You're your own best teacher: A Self-Supervised Learning Approach For Expressive Representations in Time Series Generation

Johan Vik Mathisen

June 14, 2024

In this thesis, we focus on two main objectives, which relates back to the research questions. Firstly, in Stage 1, we aim to determine whether NC-VQVAE can learn more expressive representations compared to VQVAE. Specifically, we investigate whether NC-VQVAE can achieve reconstruction performance on par with VQVAE while simultaneously enhancing downstream classification. In Stage 2, our interest lies in examining the impact of NC-VQVAE on synthetic sample quality.

Our evaluation process begins with assessing the tokenization models, focusing on their reconstruction capability and performance in downstream classification tasks. Subsequently, we then evaluate the performance of the generative models using metrics such as IS (Inception Score), FID (Fréchet Inception Distance), and CAS (Classification Accuracy Score). Additionally, visual inspections are conducted to provide further insights into the models' performance. A small ablation study on the effect of augmentation reconstruction weight on reconstruction and probe accuracy is presented towards the end.

0.1 Stage 1

In this section, we present the findings concerning the tokenization model. We find that some configuration of NC-VQVAE outperform naive VQVAE across the majority of datasets, both in terms of reconstruction quality and downstream classification, providing significant improvements in probe accuracy.

0.1.1 Reconstruction

We present the top 1 and mean reconstruction losses across the four runs in Table **??** and Table **??**, respectively.

Mean validation reconstruction error

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.217	0.127	0.134	**0.108**	0.173	0.169	0.203
ElectricDevices	**0.041**	0.067	0.044	0.049	0.105	0.042	0.049
StarLightCurves	**0.032**	0.042	0.069	0.071	0.052	0.050	0.068
Wafer	0.044	0.037	0.048	0.049	**0.035**	0.042	0.039
ECG5000	**0.048**	0.083	0.170	0.104	0.093	0.205	0.064
TwoPatterns	0.197	0.201	**0.184**	0.230	0.214	0.186	0.207
UWaveGestureLibraryAll	0.190	**0.172**	0.190	0.245	0.189	0.178	0.237
FordB	0.150	0.115	0.122	0.123	**0.114**	0.121	0.142
ShapesAll	**0.045**	0.056	0.066	0.102	0.064	0.069	0.073
SonyAIBORobotSurface1	0.402	0.509	0.494	0.491	**0.360**	0.363	0.418
SonyAIBORobotSurface2	0.623	0.622	0.618	0.640	0.487	**0.454**	0.589
Symbols	0.110	0.143	0.134	0.173	0.078	**0.067**	0.105
Mallat	0.066	0.081	0.091	0.096	0.066	0.067	**0.060**

Table 1: Mean validation reconstruction error across all 13 datasets. Results are averaged over four runs.

Top 1 validation reconstruction error

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.158	0.108	0.111	**0.087**	0.130	0.134	0.113
ElectricDevices	0.036	0.060	0.034	0.043	0.092	**0.031**	0.045
StarLightCurves	**0.026**	0.037	0.057	0.055	0.043	0.048	0.065
Wafer	0.038	0.031	0.045	0.043	**0.027**	0.031	0.038
ECG5000	**0.044**	0.069	0.156	0.084	0.080	0.181	0.056
TwoPatterns	0.181	0.184	**0.169**	0.208	0.200	0.172	0.185
UWaveGestureLibraryAll	0.159	**0.145**	0.167	0.201	0.155	0.169	0.233
FordB	0.117	0.094	0.090	0.103	**0.082**	0.094	0.102
ShapesAll	**0.035**	0.043	0.046	0.092	0.061	0.063	0.067
SonyAIBORobotSurface1	0.381	0.473	0.472	0.465	0.329	**0.328**	0.408
SonyAIBORobotSurface2	0.513	0.577	0.536	0.588	0.444	**0.414**	0.470
Symbols	0.088	0.111	0.122	0.150	0.062	**0.059**	0.090
Mallat	0.061	0.075	0.076	0.088	0.059	0.059	**0.057**

Table 2: Top 1 validation reconstruction error across all 13 datasets. Lowest value of the four runs for each model is selected.

It's evident that NC-VQVAE achieves comparable reconstruction performance to the baseline model, and certain configurations even outperform the naive VQVAE in terms of mean reconstruction loss for 9 out of 13 datasets.

In Figure **??**, we observe minimal differences in reconstruction loss across most datasets, regardless of SSL methods and augmentations. However, VIbCReg generally demonstrates slightly better performance compared to Barlow Twins, except for FordA. Additionally, the use of Gaussian augmentation introduces less regularization compared to the other augmentation methods, except for Slice and Shuffle on ECG5000. These findings suggest that incorporating a non-contrastive loss does not compromise the reconstruction capabilities compared to naive VQVAE.

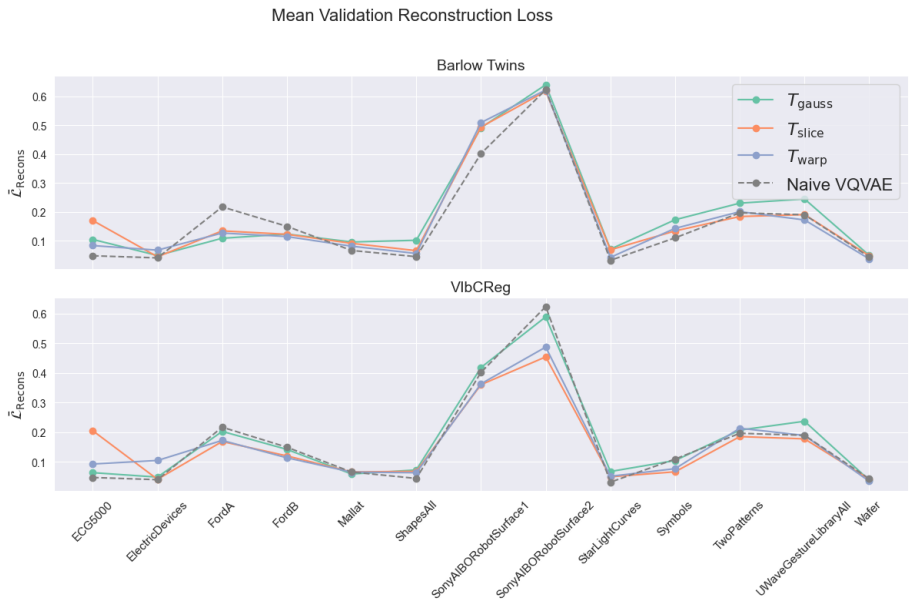


Figure 1: Mean validation reconstruction loss for the two models, compared to naive VQVAE

To explore the impact of the reconstruction loss of the augmented branch on validation reconstruction, a small ablation stydy was conducted, which is presented in Section **??**. The results indicate that the validation reconstruction loss is robust to the specific value of the augmentation reconstruction weight, indicating a minor role played.

Through investigation into the development of validation reconstruction loss during training, we observed that the right configuration for NC-VQVAE can serve as a regularizer. This is illustrated in Figure **??** which depicts the development on FordA.

0.1.2 Classification

We present the mean and top 1 downstream classification accuracy in Table **??** and Table **??**, respectively.

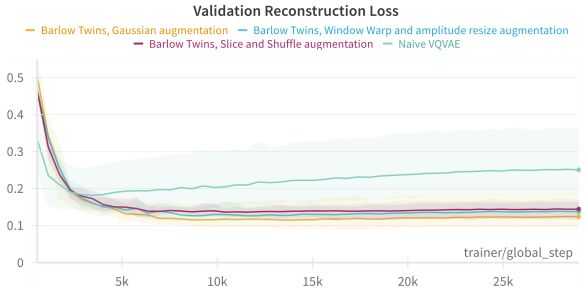


Figure 2: Development of the validation reconstruction loss for Barlow Twins and naive VQVAE on FordA during training. Averaged across all four runs.

