Online Clustered Codebook

Chuanxia Zheng, Andrea Vedaldi
Visual Geometry Group, University of Oxford

Motivation

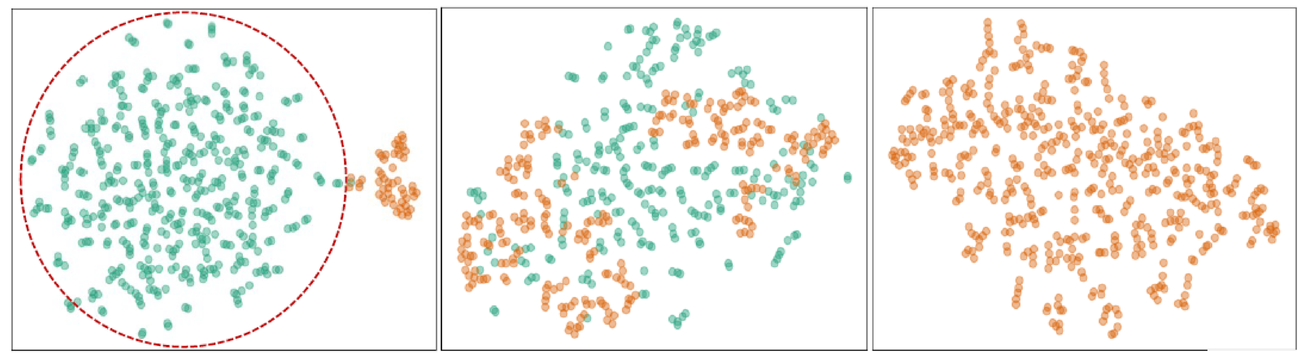
Goal: Learn a visual codebook for tasks such as generation with full codevectors utilisation.

Issue:

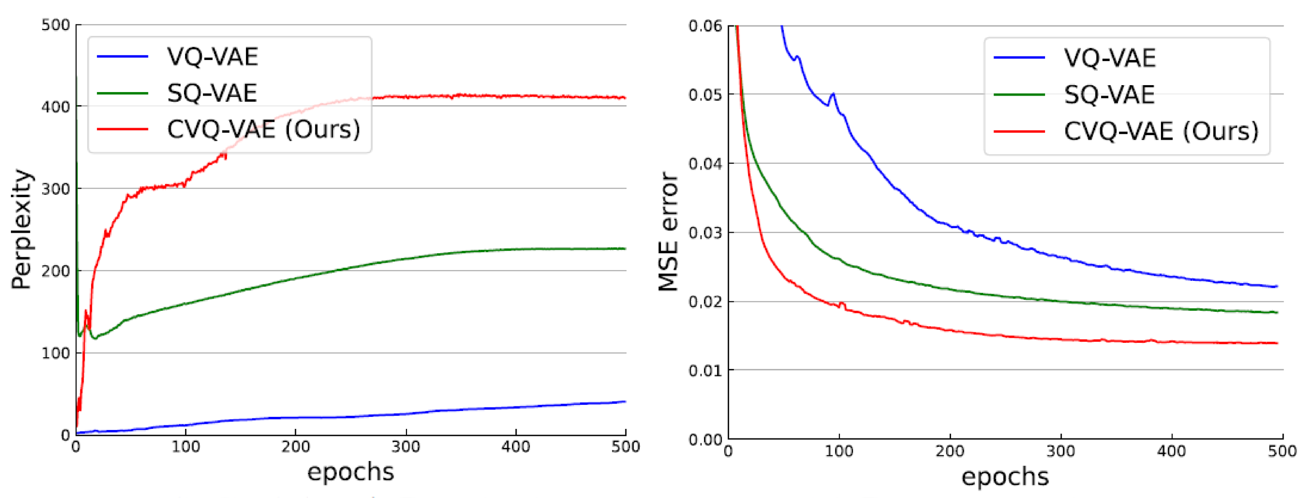
1.Codebook collapse. Only a small subset of active codebook entries are optimized

