Trading Information between Latents in Hierarchical Variational Autoencoders

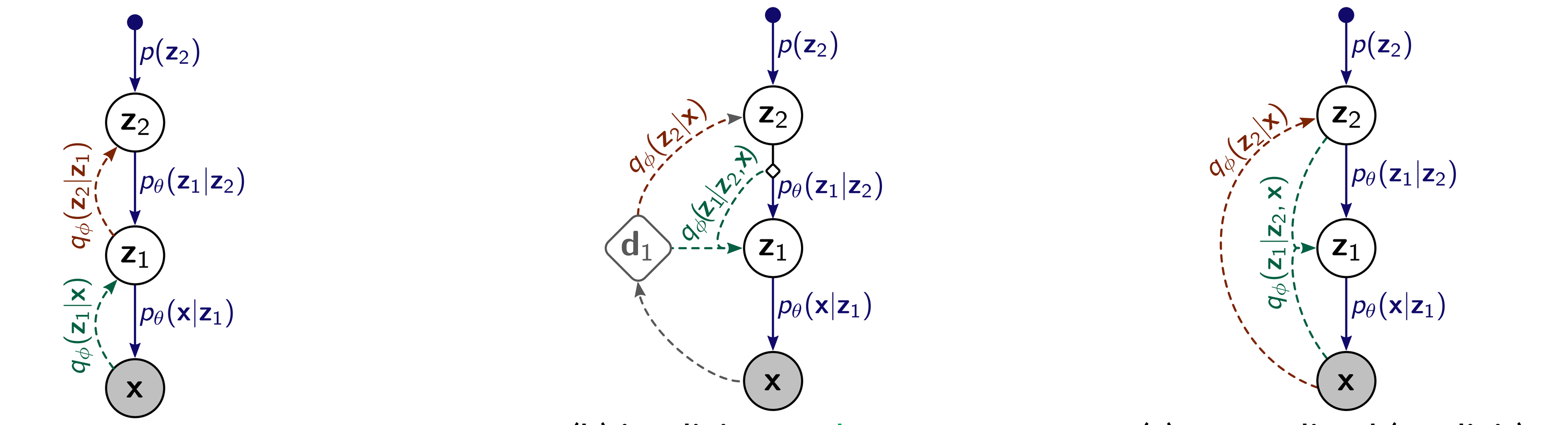
Tim Z. Xiao, Robert Bamler
University of Tübingen

SCAN ME

Three application domains of VAEs

- **Data Reconstruction tasks:** involve both the *encoder and decoder*.
- **Representation Learning tasks:** involve only the *encoder*.
- **Generative Modeling tasks:** involve only the *decoder*.

A Hierarchical Information Trading Framework



(a) bottom-up (like diffusion model) (b) implicit top-down (e.g., LVAE) (c) generalized (explicit) top-down

Figure 1: Inference and generative models for hierarchical VAEs (HVAEs) with two layers of latent variables. The diamond $\mathbf{d}_1$ in b is the result of a deterministic transformation of $\mathbf{x}$.

Generative Model:

$$p_\theta(\{z_\ell\}, x) = p_\theta(z_L)\, p_\theta(z_{L-1}|z_L)\, p_\theta(z_{L-2}|z_{L-1}, z_L) \cdots p_\theta(z_1|z_{\geq 2})\, p_\theta(x|z_{\geq 1}) \qquad (1)$$

Top-down Inference Model:

$$q_\phi(\{z_\ell\}\,|\,x) = q_\phi(z_L|x)\, q_\phi(z_{L-1}\,|\,z_L, x)\, q_\phi(z_{L-2}\,|\,z_{L-1}, z_L, x) \cdots q_\phi(z_1\,|\,z_{\geq 2}, x) \qquad (2)$$

β-VAE and rate/distortion trade-off

$$\mathcal{L}_\beta(\theta,\phi) = \mathbb{E}_{x\sim\mathbb{X}_{\text{train}}}\Big[\underbrace{\mathbb{E}_{q_\phi(\{z_\ell\}|x)}\big[-\log p_\theta(x|\{z_\ell\})\big]}_{=\text{“distortion” } D} + \beta\, \underbrace{D_{\mathrm{KL}}\big[q_\phi(\{z_\ell\}\,|\,x)\,\|\,p_\theta(\{z_\ell\})\big]}_{=\text{“rate” } R}\Big] \qquad (3)$$