Mean linear probe accuracy

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
FordA	0.70	0.74	0.83	0.84	**0.91**	**0.89**	0.80	0.83	0.80	0.74	0.87	0.86	0.76	0.78
ElectricDevices	0.35	0.41	0.35	**0.44**	0.38	0.41	**0.40**	0.42	0.33	0.38	0.36	0.39	0.39	0.43
StarLightCurves	0.87	0.89	0.93	0.93	**0.94**	**0.94**	0.88	0.88	0.92	**0.94**	0.91	0.93	0.89	0.89
Wafer	0.93	0.89	0.96	**0.94**	0.96	**0.94**	0.96	0.93	**0.97**	0.94	0.96	0.92	**0.97**	0.92
ECG5000	0.80	0.83	0.85	0.81	**0.88**	0.84	0.86	**0.84**	0.86	0.82	**0.88**	**0.84**	0.84	0.82
TwoPatterns	0.34	0.53	**0.69**	**0.91**	0.66	0.82	0.47	0.71	0.64	0.90	0.68	0.80	0.55	0.72
UWaveGestureLibraryAll	0.31	0.40	**0.62**	0.70	0.56	0.63	0.40	0.54	**0.62**	**0.73**	0.55	0.66	0.44	0.55
FordB	0.58	0.60	0.64	0.67	0.74	0.76	0.64	0.68	0.63	0.64	0.70	0.70	0.61	0.64
ShapesAll	0.29	0.30	0.49	0.55	0.53	**0.60**	0.40	0.48	0.48	0.56	**0.54**	**0.60**	0.40	0.46
SonyAIBORobotSurface1	0.56	0.68	0.54	0.70	**0.61**	**0.74**	0.53	0.70	0.48	**0.74**	0.58	0.71	0.54	0.69
SonyAIBORobotSurface2	**0.81**	**0.86**	0.77	0.79	0.80	0.80	0.80	0.81	0.77	0.85	0.80	0.85	0.80	0.85
Symbols	0.50	0.60	**0.59**	0.60	0.50	**0.66**	**0.59**	**0.66**	0.45	0.61	0.42	0.62	0.43	0.63
Mallat	0.63	0.77	0.72	0.81	0.76	0.83	0.68	0.78	**0.79**	**0.87**	0.77	0.85	0.69	**0.86**

Table 3: Summary of mean linear probe accuracy by SSL Method and Augmentation. Average across 4 seeds. Best result for KNN and SVM are highlighted in bold.

Top 1 linear probe accuracy

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
FordA	0.75	0.78	0.84	0.88	**0.93**	**0.92**	0.85	0.87	0.81	0.77	0.88	0.90	0.86	0.85
ElectricDevices	0.35	0.43	0.36	0.45	0.39	0.43	**0.45**	**0.46**	0.34	0.42	0.39	0.42	0.42	0.45
StarLightCurves	0.89	0.91	0.94	0.95	**0.96**	**0.96**	0.90	0.91	0.95	0.95	0.93	0.95	0.90	0.90
Wafer	0.94	0.89	**0.97**	**0.95**	**0.97**	**0.95**	**0.97**	0.93	**0.97**	**0.95**	**0.97**	**0.95**	**0.97**	0.94
ECG5000	0.83	0.84	0.88	0.86	**0.90**	**0.88**	**0.90**	**0.88**	0.88	0.85	0.89	0.86	0.86	0.85
TwoPatterns	0.37	0.62	**0.75**	**0.96**	0.68	0.85	0.55	0.75	0.70	0.92	0.71	0.81	0.63	0.76
UWaveGestureLibraryAll	0.34	0.43	**0.67**	0.74	0.60	0.67	0.43	0.54	**0.67**	**0.76**	0.58	0.67	0.48	0.58
FordB	0.60	0.63	0.67	0.71	**0.76**	**0.80**	0.69	0.74	0.67	0.65	0.74	0.77	0.63	0.68
ShapesAll	0.33	0.34	0.53	0.59	**0.59**	**0.65**	0.44	0.50	0.50	0.56	0.57	0.63	0.44	0.48
SonyAIBORobotSurface1	0.67	**0.80**	0.61	0.77	0.76	**0.80**	0.60	0.74	0.51	0.79	0.63	0.75	0.63	0.75
SonyAIBORobotSurface2	**0.84**	**0.89**	0.80	0.86	0.82	0.84	0.83	0.82	0.81	0.88	0.81	0.88	0.83	0.87
Symbols	0.56	0.66	0.65	0.69	0.55	0.73	**0.64**	**0.71**	0.51	0.65	0.45	0.67	0.46	0.69
Mallat	0.54	0.88	0.57	0.87	0.74	0.89	0.66	0.80	0.74	**0.92**	0.72	0.88	0.62	**0.90**

Table 4: Summary of max linear probe accuracy by SSL Method and Augmentation. Maximum value across 4 seeds. Best result for KNN and SVM are highlighted in bold.

A clear improvement in probe accuracy is observed with NC-VQVAE compared to the naive VQVAE. Across 12 out of 13 datasets, a configuration of our model performs best, with the only exception showing a negligible one percent difference for both SVM and KNN. The most significant improvements are observed in FordA, FordB, Mallat, ShapesAll, TwoPatterns, and UWaveGestureLibraryAll.

In Figure **??**, we observe that while the choice of augmentation has a substantial impact, all options lead to significantly improved probe accuracy across most datasets. Notably, both SSL methods yield comparable probe accuracies for a given augmentation, underscoring the importance of selecting appropriate augmentations.

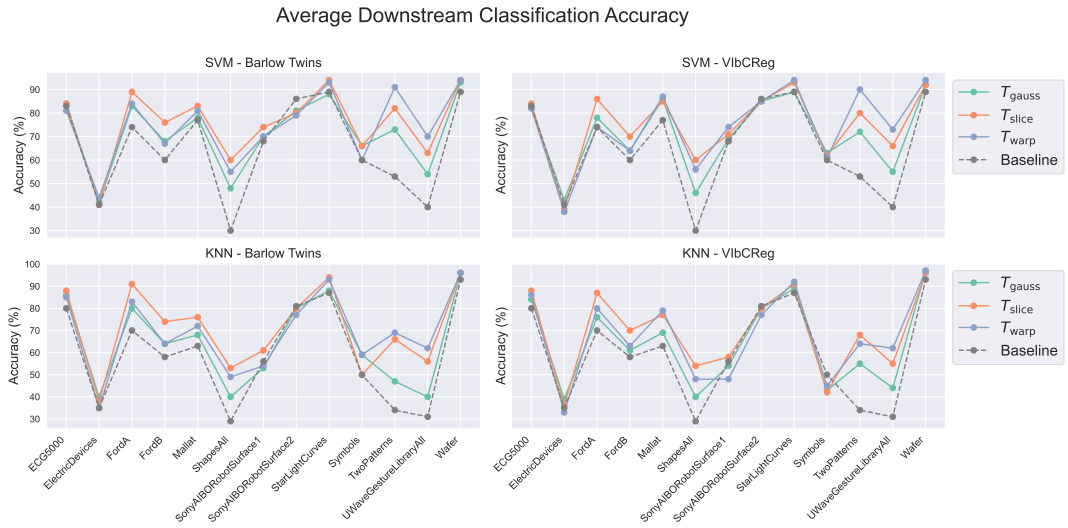


Figure 3: Mean probe accuracies.

Further analysis reveals that Slice and Shuffle, as well as Window Warp and Amplitude Resize, result in the most substantial accuracy gains, whereas Gaussian noise consistently yields less pronounced improvements. We hypothesize that since Slice and Warp often generate augmented views that differ considerably from the original, the SSL loss pushes the representations in different directions, potentially leading to better utilization of the latent space. Visualizations in Figure **??**, **??** and **??** illustrate the effect of NC-VQVAE on the discrete latent representations of FordA, TwoPatterns and UWaveGestureLibraryAll. These visualizations demonstrate that representations learned with NC-VQVAE exhibit greater structure than those of the naive VQVAE, with similar samples, typically with the same label, clustered closer together in latent space. This suggests that the SSL loss introduces semantic information into the latent representations.

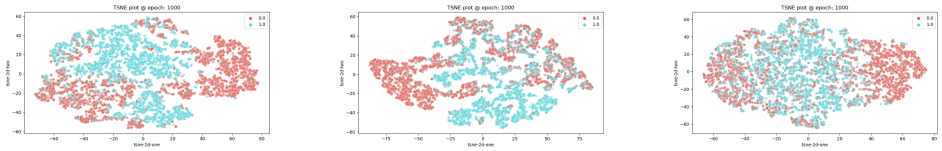


Figure 4: t-SNE plots of FordA. Barlow (left) and VIbCReg (center) with Slice and Shuffle, naive VQVAE (right).