2. Stop-gradient operator. Loss can only back propagate to the selected entries.

Green Points: "**Dead**" Codebook Entries



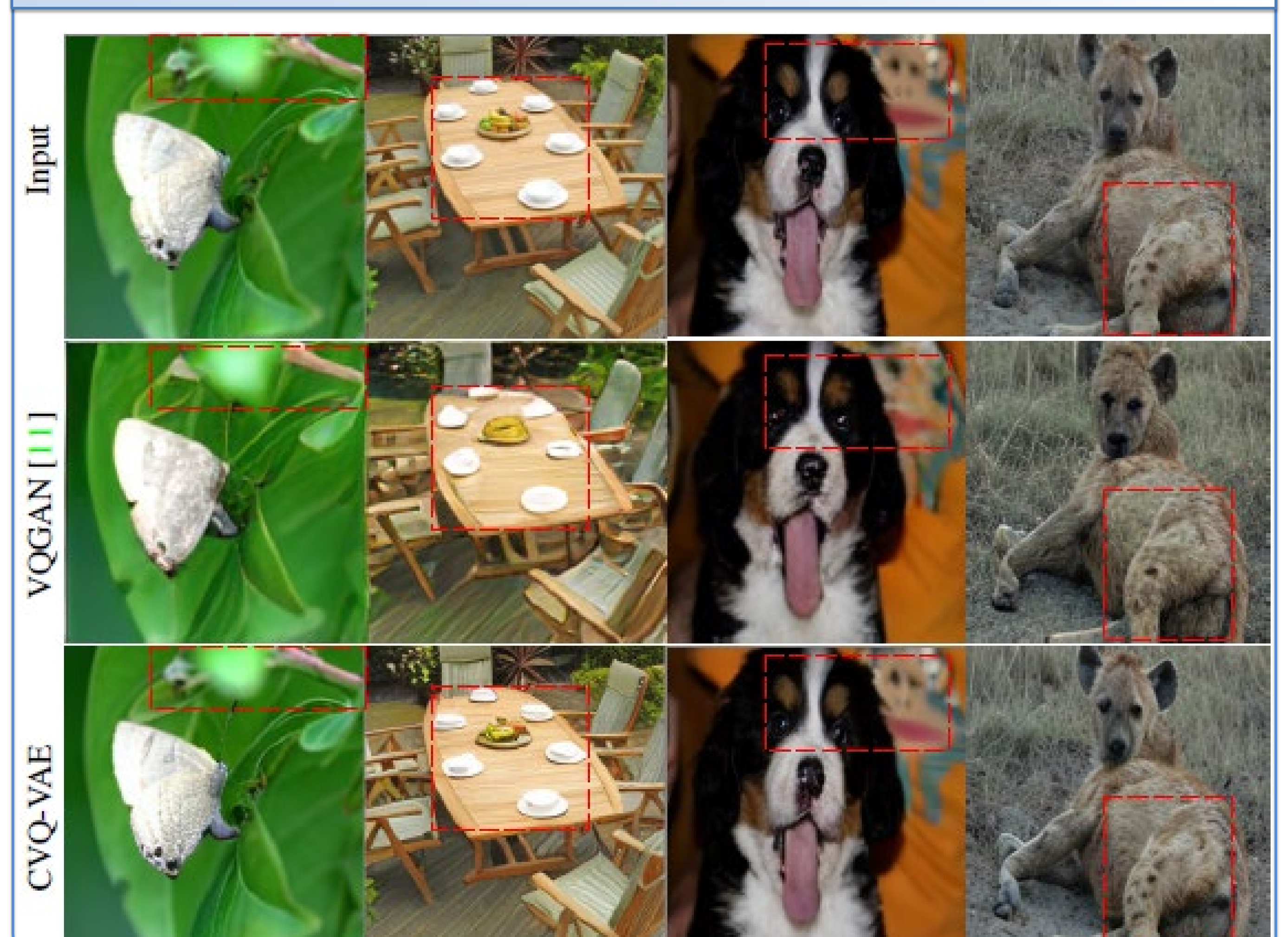
(a) VQ-VAE [37] Usage: 9.96% (b) SQ-VAE [36] Usage: 49.02% (c) CVQ-VAE Usage: 100%