For top-down inference models, the total rate R splits into a sum of layer-wise rates

$$R = \mathbb{E}_{q_\phi(\{z_\ell\}|x)}\left[\log\frac{q_\phi(z_L|x)}{p_\theta(z_L)} + \log\frac{q_\phi(z_{L-1}|z_L,x)}{p_\theta(z_{L-1}|z_L)} + \ldots + \log\frac{q_\phi(z_1|z_{\geq 2},x)}{p_\theta(z_1|z_{\geq 2})}\right] \qquad (4)$$
$$= R(z_L) + R(z_{L-1}|z_L) + R(z_{L-2}\,|\,z_{L-1}, z_L) + \ldots + R(z_1|z_{\geq 2}).$$

And control each layer's rate separately

$$\mathcal{L}_\beta(\theta,\phi) = \mathbb{E}_{x\sim\mathbb{X}_{\text{train}}}\big[D + \beta_L R(z_L) + \beta_{L-1} R(z_{L-1}|z_L) + \ldots + \beta_1 R(z_1|z_{\geq 2})\big]. \qquad (5)$$

Information-Theoretical Performance Bounds

1. For Data Reconstruction and Manipulation

$$\mathbb{E}_{p_{\text{data}}(x)}[D] \geq H[p_{\text{data}}(x)] - \mathbb{E}_{p_{\text{data}}(x)}\big[R(z_L) + R(z_{L-1}|z_L) + \cdots + R(z_1|z_{\geq 2})\big] \qquad (6)$$

2. For Representation Learning (e.g., downstream classification)

$$\text{class. accuracy} \leq f^{-1}(I_q(y; z_\ell)) \leq f^{-1}(\mathbb{E}_{p_{\text{data}}(x)}[R(z_{\geq \ell})]) \quad (\leq f^{-1}(\mathbb{E}_{p_{\text{data}}(x)}[R])) \qquad (7)$$

3. For Data Generation

Setting all β-hyperparameters in Eq. 5 to values close to 1 if a HVAE is used primarily for its generative model p_θ.

There is no "*one-size-fits-all*" HVAE!