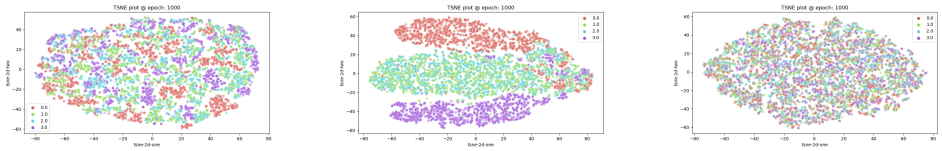


Figure 5: t-SNE plot of discrete latent representations from VIbCReg with Slice and Shuffle (left), Barlow Twins with Window Warp and Amplitude Resize (center) and naive VQVAE (right). Dataset is TwoPatterns.

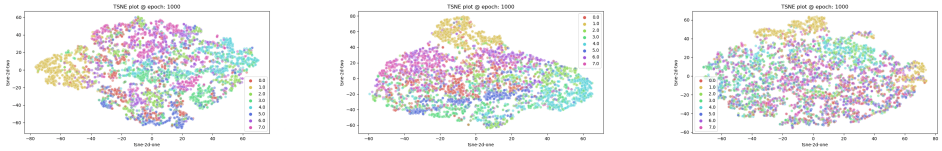


Figure 6: t-SNE plot of discrete latent representations from VIbCReg with Window Warp and Amplitude Resize (left), Barlow Twins with Window Warp and Amplitude Resize (center) and naive VQVAE (right). Dataset is UWaveGestureLibraryAll.

Summary of Stage 1

In summary, the results from stage 1 indicate that NC-VQVAE can reconstruct on par with the naive VQVAE, and in some cases improve the reconstruction loss, while substantially improving probe accuracy for most datasets. From the visual inspection of the discrete latent representations, we observed that NC-VQVAE effectively cluster samples from the same class. Addressing research question 1, we conclude that representations learned with NC-VQVAE are more expressive compared to the naive VQVAE, encoding more class-specific information without compromising reconstruction quality. Regarding research question 2, the choice of augmentation plays a pivotal role in the results, with warp and slice typically outperforming Gaussian augmentation, particularly in terms of probe accuracy. Additionally, significant variations across datasets support the hypothesis that the optimal choice of augmentations is highly dependent on the dataset.

0.1.3 Losses

We investigate some trends in the development of different loss terms during training. Notably, VIbCReg results in more easily minimizable losses, compared to Barlow Twins, and the Gaussian augmentation results in significantly easier minimization of the SSL loss as well as reducing the VQ loss. In figure **??**, we observe the typical pattern of the SSL loss during training.

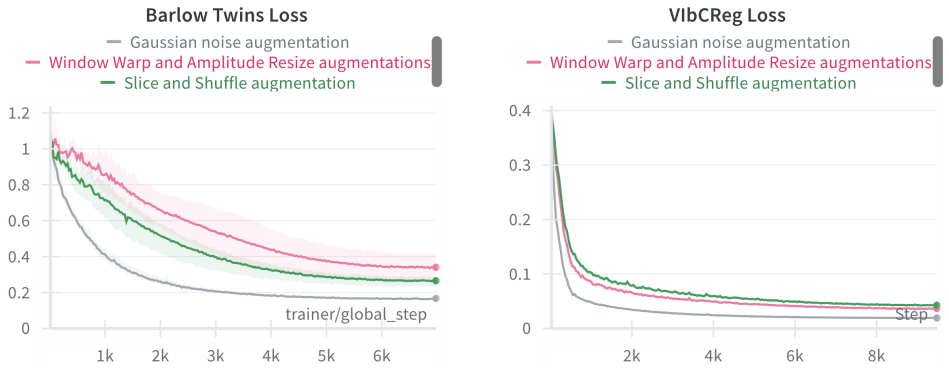


Figure 7: SSL loss during training on UWaveGestureLibraryAll. Averaged across four runs.

The relatively straightforward minimization of the losses might be attributed to the fact that the gaussian augmentation affects the samples in a predictable way. Additionally, we note that the VIbCReg loss declines more rapidly than Barlow Twins across all datasets.

Previously, in figure **??**, we observed that Gaussian augmentation typically resulted in lower probe accuracy than warp and slice. In figure **??** we see that, for datasets where probe accuracy increased significantly, the augmentations that result in a more challenging SSL loss typically has higher downstream classification accuracy. Furthermore, we observe relatively stable patterns for specific augmentations, suggesting their significant impact on probe accuracy compared to variation in SSL loss.

Reconstruction losses during training are consistently minimized across models and augmentations, with augmented reconstruction loss slightly higher in models utilizing Slice and Shuffle. Differences in VQ loss primarily stem from the codebook loss, with VIbCReg consistently exhibiting more effective minimization than Barlow Twins. Gaussian augmentation present the easiest minimization challenge, followed by Window Warp and Slice and Shuffle.

Both VIbCReg and Barlow Twins, when coupled with Gaussian augmentation, consistently perform comparably to the naive VQVAE in terms of VQ loss during

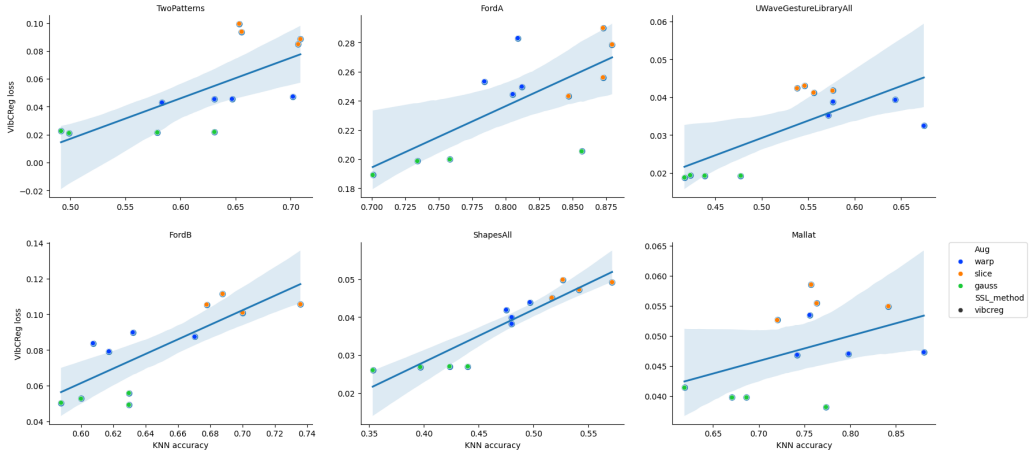


Figure 8: KNN accuracy plotted against VIbCReg loss. Each point correspond to a single run of the model. Similar tendency is shown for Barlow Twins.

training. The minimization of the codebook loss indicates that the encoder is properly aligned with the discrete latent codes. We hypothesize that when the SSL loss is not properly minimized, the encoder must adjust its weights more throughout training which keeps the encoder outputs and the discrete codes from aligning completely.

0.2 Stage 2

The generative quality of our models is assessed using FID, IS, and CAS, with all results presented in this section. However, it's important to note that for datasets with limited samples or few samples per class, the generative scores should be interpreted cautiously. Both the classifier and evaluation metrics rely on a sufficient number of samples to ensure reliability. Therefore, visual inspection is prioritized for such cases.

0.2.1 FID and IS

Tables **??** and **??** present the top 1 and mean scores across the four runs for both FID and IS. Analysis of the tables reveals that our model achieves higher IS scores for 12 out of 13 datasets and lower FID scores for 10 out of 13 datasets in both mean and top 1.

Top 1 FID and IS

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑	FID↓	IS↑
FordA	2.59	1.30	1.93	**1.51**	2.13	1.48	1.80	**1.51**	2.83	1.38	2.50	1.43	**1.66**	1.41
ElectricDevices	12.05	3.97	11.82	4.20	**8.91**	4.07	9.89	3.86	12.38	**4.23**	11.08	3.94	13.96	3.71
StarLightCurves	**0.74**	1.99	0.89	**2.43**	1.50	2.36	0.75	2.39	0.92	2.39	0.85	**2.40**	0.79	2.26
Wafer	5.27	**1.39**	3.31	1.29	3.82	**1.26**	2.77	1.35	3.33	1.29	3.60	1.30	**2.52**	1.34
ECG5000	1.56	2.01	2.43	2.02	2.27	2.00	2.15	2.02	2.15	**2.03**	2.21	2.00	**1.52**	2.02
TwoPatterns	3.63	2.47	3.59	2.65	2.74	2.73	**2.24**	2.70	3.45	2.64	2.90	2.70	**2.19**	**2.77**
UWaveGestureLibraryAll	8.16	2.24	6.45	2.94	**6.26**	**3.13**	7.31	2.79	6.52	2.99	6.33	3.06	7.09	2.79
FordB	2.92	1.52	2.10	1.52	2.44	1.61	1.93	**1.67**	1.76	1.65	2.12	1.64	**1.66**	1.52
ShapesAll	**21.35**	4.32	35.89	**5.22**	29.61	5.16	27.91	4.83	30.03	4.95	31.59	4.92	27.20	4.94
SonyAIBORobotSurface1	18.21	1.27	26.20	1.32	28.90	1.28	21.63	1.32	21.98	1.36	25.20	1.38	**15.73**	**1.55**
SonyAIBORobotSurface2	3.85	1.69	2.50	1.82	3.34	1.79	**0.82**	1.82	2.61	1.81	2.75	1.83	1.24	**1.84**
Symbols	8.50	2.43	5.86	3.20	7.39	2.82	**4.25**	**3.50**	6.78	3.39	7.21	3.23	8.21	3.30
Mallat	**1.31**	3.41	2.01	3.67	2.24	3.72	1.85	3.66	1.87	3.34	2.30	3.05	**1.31**	**3.92**