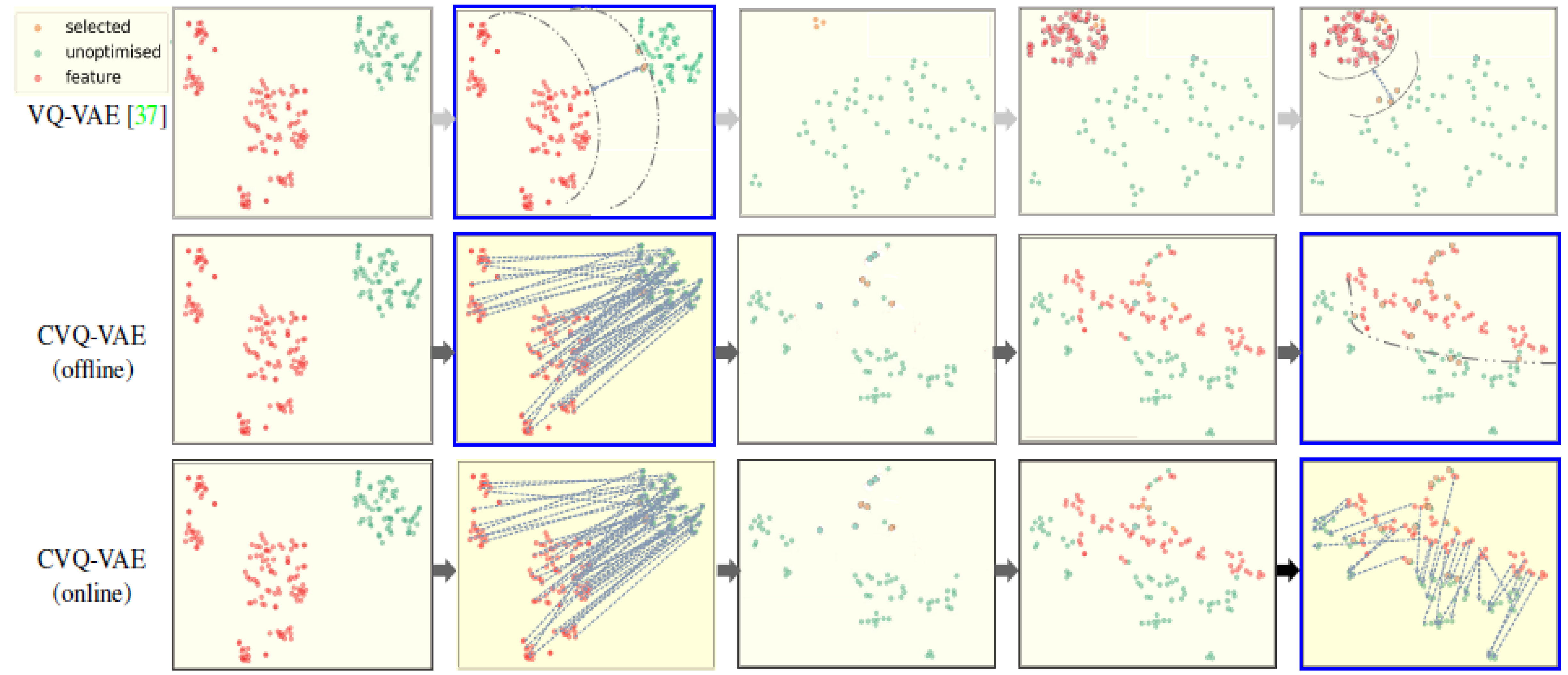
(d) Codebook Perplexity (e) Reconstruction error

How to ensure the embedded features and codebook entries closely adhere to the same distribution?

Stage-1: Perceptual Compression



Contribution: Online Clustered Codebook Optimisation



(a1) Distribution (b1) Code vector update (c1) Updated codebook (a2) Distribution (b2) Code vector update

Codebook Usage:

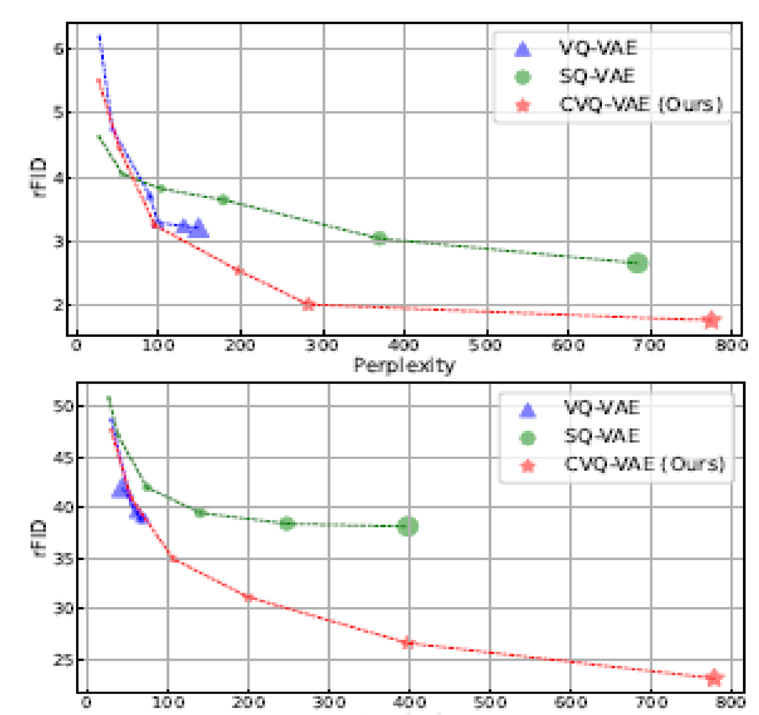
$$N_k^{(t)} = N_k^{(t-1)} \cdot \gamma + \frac{n_k^{(t)}}{Bhw} \cdot (1-\gamma)$$

