Improving Generation Quality of Long-Tailed Diffusion via Disentangled Latent Representations

Diffusion models have achieved impressive performance in generating high-quality and diverse synthetic data. However, their success typically assumes a class-balanced training distribution. In real-world settings, multi-class data often follow a long-tailed distribution, where standard diffusion models struggle—producing low-diversity and lower-quality samples for tail classes. While this degradation is well-documented, its underlying cause remains poorly understood.

We investigate the behavior of diffusion models trained on image long-tailed datasets and identify a key issue: the latent representations from the bottleneck layer of the U-Net for tail class subspaces exhibit significant overlap with those of head classes, leading to feature borrowing and poor generation quality. This representation entanglement is not merely due to limited data per class, but the relative class imbalance significantly contributes to this phenomenon. Through extensive visualizations using t-SNE and UMAP on various long-tailed datasets, we demonstrate how tail-class representations are absorbed into dominant clusters, undermining the model's ability to preserve class-specific features.

We propose **CO**ntrastive **R**egularization for **A**ligning **L**atents **(CORAL)**, a contrastive latent alignment framework that leverages supervised contrastive losses to encourage well-separated latent class representations. CORAL augments the encoder of the denoising U-Net with a projection head applied to the bottleneck output. The resulting projected embeddings are trained using a supervised contrastive loss, which is then combined with the standard diffusion objective through a time-dependent weighting scheme. CORAL regularizes the internal feature space of the diffusion model itself, where representation entanglement arises. This encourages the model to pull together representations of samples from the same class while pushing apart those from different classes, thus promoting class-wise separation in the latent space where it matters most for generation quality.

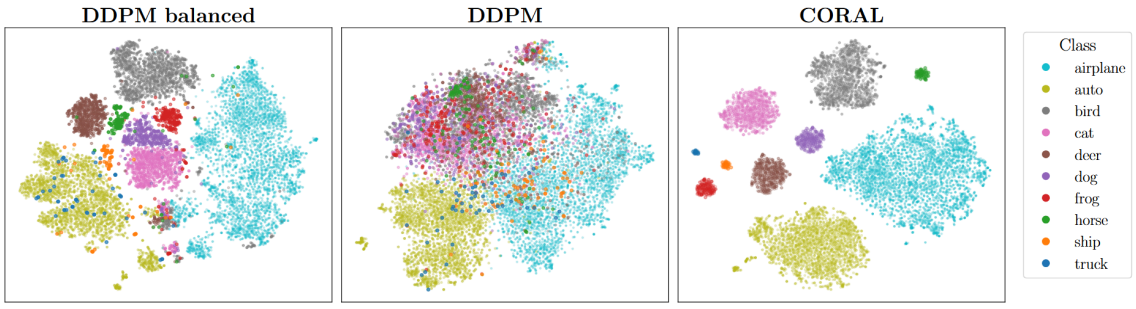

Figure: t-SNE plots. Balanced CIFAR10 (left), imbalanced CIFAR10-LT (middle), and CORAL (right).

This work reveals representation entanglement as a cause of poor tail-class generation in diffusion models and demonstrates that contrastive regularization of internal latent representations provides a principled solution for improving generative performance on imbalanced datasets, with broader implications for semantically meaningful representations and robustness in generative modeling.

We evaluate CORAL on CIFAR10-LT, CIFAR100-LT, CelebA-5 and ImageNet-LT. CORAL significantly outperforms state-of-the-art methods across all datasets and standard image generation metrics. CORAL achieves improved results while requiring minimal architectural overhead, demonstrating that direct latent space intervention is more effective than ambient space modifications. Qualitative analysis shows CORAL generates samples with improved diversity and visual fidelity while improving tail-class generation.