Table 5: Summary of FID and IS scores by SSL Method and Augmentation. Best achieved results are highlighted in bold

Mean FID and IS

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑	FID↓	IS↑
FordA	5.15	1.16	2.59	1.41	2.36	**1.45**	**2.28**	**1.45**	3.01	1.34	2.90	1.41	3.73	1.29
ElectricDevices	13.48	3.75	16.51	3.95	**10.20**	3.93	11.54	3.75	13.99	**4.17**	11.82	3.85	15.20	3.55
StarLightCurves	**1.01**	1.93	1.29	2.35	1.91	2.32	1.08	2.25	1.07	2.35	1.19	**2.36**	1.05	2.22
Wafer	5.72	**1.33**	3.70	1.25	4.20	1.24	2.85	1.31	3.67	1.26	3.86	1.26	**2.84**	1.31
ECG5000	**1.62**	1.94	2.61	**2.00**	2.56	1.98	2.47	**2.00**	2.60	1.99	2.39	**2.00**	1.76	1.99
TwoPatterns	4.04	2.41	4.00	2.54	2.96	2.66	**2.44**	**2.67**	4.05	2.56	3.15	2.66	2.62	**2.67**
UWaveGestureLibraryAll	8.48	2.13	6.77	2.86	6.64	2.96	7.35	2.73	6.80	2.91	**6.49**	**2.99**	7.34	2.72
FordB	4.05	1.28	2.66	1.48	3.49	1.50	2.88	**1.52**	**2.49**	1.48	3.07	1.51	3.04	1.31
ShapesAll	**27.64**	4.22	38.22	5.07	32.54	**5.04**	32.25	4.56	36.59	4.72	35.79	4.76	31.56	4.71
SonyAIBORobotSurface1	23.71	1.20	30.65	1.22	31.97	1.21	25.29	1.28	26.11	1.32	28.20	1.32	**18.61**	**1.44**
SonyAIBORobotSurface2	5.42	1.62	3.35	1.77	4.41	1.74	**1.78**	**1.81**	4.43	1.74	3.32	1.79	2.36	1.79
Symbols	13.62	1.99	9.78	2.92	9.78	2.67	8.61	3.14	8.84	3.20	9.74	3.03	**8.58**	**3.24**
Mallat	2.09	3.01	2.54	3.29	3.68	2.94	2.12	3.53	2.11	3.18	2.40	2.96	**1.65**	**3.72**

Table 6: Summary of FID and IS scores by SSL Method and Augmentation. Best mean achieved FID and IS are highlighted in bold

In Figure **??**, an overview of the results highlights that both Barlow Twins and VIbCReg generally yield better samples than the naive VQVAE in terms of FID and IS. There is a difference in VIbCReg and Barlow Twins, in that VIbCReg is more robust to the particular choice of augmentation, which is particularly evident when looking at the Slice and Shuffle augmentation. Furthermore, the use of Gaussian augmentation leads to the most significant improvements across most datasets. The high IS scores suggest that NC-VQVAE captures the conditional distributions more accurately than the naive VQVAE. The improved FID scores indicate that the synthetic samples more closely resemble the test data. This will be explored further in Section **??**. The moderate decrease in FID, compared to the increase in IS, could indicate that the generated samples does not generalize too well to the test data. However, the discrete latent representations from NC-VQVAE offer additional class-specific information, as evidenced by the improved downstream

classification accuracy observed in stage 1. This supplementary class-specific information appears to assist the prior learning process in capturing class conditional distributions.

Figure 9: Mean FID and IS scores for Barlow Twins and VIbCReg VQVAE. FID is plotted on a log scale because of the large difference in values across datasets.

It is important to note that the FID and IS scores are calculated using the SupervisedFCN, which is trained on the UCR Archive. Consequently, there may be a bias towards samples that mimic the training data.

0.2.2 CAS

We present the mean CAS for all models across datasets in Table **??**.

Mean CAS

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.864	0.884	**0.902**	0.878	0.864	0.895	0.870
ElectricDevices	0.614	0.588	0.607	0.599	**0.618**	0.610	0.594
StarLightCurves	0.960	0.953	0.955	**0.965**	0.962	0.954	0.964
Wafer	0.976	0.977	0.978	0.968	0.979	0.976	**0.984**
ECG5000	0.866	0.881	0.863	0.880	0.877	0.892	**0.910**
TwoPatterns	0.808	0.770	0.788	**0.847**	0.715	0.781	0.846
UWaveGestureLibraryAll	0.333	0.300	0.367	0.313	0.360	**0.401**	0.383
FordB	0.725	0.748	**0.756**	0.741	0.750	0.738	0.750
ShapesAll	0.361	0.344	0.329	**0.420**	0.379	0.367	0.404
SonyAIBORobotSurface1	0.975	0.933	0.957	0.979	0.982	0.976	**0.985**
SonyAIBORobotSurface2	0.929	0.956	0.951	0.969	0.960	**0.970**	0.964
Symbols	0.956	0.929	0.930	0.930	0.969	**0.974**	0.963
Mallat	0.471	0.642	0.563	0.661	0.827	0.876	**0.908**

Table 7: Mean CAS score across datasets. Results averaged across four runs.

We observe that some configuration of NC-VQVAE outperform the naive VQVAE on all datasets. Additionally, VIbCReg with gaussian augmentation outperforms the baseline on 12 out of 13 datasets. Generally the difference in CAS is small for the different models, with the exception of Mallat, where VIbCReg with Gaussian augmentaion leads to an improvement of 0.437 compared to baseline. This spesific case will be investigated further in Section **??**.

0.2.3 Prior Loss

In Figure **??**, we illustrate the evolution of the validation prior loss during training. We observe that naive VQVAE outperforms all configurations of NC-VQVAE, and this pattern observed is representative for most datasets. The exception being SonyAIBORobotSurface1 and 2, where Barlow Twins with warp augmentation exhibits slightly better performance. These observations suggest that minimizing the validation prior loss does not correlate with improved generative scores.

0.2.4 The influence of stage 1 on stage 2

We aim to address research questions 3 and 4, which focus on how expressive representations influence the quality of synthetic samples and the role of augmentations. To explore this, we examine the relationship between probe accuracy and FID and IS metrics. Specifically, we focus on datasets where NC-VQVAE demonstrated a significant increase in probe accuracy, namely FordA, FordB, Mallat, ShapesAll, TwoPatterns, and UWaveGestureLibraryAll.