Running average updates:

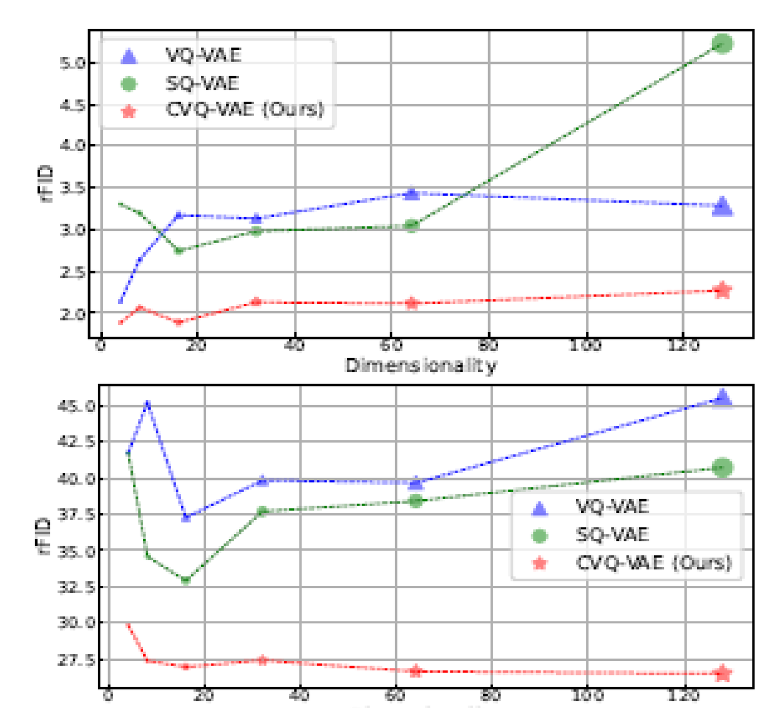
$$\alpha_k^{(t)} = \exp^{-N_k^{(t)} K \frac{10}{1-\gamma} - \epsilon},$$

$$e_k^{(t)} = e_k^{(t-1)} \cdot (1-\alpha_k^{(t)}) + \hat{z}_k^{(t)} \cdot \alpha_k^{(t)}$$

Stage-1: Ablation Study on Image Quantization



(a) **Codebook size.** The blobs' size is proportional to the number of codebook vectors {32, 64, 128, 256, 512, 1024}. The larger size naturally leads to better results in our CVQ-VAE.



(b) **Codebook dimensionality.** The blob's size refers to the dimensionality of codebook vectors {4,8,16,32,64,128}. The higher dimensionality does not ensure a better representation.

Method	Dataset	rFID↓ (offline)	rFID↓ (online)
random	MNIST	3.20	2.27
unique	MNIST	2.84	2.24
probability	MNIST	2.78	**2.23**
closest	MNIST	**2.51**	2.59
random	CIFAR10	34.49	26.04
unique	CIFAR10	36.99	26.02
probability	CIFAR10	**31.10**	26.62
closest	CIFAR10	32.31	**25.99**

(c) **Anchor sampling methods.** The choice of anchor sampling method has a significant impact on offline (one-time) feature initialization, while the online clustered method is robust for various samplings.

