References

[1] Denis Agniel, Isaac S Kohane, and Griffin M Weber. 2018. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj* 361 (2018).

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).

[3] Raquel Aoki, Frederick Tung, and Gabriel L Oliveira. 2022. Heterogeneous multi-task learning with expert diversity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 6 (2022), 3093–3102.

[4] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. 2018. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075* (2018).

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854.

[6] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems* 31 (2018).

[7] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 6085.

[8] Xiwen Chen, Peijie Qiu, Wenhui Zhu, Huayu Li, Hao Wang, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. 2024. TimeMIL: Advancing Multivariate Time Series Classification via a Time-aware Multiple Instance Learning. *arXiv preprint arXiv:2405.03140* (2024).

[9] MinGyu Choi and Changhee Lee. 2023. Conditional Information Bottleneck Approach for Time Series Imputation. In *The Twelfth International Conference on Learning Representations*.

[10] Andrea Cini, Ivan Marisca, and Cesare Alippi. 2021. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. *arXiv preprint arXiv:2108.00298* (2021).

[11] Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2024. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[12] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica* 6, 6 (2019), 1293–1305.

[13] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. 2019. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems* 32 (2019).

[14] Wenjie Du, David Côté, and Yan Liu. 2023. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications* 219 (2023), 119619.

[15] Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. 2024. Dynamic Multimodal Information Bottleneck for Multimodality Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 7696–7706.

[16] Ian Fischer. 2020. The conditional entropy bottleneck. *Entropy* 22, 9 (2020), 999.

[17] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. 2020. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*. PMLR, 1651–1661.

[18] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).

[19] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34, 6 (2020), 1936–1962.

[20] SeungHyun Kim, Hyunsu Kim, Eunggu Yun, Hwangrae Lee, Jaehun Lee, and Juho Lee. 2023. Probabilistic imputation for time-series classification with missing data. In *International Conference on Machine Learning*. PMLR, 16654–16667.

[21] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[22] Namkyeong Lee, Dongmin Hyun, Gyoung S Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. 2023. Conditional graph information bottleneck for molecular relational learning. In *International Conference on Machine Learning*. PMLR, 18852–18871.

[23] Zhe Li, Zhongwen Rao, Lujia Pan, and Zenglin Xu. 2023. Mts-mixers: Multivariate time series forecasting via factorized temporal and channel mixing. *arXiv preprint arXiv:2302.04501* (2023).

[24] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2024. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems* 36 (2024).

[25] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

[26] Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. 2024. TimeX++: Learning Time-Series Explanations with Information Bottleneck. *arXiv preprint arXiv:2405.09308* (2024).

[27] Donghao Luo and Xue Wang. 2024. Moderntcn: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*.

[28] Qianli Ma, Sen Li, and Garrison W Cottrell. 2020. Adversarial joint-learning recurrent neural network for incomplete time series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 1765–1776.

[29] Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia* 25 (2022), 4121–4134.

[30] Xiaoye Miao, Yangyang Wu, Jun Wang, Yunjun Gao, Xudong Mao, and Jianwei Yin. 2021. Generative semi-supervised learning for multivariate time series imputation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 8983–8991.

[31] Kohei Miyaguchi, Takayuki Katsuki, Akira Koseki, and Toshiya Iwamori. 2022. Variational inference for discriminative learning with generative modeling of feature incompletion. In *International Conference on Learning Representations*.

[32] Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I Webb, Germain Forestier, and Mahsa Salehi. 2024. Deep learning for time series classification and extrinsic regression: A current survey. *Comput. Surveys* 56, 9 (2024), 1–45.

[33] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[35] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*. IEEE, 108–109.

[36] Patrick Schäfer. 2015. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery* 29 (2015), 1505–1530.

[37] Sangwoo Seo, Sungwon Kim, Jihyeong Jung, Yoonho Lee, and Chanyoung Park. 2024. Self-Explainable Temporal Graph Networks based on Graph Information Bottleneck. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2572–2583.