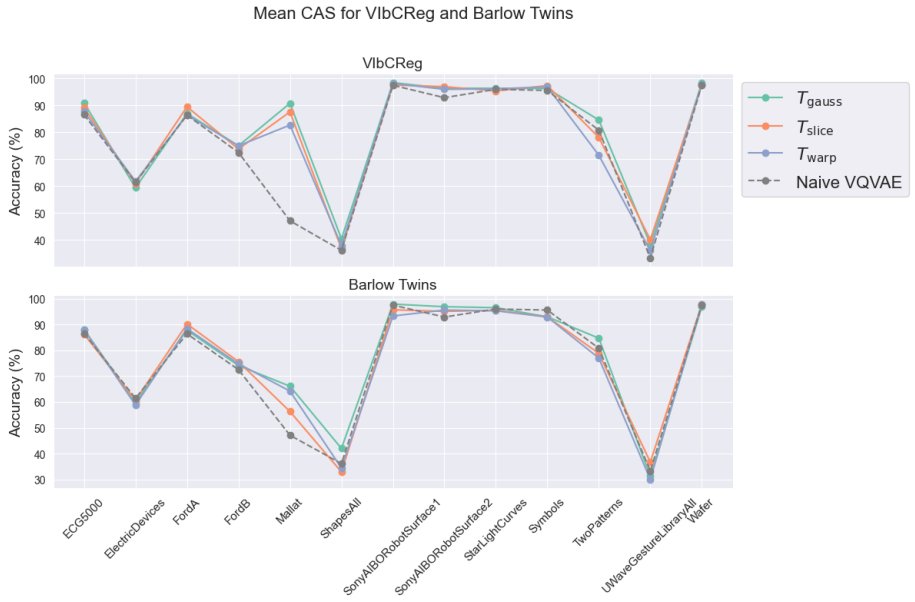


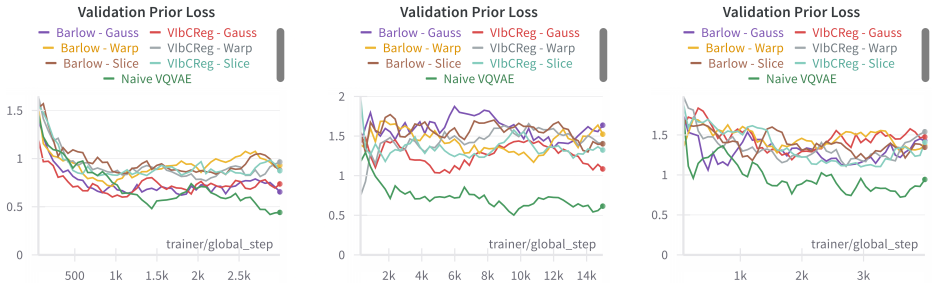
Figure 10: Mean CAS across all datasets.

In Figure **??** and **??**, we present scatterplots of KNN against FID and IS, respectively, along with the least square regression line. Similar trends are observed in the corresponding plots with SVM accuracy. From Figure **??**, we observe a trend, where higher probe accuracy correlates with higher IS. Upon closer inspection, we see a pattern of the prominent effect of augmentations. For each specific augmentation, the correlation between KNN and IS is close to 0. It appears that augmentations leading to higher KNN accuracy tend to result in higher IS scores, though the pattern is not consistent across all datasets. In Figure **??**, we notice that the specific augmentation serves as a better indicator of FID score than KNN accuracy. Both figures also illustrate the sensitivity of model performance to initialization, particularly in terms of probe accuracy.

0.2.5 Visual inspection

In the following we present generated samples from naive VQVAE and NC-VQVAE for selected datasets. The ground truth, both test and train, are additionally included in order to better make sense of the IS, FID and CAS scores.

Some datasets, such as FordA and FordB, are poorly suited for this type of visual inspection, as illustrated previously in Figure **??**. As a result, the selection of datasets is primarily based on how well they lend themselves to this type of presentation. For each figure in the following sections, 50 samples are generated from each model. For the ground truth, we plot a subset of 50 randomly selec-



ShapesAll FordA TwoPatterns

Figure 11: Validation prior loss during training for selected datasets. Averaged across all runs. Results averaged using exponential moving average with decay 0.7.

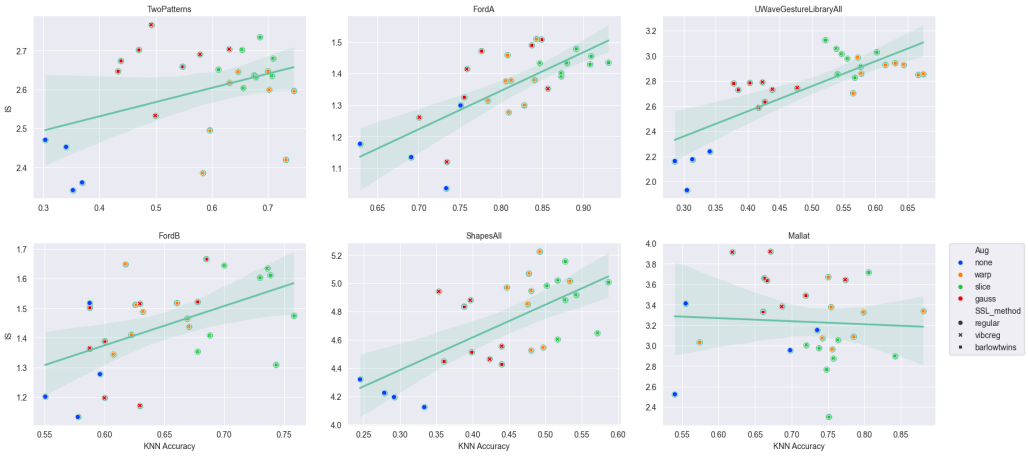


Figure 12: KNN plotted against Inception Score on the subset of datasets with significant improvement in probe accuracy.

ted samples, or the entire set if the dataset contain less than 50 samples. For the datasets with complex distribution, or many classes, it is very difficult to visually asses the unconditional distribution. Thus, we mainly provide class conditional samples. We surprisingly only observe minor differences in the generated samples from NC-VQVAE trained with different augmentations.

Typically naive VQVAE has more difficulty with capturing the global consistency of the samples when training samples are scarce and diverse, as seen on ShapesAll and Symbols. In contrast, our method will tend to overfit in these cases. The overfitting issue is most prominent in the class conditional distributions.

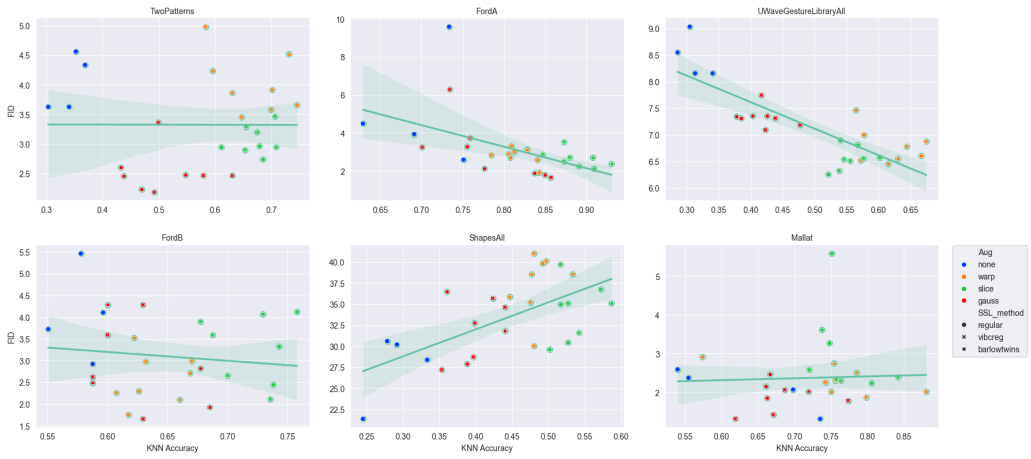


Figure 13: KNN plotted against Fréchet Inception Distance on the subset of datasets with significant improvement in probe accuracy.

ECG5000

In Figure **??** we present generated samples from naive VQVAE and NC-VQVAE trained with Window Warp and Amplitude Resize augmentations on ECG5000.

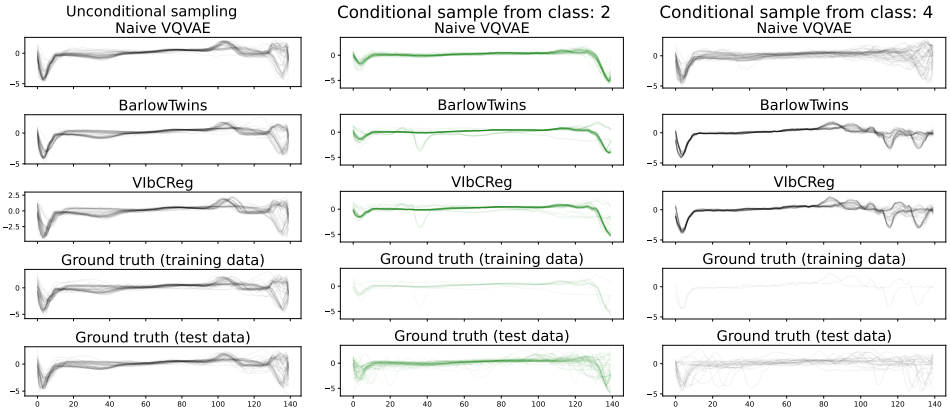


Figure 14: Dataset: ECG5000. Barlow and VIbCReg both trained with window warp and amplitude resize augmentations. 50 samples from each model.

