $\mathbf{z}_1$ and $\mathbf{z}_2$ in the two-variable ELBO (see Eq. 2) would be to augment the ELBO with the terms:

$$+D_{\mathrm{KL}}\left(q_{\phi_2} \,\|\, p_1\right) + D_{\mathrm{KL}}\left(q_{\phi_1} \,\|\, p_2\right). \tag{9}$$

However, early experiments with this approach yielded poor results, in fact, performance was worse than with LoRA. From a theoretical standpoint, adding such terms to the two-variable ELBO effectively cancels out q_{ϕ_2} and q_{ϕ_1} from the objective, leading instead to a direct repulsion between the priors p_1 and p_2, which is not desirable. Other similar approaches such as directly repelling q_{ϕ_1} and q_{ϕ_2} suffered from the same issue. We found that all overly symmetric and direct formulations, including two-term symmetric variants, were ultimately unfruitful. To avoid this cancellation effect, we instead propose an indirect way to introduce repulsion between $\mathbf{z}_1$ and $\mathbf{z}_2$ by introducing a cross-term between the parametric encoder q_{ϕ_2} and the latent distribution p_1. This solution is theoretically grounded: we show that it induces a geometric separation, measured through a Wasserstein upper bound, between the two encoders q_{ϕ_1} and q_{ϕ_2}. It is also supported by experimental results, outperforming LoRA across all tested modalities.

F Additional Insights in FVAE-LoRA

Regarding the objective in Eq. 4, our proposed loss is a novel objective derived from and inspired by the Evidence Lower Bound, but it is not a strict lower bound on the marginal log-likelihood $\log p(\mathbf{x})$. By introducing the repulsive regularization term Γ, we modify the standard ELBO to enforce factorization between the latent spaces. This term is essential for the method's success, but it means the objective no longer serves as a formal lower bound on the data log-likelihood in the traditional VAE sense.

FVAE-LoRA intentionally sacrifices static weight merging to enable a more powerful dynamic, input-dependent adaptation. By computing the adaptation specifically for each input $\mathbf{x}$, our model learns more robust and fine-grained representations. We believe this dynamic mechanism is the key to its performance edge, a capability validated by our strong results on the spurious correlation benchmarks. This trade-off is therefore central to achieving the higher performance and robustness we demonstrate.

Note that simply reducing the rank of LoRA is a simple and efficient form of regularization, however It compresses all information flowing through the adapter, without distinguishing between features that are useful, irrelevant, or even detrimental to the downstream task. Our hypothesis is that large foundation models, pretrained on vast and general datasets, contain rich and entangled set of features. For any specific downstream task, some features are highly relevant (the "signal"), some are irrelevant but harmless, and some are actively harmful. The most prominent example of these detrimental features are spurious correlations (e.g., a water background being correlated with a "waterbird" label). A standard fine-tuning process, which optimizes a task-specific loss, may still latch onto these spurious features because they are prevalent in the training data and help minimize the training loss. This leads to poor generalization on data where that correlation is broken. This is why FVAE-LoRA is designed to be a more intelligent filter. Its goal is not just to compress, but to actively separate and isolate these different types of information. By using two latent spaces ($\mathbf{z}_1$ and $\mathbf{z}_2$) and our novel factorization objective, we encourage the model to encode task-salient, causal information in $\mathbf{z}_1$ while relegating the residual, non-essential, or spurious information to $\mathbf{z}_2$. The most direct validation of this rationale is in our spurious correlation experiments (Section 4.2). These results show that FVAE-LoRA is significantly more robust to misleading features than standard LoRA, confirming that it successfully learns to rely on the core features isolated within $\mathbf{z}_1$. This ability to "denoise" the adaptation is why it ultimately achieves better and more reliable performance.

G A Practical Guide on Hyperparamters Selection

The factorization in FVAE-LoRA is governed by a subtle equilibrium between reconstruction and regularization, enforced by our ELBO objective. The key hyperparameters, β and δ, control this balance.

β and the Task-Salient Space ($\mathbf{z}_1$). The β parameter controls the KL divergence on $\mathbf{z}_1$, our task-salient latent space. Its role is critical, as it enforces a structured and efficient representation of the task-salient features. To understand its impact, we experimented with a wide range of values.

A significantly lower value, such as $\beta = 0.1$, led to a drastic degradation in performance across all tasks. This is because a near-zero β effectively removes the KL divergence term, freeing the encoder for $\mathbf{z}_1$ to learn an unconstrained and arbitrarily complex representation. This removes the crucial pressure for the learned posterior $q_{\phi_1}(\mathbf{z}_1|\mathbf{x})$ to align with the prior $p_1(\mathbf{z}_1)$, leading to overfitting and a loss of generalization. This result is not merely a poor tuning choice; it is critical evidence that enforcing this prior alignment is essential for learning a robust and meaningful task-salient space.

Conversely, we explored a much higher value of $\beta = 100$. While this yielded marginal improvements over $\beta = 10$ on some specific tasks, the gains were not significant enough to justify such a strong constraint. An overly large β can create an information bottleneck, punishing the model so heavily for deviating from the prior that it struggles to encode sufficient task-specific information in $\mathbf{z}_1$.