[38] Kamran Shaukat, Talha Mahboob Alam, Suhuai Luo, Shakir Shabbir, Ibrahim A Hameed, Jiaming Li, Syed Konain Abbas, and Umair Javed. 2021. A review of time-series anomaly detection techniques: A step to future perspectives. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 1*. Springer, 865–877.

[39] Satya Narayan Shukla and Benjamin M Marlin. 2021. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318* (2021).

[40] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.

[42] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).

[43] Denis Ullmann, Olga Taran, and Slava Voloshynovskiy. 2023. Multivariate Time Series Information Bottleneck. *Entropy* 25, 5 (2023), 831.

[44] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[45] Slava Voloshynovskiy, Mouad Kondah, Shideh Rezaeifar, Olga Taran, Taras Holotyak, and Danilo Jimenez Rezende. 2019. Information bottleneck through variational glasses. *arXiv preprint arXiv:1912.00830* (2019).

[46] Jun Wang, Wenjie Du, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. 2024. Deep learning for multivariate time series imputation: A survey. *arXiv preprint arXiv:2402.04059* (2024).

[47] Xue Wang, Tian Zhou, Qingsong Wen, Jinyang Gao, Bolin Ding, and Rong Jin. 2024. CARD: Channel aligned robust blend transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*.

[48] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. 2013. Strategies for handling missing data in electronic health record derived data. *Egems* 1, 3 (2013).

[49] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186* (2022).

[50] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. *Advances in Neural Information Processing Systems* 33 (2020), 20437–20448.

[51] Jingyun Xiao, Ran Liu, and Eva L Dyer. 2024. GAFormer: Enhancing Timeseries Transformers Through Group-Aware Embeddings. In *The Twelfth International Conference on Learning Representations*.
[52] Duo Xu and Faramarz Fekri. 2018. Time series prediction via recurrent neural networks with the information bottleneck principle. In *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 1–5.
[53] Pengshuai Yao, Mengna Liu, Xu Cheng, Fan Shi, Huan Li, Xiufeng Liu, and Shengyong Chen. 2024. An End-to-End Model for Time Series Classification In the Presence of Missing Values. *arXiv preprint arXiv:2408.05849* (2024).
[54] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2114–2124.
[55] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. 2021. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357* (2021).
[56] Rundong Zuo, Guozhong Li, Byron Choi, Sourav S Bhowmick, Daphne Ngar-yin Mah, and Grace LH Wong. 2023. SVP-T: a shape-level variable-position transformer for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11497–11505.

A EXPERIMENTAL DETAILS

A.1 Datasets

All datasets are introduced as follows:

Multivariate Time Series: We selected 10 multivariate datasets from the UEA Time Series Classification Archive, covering tasks such as gesture recognition, action recognition, audio recognition, and heartbeat monitoring. The datasets include: EthanolConcentration (EC), FaceDetection (FD), SelfRegulationSCP1 (SCP1), SpokenArabicDigits (SAD), SelfRegulationSCP2 (SCP2), JapaneseVowels (JV), UWaveGestureLibrary (UW), Handwriting (HW), PEMS-SF (PM), and Heartbeat (HB). Table 6 presents the statistics of these datasets.

Univariate Time Series: We selected 8 univariate datasets from the UCR Time Series Classification Archive, following AJRNN, covering various fields and sample sizes. The datasets include: CBF, CinC ECG (C-ECG), FaceAll (FA), ProxPhxAgeGp (PPAG), ScreenType (ST), SonyRobot Sur (SR Sur), SonyRobot Sur2 (SR Sur2), and UWavGestAll (UWGA). Table 6 presents the statistics of these datasets.

Real-World Datasets: The statistics of DodgerLpDay (DodgerLD), DodgerLpGame (DodgerLG), DodgerLpWend (DodgerLW), MelPedestrian (MP), and PAM is referenced in Table 7.

A.2 Baselines

All baselines are introduced as follows:

TARNet: Task-AwareReconstruction Network,a new model using Transformers to learn task-aware data reconstruction that augments end-task performance.

TimesNet: TimesNet transforms 1D time series into 2D tensors to model complex temporal variations, achieving state-of-the-art results in forecasting, imputation, classification, and anomaly detection.