We see some evidence that VIbCReg maintains more variability than Barlow Twins, while both have good mode coverage. In class 4, where the training data only consists of 2 samples, both Barlow Twins and VIbCReg catches the pattern, while producing some variation which resemble the training samples. The naive VQVAE samples does not capture this distribution, but when compared to the test data, it is more similar. This could explain the minor increase in CAS for NC-VQVAE. Looking at the unconditional sample, it is not evident why naive VQVAE

performs better in terms of FID score than NC-VQVAE.

Mallat

Mallat is a simulated dataset, where the classes have very little variability and training and test distribution are almost indistinguishable, except for sample size.

We observe that VIbCReg is superior in capturing the variability, compared to Barlow Twins and naive VQVAE. This is most evident in the first 300 timesteps of class 5 in Figure **??**. Looking as class 7, we see Barlow Twins completely collapsing, essentially producing the same sample over and over. These figures explain the significant increase in CAS seen in Figure **??**, particularly for VIbCReg. It also explains why VIbCReg with Gaussian augmentation both increases IS and reduces FID.

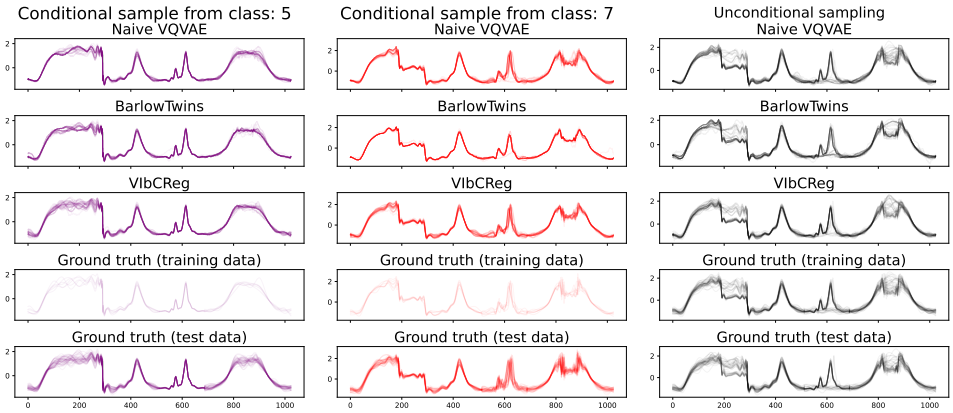


Figure 15: Class conditional distribution for some selected classes of Mallat, in addition to unconditional samples. Barlow and VIbCReg both trained with Gaussian augmentation.

By inspecting the PCA plots of both data space and the discrete latent representations of samples from Mallat, compared to synthetic samples form VIbCReg and Barlow Twins in Figure **??**, we see a clear case of representation collapse for Barlow Twins. We hypothesize that the variance term in VIbCReg assists in maintaining variability in the representations. Making it more effective in avoiding this type of collapse.

Symbols

The Symbols dataset consists of several distinct, but simple, patterns. The dataset is very small, where each class contsists of less than 5 training samples.

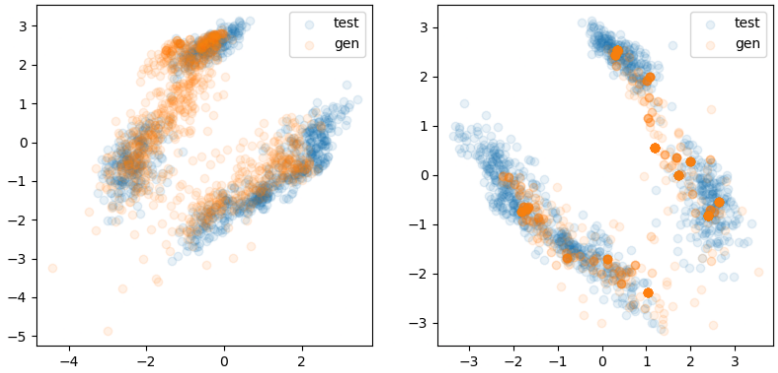


Figure 16: PCA of discrete latent representation from Mallat. Both VIbCReg (left) and Barlow Twins (right) are trained with gaussian augmentation.

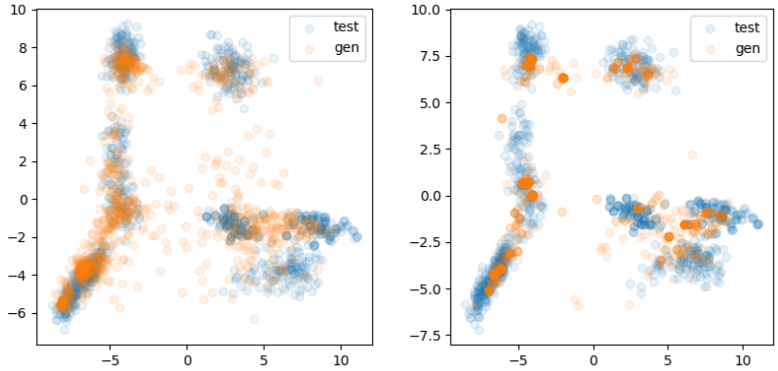


Figure 17: PCA of generated time series from Mallat. Both VIbCReg (left) and Barlow Twins (right) are trained with gaussian augmentation.

In Figure **??**, particularly the unconditional sample, we see that naive VQVAE does not capture the entire underlying distribution, some classes are not represented/not recognizable, while global consistency for the sinusoids are poor, particularly towards the end. In class 3, we observe that both Barlow Twins and VIbCReg mimic the training data to a large degree. At first glance the naive VQVAE looks to produce the most desirable distribution, but upon closer inspection we see an excessive amount of noise and lack of consistency.

The IS on Symbols, for both VIbCReg and BarlowTwins, is substantially higher than naive VQVAE. While this is not surprising after inspecting the samples, it exposes an issue with the IS metric. It fails to take intraclass diversity into account, and is therefore oblivious to overfitting.

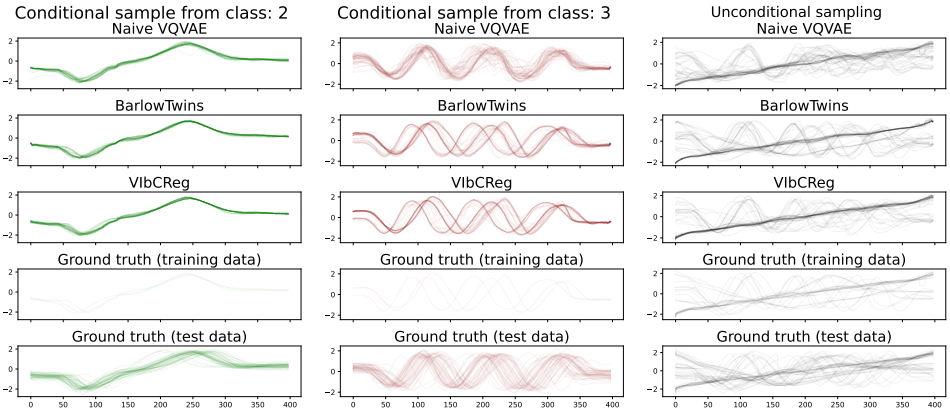


Figure 18: Class conditional distribution for some selected classes of Symbols. Barlow and VIbCReg both trained with gaussian augmentation.

ShapesALL

The dataset ShapesAll consists of 60 classes, with 10 samples within each class. Each class has distinct patterns, with varying complexity.

In Figure **??**, we observe clearly that naive VQVAE struggles with capturing the global consistency of the samples. We too observe that Barlow Twins mimic the training slightly more closely than VIbCReg, which provides some insight as to why Barlow Twins improves CAS by about 10 percent compared to naive VQVAE. Both Barlow Twins and VIbCReg improve IS, but fail to improve FID. By inspecting the samples, it is not evident why NC-VQVAE fails to improve FID. As FID is calculated from unconditional samples, the issue is most likely due to a issue not observable from the conditional samples. As the dataset has 60 classes, it is probable that an insufficient number of synthetic samples were generated to represent the large variability in the unconditional distribution.

In Symbols and ECG5000 we saw that NC-VQVAE can overfit when there are very few samples in a class. Reassuringly, the mode coverage is quite good, and similar tendencies are observed for ShapesAll as well.

UWaveGestureLibraryAll

The dataset UWaveGestureLibraryAll contains time series with distinct discontinuities and sharp changes in modularity. As noted in [**TimeVQVAE**], such datasets are challenging to model.