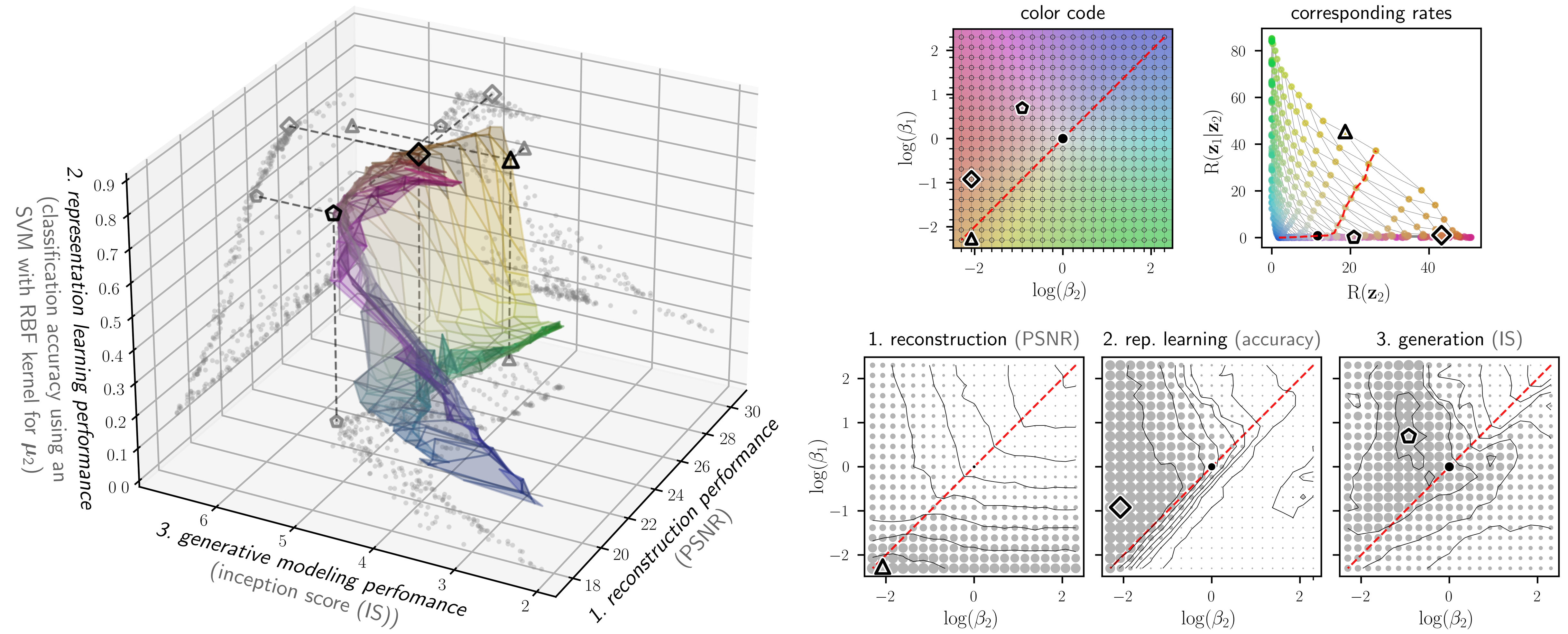
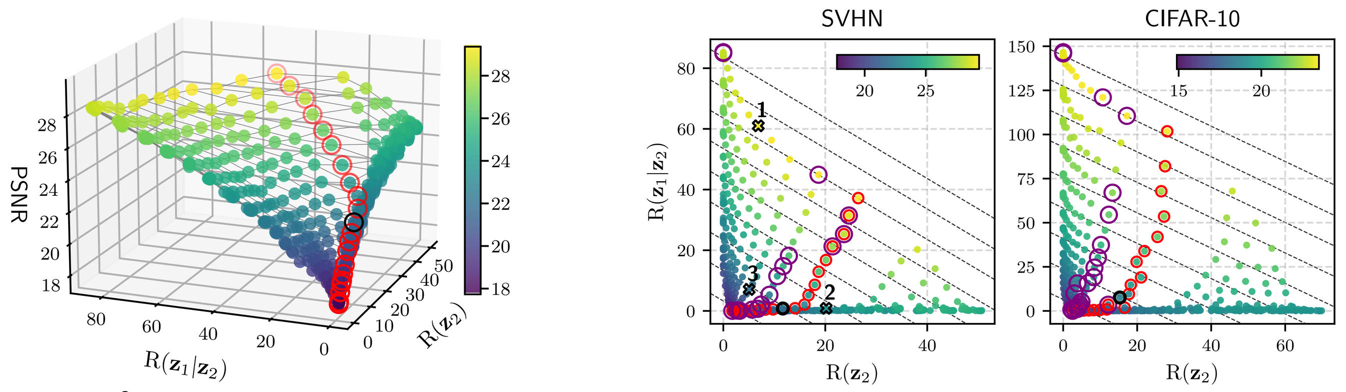


Figure 2: Left: trade-off between performance on SVHN. Right: color code, corresponding layer-wise rates (Eq. 5), and individual performance landscapes (size of dots $\propto$ performance). △: best data reconstruction; ⋄: best representation learning; ⬠: best generative modeling. Note that performance landscapes for the three tasks differ strongly; neither a standard VAE (marked "•") nor a conventional β-VAE (dashed red lines) result in optimal performances.

1. Data Reconstruction



(a) R^2/D surface (SVHN) (b) PSNR-rates comparison in 2d.

Figure 3: PSNR-rate trade-off. "○" mark $\beta_2 = \beta_1 = 1$; "○" mark $\beta_2 = \beta_1$; and "○" mark optimal models (refer to Figure 7) along constant total rate (dashed diagonal lines). Crosses point to columns in Figures 5.