ModernTCN: ModernTCN revitalizes convolution for time series analysis by modernizing traditional TCN, achieving state-of-the-art performance across five tasks while maintaining efficiency, and demonstrating larger effective receptive fields to unlock the full potential of convolution.

TimeMIL: TimeMIL is a novel multiple instance learning framework for multivariate time series classification that leverages time-aware MIL pooling and a tokenized transformer with learnable wavelet positional tokens.

GRU-D: GRU-D employs gated recurrent units with decay mechanisms and binary masking to fill missing values.

BRITS: BRITS is a bidirectional recurrent neural network-based method for missing value imputation in time series data, which directly learns missing values without imposing strong assumptions, handles multiple correlated missing values.

mTAND: mTAND is a deep learning framework for irregularly sampled multivariate time series, using continuous time embeddings and attention mechanisms to create fixed-length representations.

AJRNN: AJRNN is an end-to-end model for incomplete time series classification that integrates imputation and classification using adversarial training to reduce error propagation.

RainDrop: RainDrop is a graph neural network designed for irregularly sampled and multivariate time series, using a novel message-passing operator to model time-varying sensor dependencies and predict misaligned readouts.

G-MS: G-MS is an end-to-end neural network for incomplete time series classification that unifies data imputation and representation learning, prioritizing classification performance over imputation accuracy and leveraging a multi-scale feature learning module to extract useful information

GPVAE: GPVAE is a deep sequential latent variable model that leverages a Gaussian process for smooth temporal evolution of lower-dimensional representations and a VAE approach with structured variational approximation to achieve non-linear dimensionality reduction and data imputation.

SAITS: SAITS is a novel method for missing value imputation in multivariate time series based on the self-attention mechanism, which learns missing values from a weighted combination of two diagonally-masked self-attention blocks through joint optimization.

B SUPPLEMENTARY EXPERIMENTS.

B.1 Complete result on Multivariate Datasets

Here, we present the complete results of Section 5.2, which includes the performance of HCIB and 16 baselines across 10 multivariate datasets. Specifically, we provide the full outcomes for each dataset under six different missing rates. As shown in Table 8.

B.2 Complete result on Univariate Datasets

Here, we present the complete results of Section 5.3, which includes the performance of HCIB and 5 baselines across 8 univariate datasets. Specifically, we provide the full outcomes for each dataset under four different missing rates. As shown in Table 9.

B.3 Hyperparameter Sensitivity Analysis

We conducted hyperparameter sensitivity experiments on the SCP1 and JV datasets, focusing on the the embedding size for imputation, the embedding size for classification, and the number of layers in the classification encoder. The results are summarized as shown in Figure 5:

Table 6: The statistics of Datasets.

Dataset	Sample Numbers (train set, test set)	Variable Number	Series Length
EthanolConcentration	(261, 263)	3	1751
FaceDetection	(5890, 3524)	144	62
Handwriting	(150, 850)	3	152
Heartbeat	(204, 205)	61	405
JapaneseVowels	(270, 370)	12	29
PEMS-SF	(267, 173)	963	144
SelfRegulationSCP1	(268, 293)	6	896
SelfRegulationSCP2	(200, 180)	7	1152
SpokenArabicDigits	(6599, 2199)	13	93
UWaveGestureLibrary	(120, 320)	3	315
CBF	(30, 900)	1	128
CinC ECG	(40, 1380)	1	1369
FaceAll	(560, 1690)	1	131
ProxPhxAgeGp	(400, 205)	1	80
ScreenType	(375, 375)	1	720
SonyRobot Sur	(20, 601)	1	70
SonyRobot Sur2	(27, 953)	1	65
UWavGestAll	(896, 3582)	1	945

Table 7: The statistics of Real-World Datasets.