In Figure **??** a selected subset of classes are illustrated. We observe upon close inspection that VIbCReg maintains variability in the samples to a greater degree than Barlow Twins, as well at slightly better capturing the dead-spots following

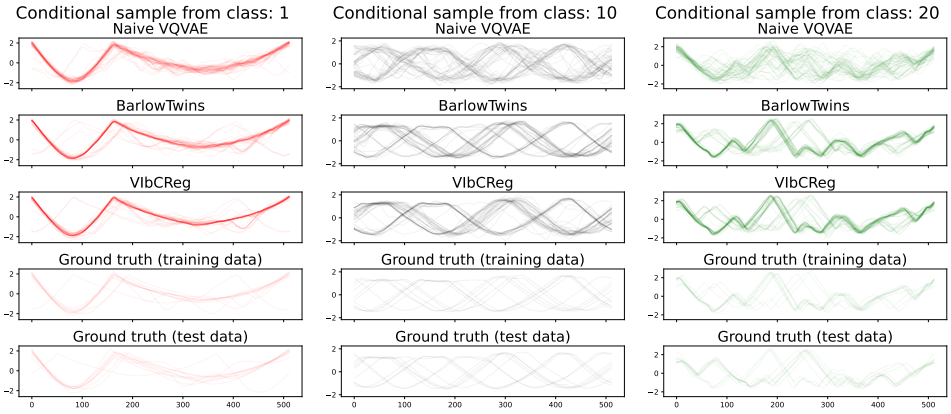


Figure 19: Class conditional distribution for some selected classes of ShapesAll. Barlow and VIbCReg both trained with gaussian augmentation.

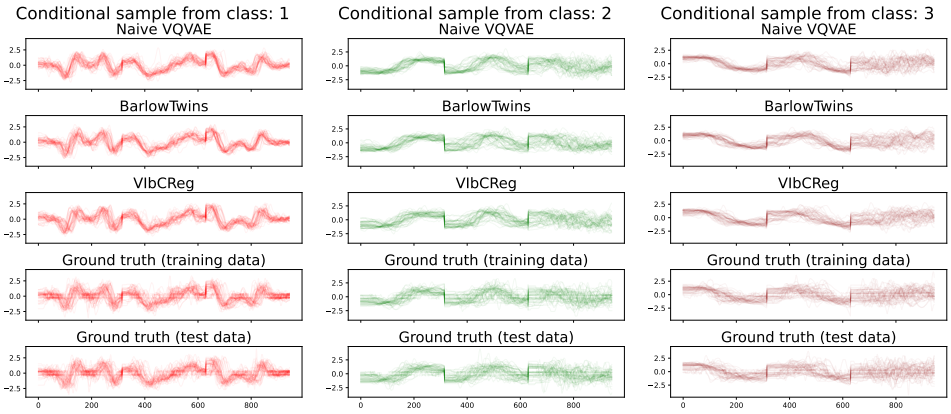
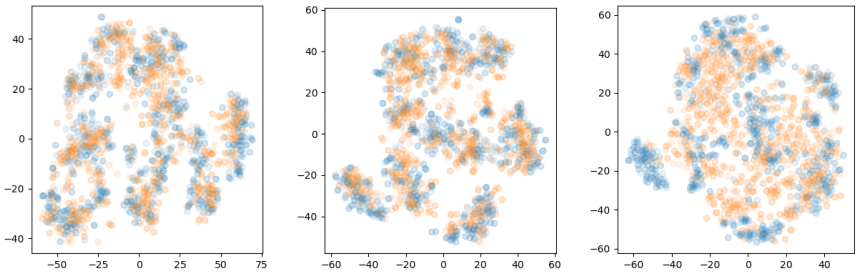


Figure 20: Class conditional distribution for some selected classes of UWaveGestureLibraryAll. Barlow and VIbCReg both trained with window warp and amplitude resize augmentations.

the discontinuities. By investigating the t-SNE plots in Figure **??** it becomes more evident that NC-VQVAE captures the distribution better than naive VQVAE.

Summary of Stage 2

In summary, the results from stage 2 suggest that NC-VQVAE offers advantages over the naive VQVAE in capturing conditional distributions, as indicated by higher IS and CAS scores. Additionally, NC-VQVAE demonstrates improved sample quality compared to ground truth data, with lower FID scores. Visual inspections further reveal that NC-VQVAE achieves better mode coverage and captures global sample consistency to a greater extent than its naive counterpart. Additionally, from the visual inspection we observe that the quality of the generated samples often are much higher, despite a moderate decrease in FID.



(a) VIbCReg **(b)** Barlow Twins **(c)** Naive VQVAE

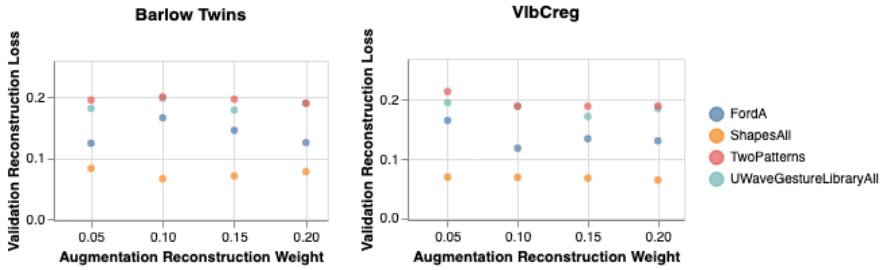
Figure 21: t-SNE of generated (orange) and test data (blue). Dataset: UWaveGestureLibraryAll

Addressing research question 3, we conclude that the expressive representations learned from NC-VQVAE contribute to learning class-specific details and enhancing the quality of synthetic samples. However, it's worth noting that NC-VQVAE is prone to overfitting when faced with small datasets or classes with few samples, whereas the naive VQVAE struggle to capture global consistencies effectively.

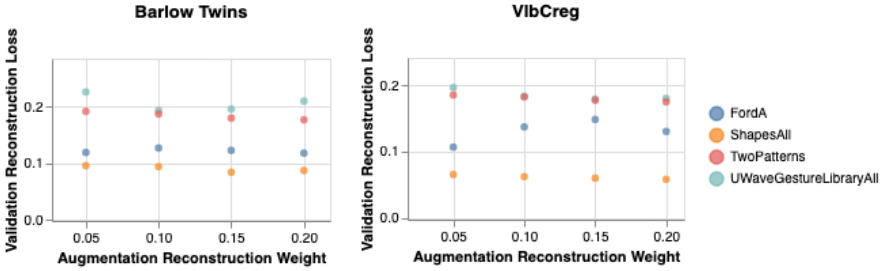
Regarding research question 4, we observe that the generative performance is less sensitive to the choice of augmentations compared to the downstream classification accuracy observed in stage 1. Nonetheless, Gaussian noise yields the least variability in performance compared to the baseline, outperforming the naive VQVAE consistently in terms of IS, FID, and CAS metrics.

0.3 Ablation Study

During the development of our model, we investigated the effect of the augmentation reconstruction weight, ζ, both on validation reconstruction and downstream classification. This experiment was carried out on a small scale, focusing on two sets of augmentations: Window Warp and Amplitude Resize, and Slice and Shuffle. We conducted our experiment on a subset of the UCR Archive consisting of FordA, ShapesAll, TwoPatterns and UWaveGestureLibraryAll. The tokenization model underwent training for 250 epochs, with all other parameters maintained as in the main experiment setup. We explored the effect of ζ across a range of values, specifically $\zeta \in \{0.05, 0.10, 0.15, 0.20\}$. The influence on reconstruction is visualized in Figure **??**, while the impact on KNN and SVM accuracy is presented in Figure **??** and **??**, respectively.

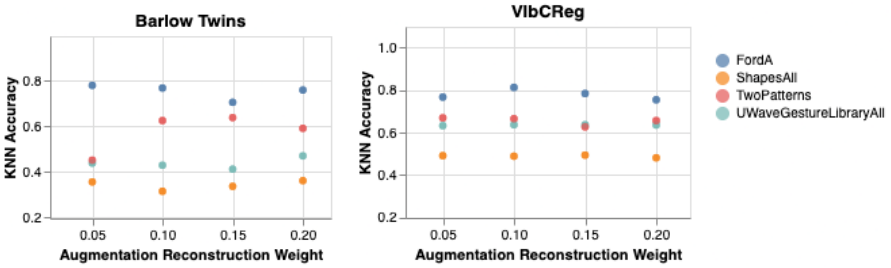


Augmentation: Window Warp and Amplitude Resize.