2. Representation Learning

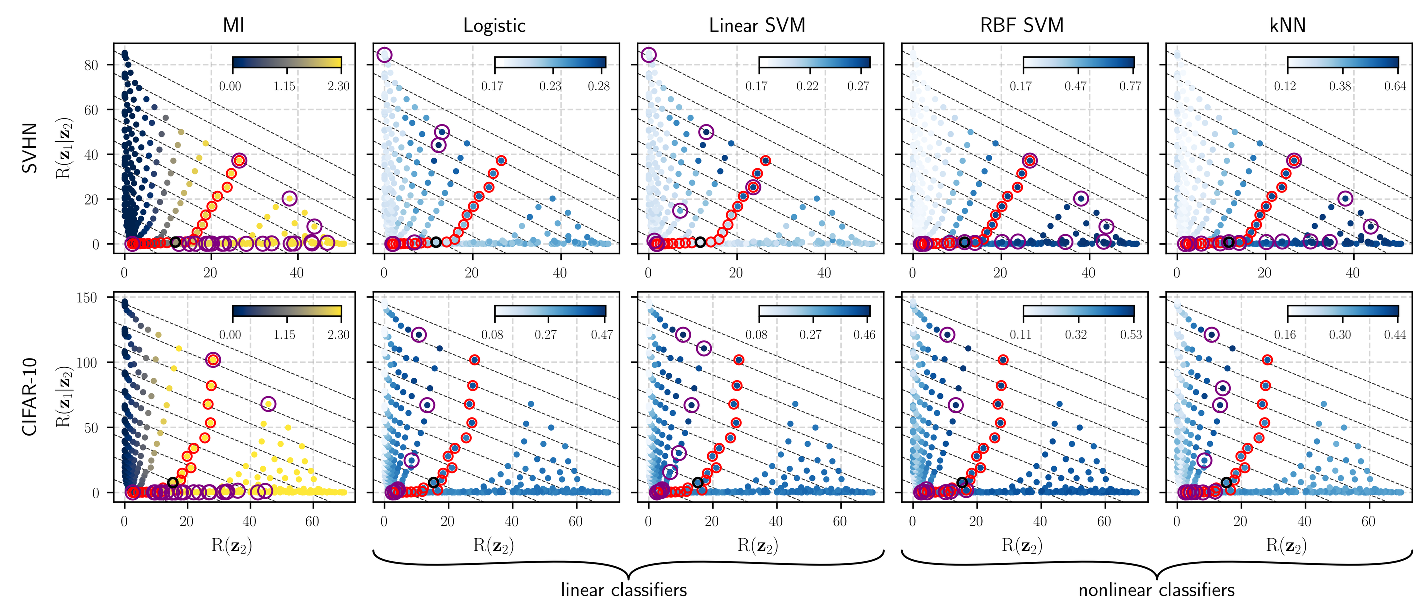


Figure 4: Mutual information (MI) $I_q(y; z_2)$ and classification accuracies as a function of layer-wise rates $R(z_2)$ & $R(z_1|z_2)$. Classifiers are conditioned on $\mu_2 := \arg\max_{z_2} q(z_2|x)$. Simple (linear) classifiers perform best on low $R(z_2)$.