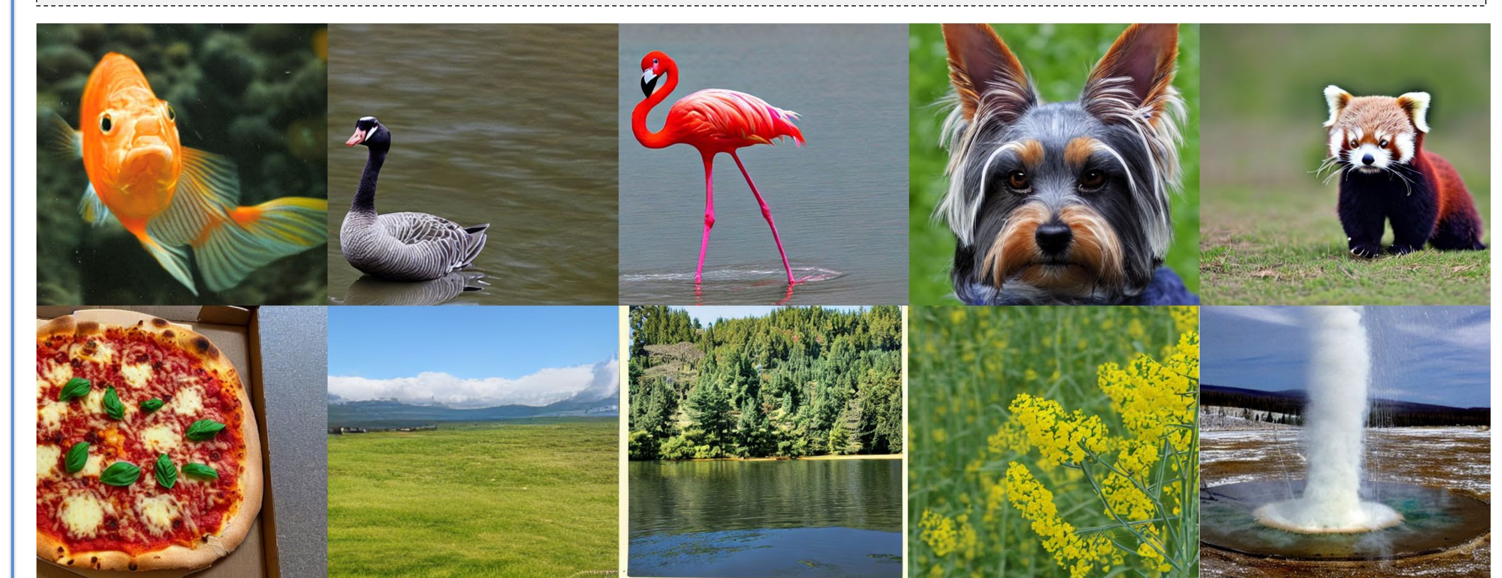
Methods	MNIST (28×28) SSIM↑	MNIST LPIPS↓	MNIST rFID↓	CIFAR10 (32×32) SSIM↑	CIFAR10 LPIPS↓	CIFAR10 rFID↓	FFHQ (256×256) SSIM↑	FFHQ LPIPS↓	FFHQ rFID↓
near codevectors [39]	0.9790	0.0270	3.17	0.8553	0.2553	41.08	0.7282	0.1085	4.31
hard encoded features [8]	0.9814	0.0243	2.25	0.8988	0.1978	29.16	0.7646	0.0870	3.91
running average (ours)	**0.9823**	**0.0236**	**2.23**	**0.8991**	**0.1897**	**26.62**	**0.8193**	**0.0603**	**2.94**

(d) **Codebook reinitialization methods.** In previous works [39, 8], each code entry is associated only with a single feature.

Stage-2: High-Fidelity Image Generation

Unconditional Generation on LSUN

Class-conditional Generation on ImageNet



Quantitative Results on Image Generation

Methods	FID↓ Churches	FID↓ Bedrooms
StyleGAN [19]	4.21	2.35
DDPM [16]	7.89	4.90
ImageBART [10]	7.32	5.51
Projected-GAN [35]	**1.59**	**1.52**
LDM [32]-8*	4.02	-
LDM [32]-4	-	2.95
LDM [32]-8 (reproduced)	4.15	3.57
CVQ-VAE-LDM [32]-8	3.86	3.02

Model	FFHQ Steps	FFHQ FID↓	ImageNet Steps	ImageNet FID↓
RQVAE [22] CVPR'2022	256	10.38	1024	7.55
MoVQ [44] NeurIPS'2022	1024	8.52	1024	7.13
SQ-VAE [33] ICML'2022	200	5.17	250	9.31
LDM-4 [31] CVPR'2022	200	4.98	250	10.56
CVQ-VAE (ours)	200	**4.46**	250	**6.87**

Other prior related works

1. Zheng, C., Vuong, T. L., Cai, J., & Phung, D. Movq: Modulating quantized vectors for high-fidelity image generation. NeurIPS, 2022.. (MoVQ was reported means a lot to Kandinsky2.1., Github)

2. Vuong, T. L., Le, T., Zhao, H., Zheng, C., Harandi, M., Cai, J., & Phung, D. Vector Quantized Wasserstein Auto-Encoder. ICML 2023.