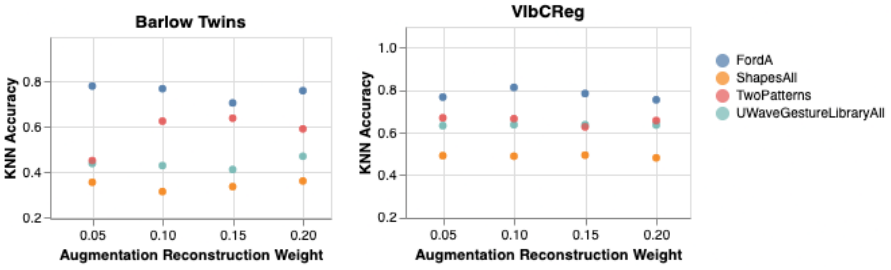


Augmentation: Slice and Shuffle.

Figure 22: The effect of the augmentation reconstruction weight on validation reconstruction. Results averaged across 2 runs.

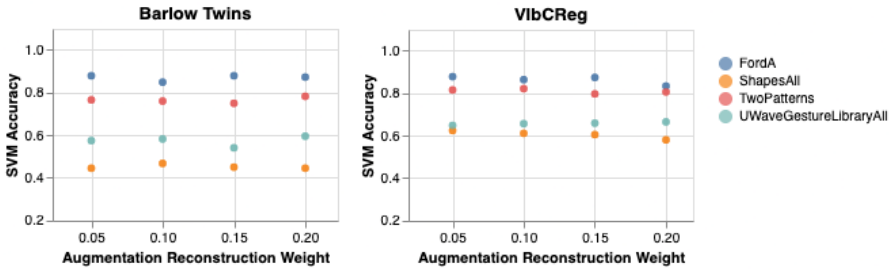


Augmentation: Window Warp and Amplitude Resize.

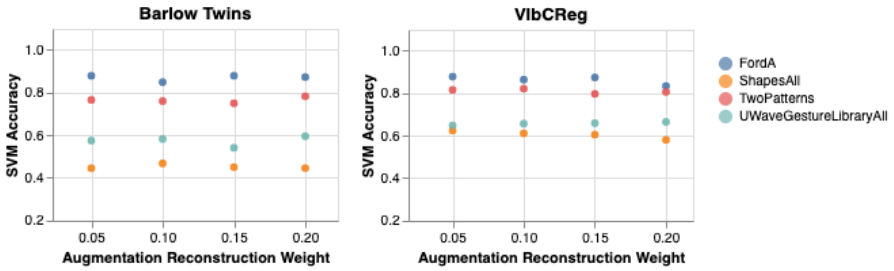


Augmentation: Slice and Shuffle.

Figure 23: The effect of the augmentation reconstruction weight on KNN accuracy. Results averaged across 2 runs.



Augmentation: Window Warp and Amplitude Resize.



Augmentation: Slice and Shuffle.

Figure 24: The effect of the augmentation reconstruction weight on SVM accuracy. Results averaged across 2 runs.

Observing Figure **??**, we note that the reconstruction loss remains relatively stable across most datasets, exhibiting robustness to variations in the augmentation reconstruction weight. However, FordA displays some degree of variability, and the variability is augmentation dependent. This suggests that when optimization of model performance could benefit from jointly optimizing choice of augmentation and augmentation reconstruction weight.

Examining Figure **??** and **??**, we observe that augmentation reconstruction weight affects KNN to a larger degree than SVM, indicating that the weight introduces differences in local structure of the discrete latent space, while the global structure is less affected. Furthermore, it's evident that VIbCReg consistently produces high probe accuracies and demonstrates greater robustness compared to Barlow Twins across different augmentation reconstruction weights.

0.4 Discussion

Difficulty in assessing generative samples

The difficulty in visually assessing generative performance for datasets with complicated distributions is still highly relevant. One would like to be able to confidently rely on the evaluation metrics when assessing the model performance, but as observed, the generative metics are flawed. We observe a discrepancy in the

improvement in generative scores and the quality observed in the visual inspection.

Many datasets in the UCR Archive and our selected subset is heavily imbalanced. As the SupervisedFCN used to evaluate FID and IS is trained on these imbalanced datasets, they may underrepresented the quality of synthetic samples from the minority classes. We have from the visual inspection seen that NC-VQVAE is superior to naive VQVAE in its ability to capture the conditional distributions, even then the sample size is very small. Additionally, the SupervisedFCN may be better at recognizing and evaluating features of the dominant classes and worse at recognizing features of minority classes. This can lead to biased scores that favor synthetic samples resembling the overrepresented classes.

The current metrics are dependent on high quality data with large sample sizes, balanced classes etc. There is currently a lack of large high quality datasets in the time series domain, and something akin to ImageNet for time series would be beneficial for the community.

The role of augmentations

We observe that both probe accuracies and generative scores are dependent on the particular choice of augmentations. The probe accuracies are generally better when augmentation are clearly different from the original view, as for warp and slice, while the generative scores has a tendency to follow the opposite pattern, though not as clearly. Additionally, we have observed that the optimal choice of augmentations is highly dataset dependent, and a systematic analysis of different augmentations on specific datasets is needed to understand the dynamics better.

Temporal vs frequency influence of augmentations

The Gaussian augmentation introduces a HF component, though its influence in the time domain is visually clear, its effect on the spectrograms is minor, as the LF components typically has much larger magnitude. Window Warp and Amplitude Resize mainly alter LF components, often changing the exact location of dominant frequencies on the time axis, but not their order. The effect of Slice and Shuffle is variable, but has a tendency to create sharp discontinuities, which in many cases introduce a significant HF component.

We assessed and chose our different sets of augmentations based on their effect on the temporal representation of the time series. As we model the time frequency domain, future work should more thoroughly investigate the effects of augmentation on spectrograms. Additionally, all models considered compress the input only along temporal axis in the encoder, which in a sens puts more emphasis in the frequency components rather than the exact temporal structure.

Differences in Barlow Twins and VIbCReg

From the visual inspections we observed that VIbCReg maintain the variability in the generated samples a bit better than Barlow Twins. We hypothesize that it is due to the variance term present in VIbCReg, and wonder whether increasing its weight might assist in producing more diverse samples.

In terms of the generative scores, VIbCReg is more robust to choice of augmentations, highlighted by the their different response to the Slice and Shuffle augmentation. This can likely be attributed the difference in variance/covariance regularization on the two models. Since VIbCReg regularizes each branch separately, it responds better to situations where the input of each branch differs significantly.

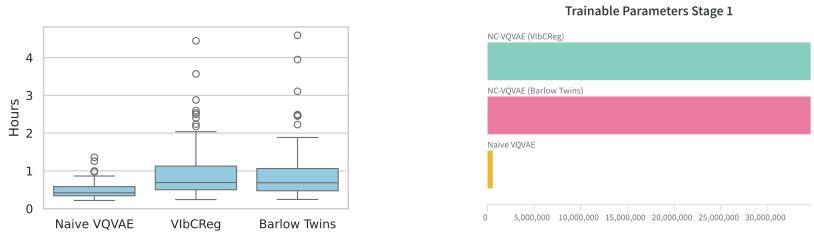
VIbCReg generally demonstrated slightly better performance in terms of validation reconstruction than Barlow Twins. In the ablation study we observed that VIbCReg was more robust to augmentation reconstruction weight, both in terms of the validation reconstruction and downstream classification.

Overfitting problem

We have observed on multiple occasions that when the sample size is small and the patterns in the data are simple NC-VQVAE has a tendency to overfit, and memorize the training data to a large degree. In our experiments we have not employed any form of dropout or other regularization techniques when training on small datasets, therefore this issue is to be expected. As the naive VQVAE in the same cases fails to capture the global consistency, we see our model as a step in the right direction, but stress the need for regularization. Additionally, investigating prior learning models and sampling procedures tailored for the more expressive representations, leveraging the encoded semantic information, could be beneficial.

Training Time and Model Size

The addition of the self-supervised loss in NC-VQVAE comes at a cost in terms of model size, and consequently training time. The added complexity is a result of the high dimensionality of the projector used in both Barlow Twins and VIbCReg. As the projector is discarded after stage 1, the increased complexity only affect training of the tokenization model. Examining Figure **??**, we observe a significant increase in training time and trainable parameters.



(a) Training time distributions for Stage 1.

(b) Number of parameters in Stage 1 for models trained on Wafer.

Figure 25: Overview of training time and model size for Stage 1.