3. Sample Generation

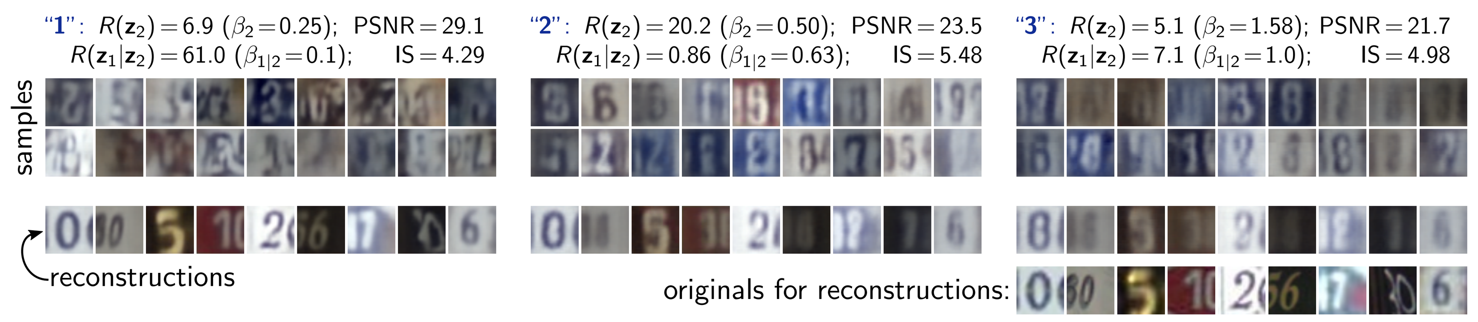


Figure 5: Samples (top) and reconstructions (bottom) from 3 different models (blue column labels "1", "2", and "3" from left to right correspond to crosses "1", "2", and "3" in Figures 3b & 6). Consistent with PSNR and IS metrics, model "1" produces poorest samples but best reconstructions.

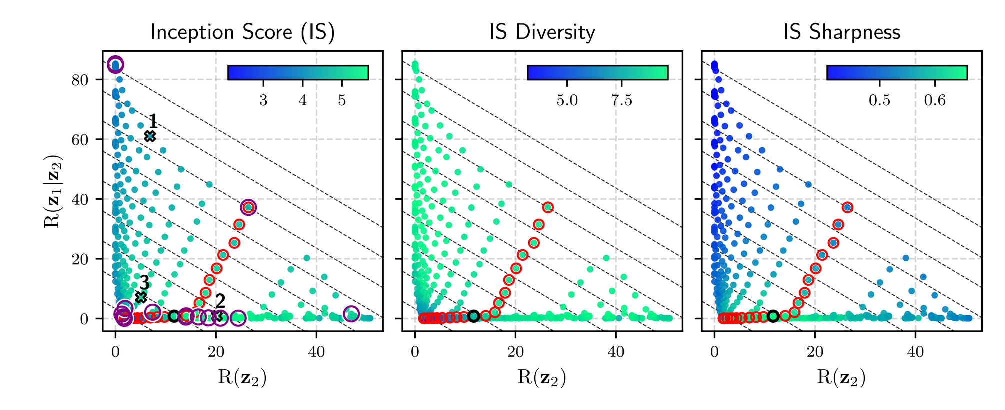


Figure 6: Sample generation performance, measured in Inception Score (Eq. 8) and its factorization into diversity and sharpness (Eq. 9) as a function of layer-wise rates on SVHN data. Increasing the rate $R(z_1|z_2)$ of lower-level latents increases sharpness, while high-level latents seem to be more important for diversity.

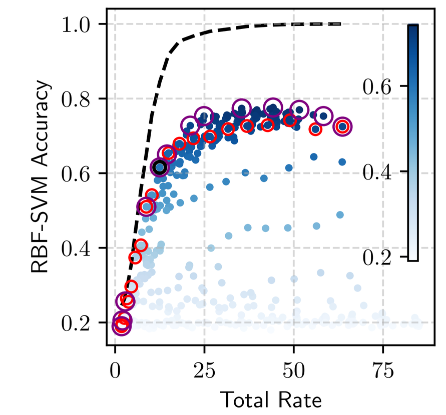


Figure 7: RBF-SVM classification accuracies on μ_2. Dashed line shows bound (Eq. 7).

Inception Score:

$$\text{IS} = \exp\left\{\mathbb{E}_{p_\theta(x)}\big[D_{\mathrm{KL}}[p_{\text{cls.}}(y|x)\,\|\,p_{\text{cls.}}(y)]\big]\right\} \qquad (8)$$
$$= e^{H[p_{\text{cls.}}(y)]} \times e^{-\mathbb{E}_{p_\theta(x)}[H[p_{\text{cls.}}(y|x)]]} \qquad (9)$$

Presentation on 11th International Conference on Learning Representations (ICLR), 2023