This evidence from both extremes reveals a necessary balance. The optimal values, which we found to be in the range of 1 to 10, are large enough to enforce a structured, regularized space but not so large as to prevent the learning of useful features.

δ and Latent Space Separation. The δ parameter controls the strength of our repulsive regularizer, Γ, which is the primary mechanism for enforcing factorization between the task-salient space ($\mathbf{z}_1$) and the residual space ($\mathbf{z}_2$). Our empirical results consistently show that $\delta = 1$ provides sufficient repulsive force to achieve this separation effectively, as was demonstrated in the spurious correlation experiments. We recommend $\delta = 1$ as a robust and generally optimal default.

For practical application, users can start with $\delta = 1$ and tune β (typically between 1 and 10) to adjust the regularization on the learned task-salient features. This is a crucial point for the practical application of our method.

H Computational Cost Analysis

Our empirical results on image classification tasks show that the training time for FVAE-LoRA is approximately 30% higher than that of the strong DoRA baseline. This increase is primarily due to the additional forward and backward pass through the VAE's decoder during the training phase. However, the inference-time overhead is significantly lower because only the lightweight $\mathbf{z}_1$ encoder is used.

I Future Work

A particularly exciting avenue for future work lies in exploiting the inherent generative capabilities of the FVAE framework. Key directions will include exploiting the generative capabilities of the FVAE decoder for principled data augmentation, applying our latent factorization principle to other PEFT methods beyond LoRA, exploring approximate high-rank adaptation methods like HiRA, and exploring architectural enhancements such as allocating adaptive parameter budgets or different latent space ranks to different layers.

J Limitations

While FVAE-LoRA demonstrates promising results across diverse modalities, several limitations of the current work remain. First, the LoRA rank is fixed to a value of 16 across all experiments. Although this ensures consistent parameter budgets, it may not represent the optimal configuration for each task or domain, potentially limiting performance. Second, FVAE-LoRA and all baselines are applied only to the query and key matrices of the transformer models. This restricted application may overlook potential gains from adapting other components such as value matrices or feedforward layers.

Furthermore, detailed hyperparameter settings and VAE-specific architectural choices are provided in the appendix. This separation may hinder reproducibility, for readers interested in extending the approach. Nevertheless, the code will be open-sourced after publication.

In addition, while modality-specific baselines are chosen with the goal of providing meaningful comparisons, we do not evaluate against stronger non-LoRA or non-factorized alternatives, which may offer a more comprehensive picture of relative performance. Further, the paper does not report the computational cost of training or inference, which is important for assessing the practical deployment potential of the method, especially in resource-constrained environments.

Finally, a practical limitation of FVAE-LoRA is that its adapter weights cannot be merged back into the original base model after training. This stands in contrast to some LoRA-based methods that allow for such weight merging, which can simplify inference or reduce model complexity at deployment time.

K Broader Impacts

FVAE-LoRA advances parameter-efficient fine-tuning by enabling a factorized latent representation. Potential positive impacts include more effective and robust model adaptation across modalities, potentially leading to improved performance, resource efficiency, and more reliable AI systems, especially in handling spurious correlations, as demonstrated in our experiments. This can also enhance accessibility to powerful AI capabilities for a wider range of researchers and developers.

However, techniques that improve the adaptability of large foundation models also carry inherent risks. Easier and more effective fine-tuning could lower barriers for misuse in sensitive areas such as the generation of sophisticated disinformation or the development of enhanced surveillance tools. While FVAE-LoRA aims to disentangle task-salient information, it does not inherently mitigate biases that may be present in the original pre-trained models or the fine-tuning data. Indeed, such biases could potentially be concentrated or even amplified within the task-salient latent space if not proactively identified and addressed.

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 main claims in the abstract and introduction accurately reflect the paper's contribution and scope.

   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: The authors provide sufficient information regarding limitations of the current work in the appendix J.

   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: The authors provide for all theoretical results the full set of assumptions accompanied by their proofs.

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 authors provide all the necessary information need to reproduce the experiments, including hyperparameters, optimizer, model architectures, etc. in the appendix.

   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: [No]

Justification: Access to the data and code will be provided in a following step.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details are provided throughout the paper and any remaining details and hyperparameters are provided in the appendix.

   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: The authors provide standard deviation for all experiments across different runs.

   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: [No]

   Justification: The authors have not provided this information at the current stage.

   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: The paper raises no ethical concerns and complies with the NeurIPS guidelines. No high-risk applications or data are involved, and fairness and privacy considerations are acknowledged where applicable.

   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: The authors include a "Broader Impacts" section that discusses potential benefits and risks in the appendix.

    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 models released do not pose potential for misuse.

    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: All datasets and code used are properly cited, and licenses are referenced. There is no evidence of license violations.

    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` 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: [Yes]

    Justification: Any new assets introduced include usage instructions, and documentation appears sufficient for independent use.

    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: There are no experiments related to crowdsourcing and research with human subjects.

    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: There is no research with human subjects

    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: [NA]

    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.