Dataset	Sample Numbers (train set, (val set), test set)	Variable Number	Series Length	Missing rate
DodgerLpDay	(78, 80)	1	288	(14%, 4%)
DodgerLpGame	(20, 128)	1	288	(15%, 8%)
DodgerLpWend	(20, 138)	1	288	(10%, 9%)
MelPedestrian	(1194, 2439)	1	24	(5%, 5%)
PAM	(4266, 534, 533)	17	600	(60%, 60%, 60%)

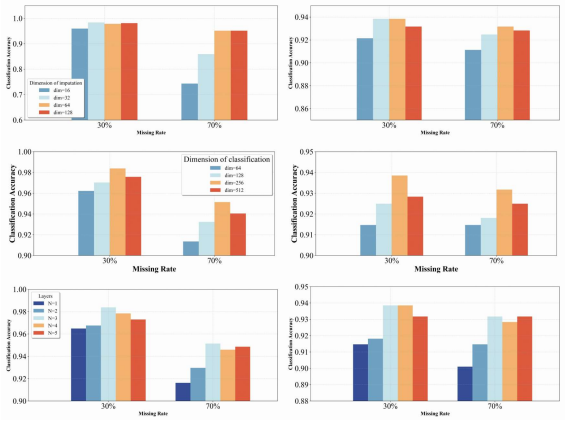

Figure 5: Hyperparameter (the embedding size for imputation, the embedding size for classification, and the number of layers of the classification encoder) ensitivity experiments in SCP (left) and JV (right).

Effect of Imputation Embedding Size: When the missing rate is low, the size of the imputation embedding has little impact on classification performance, and good imputation results can be achieved with a small number of parameters. However, when the missing rate is high, increasing the embedding size appropriately improves classification performance. This is because recovering missing information and retaining enriched information become more challenging under higher missing rates.

Effect of Classification Embedding Size: The influence of classification embedding size remains consistent across different missing rates. Properly increasing the embedding size can improve classification performance, but excessively large embedding sizes may lead to overfitting, which negatively impacts classification performance.

Effect of the Number of Encoder Layers: The influence of the number of layers is also consistent across different missing rates. A smaller number of layers may fail to capture the complete temporal and variable dependencies. Increasing the number of layers appropriately enhances classification performance. However, after a certain point (e.g., N=3), further increasing the number of layers does not lead to additional improvements in classification performance.

Table 8: Complete Performance comparison results of all baselines and HCIB on multivariate datasets.

Method		Our	TM+GP	TM+SA	TM	MT+GP	MT+SA	MT	TN+GP	TN+SA	TN	TA+GP	TA+SA	TA	RD	AJ	mT	GD
EC	30%	**0.338**	0.312	0.323	0.335	0.266	0.304	0.278	0.285	0.278	0.289	0.259	0.255	0.278	0.262	0.270	0.304	0.300
	40%	0.335	0.327	**0.342**	0.338	0.274	0.278	0.289	0.293	0.293	0.278	0.308	0.262	0.289	0.281	0.255	0.278	0.300
	50%	0.331	**0.350**	0.346	0.346	0.289	0.270	0.308	0.285	0.281	0.323	0.285	0.278	0.274	0.319	0.289	0.289	0.251
	60%	0.342	**0.346**	0.327	0.327	0.293	0.285	0.285	0.281	0.270	0.293	0.274	0.289	0.266	0.319	0.274	0.293	0.285
	70%	0.331	0.319	**0.338**	0.278	0.274	0.278	0.304	0.300	0.300	0.278	0.278	0.281	0.278	0.300	0.236	0.304	0.292
	80%	**0.346**	0.274	0.327	0.327	0.270	0.274	0.312	0.297	0.262	0.274	0.274	0.289	0.266	0.281	0.266	0.312	0.292
FD	30%	0.686	0.677	0.669	0.672	0.676	**0.687**	0.674	0.643	0.665	0.654	0.610	0.617	0.595	0.591	0.563	0.541	0.555
	40%	**0.681**	0.644	0.677	0.669	0.674	0.671	0.673	0.674	0.654	0.661	0.598	0.602	0.593	0.579	0.553	0.541	0.569
	50%	0.674	0.652	0.671	0.651	0.665	**0.675**	0.660	0.634	0.652	0.641	0.579	0.595	0.582	0.593	0.558	0.539	0.541
	60%	**0.674**	0.658	0.647	0.647	0.658	0.663	0.659	0.625	0.641	0.645	0.562	0.580	0.562	0.579	0.562	0.543	0.531
	70%	**0.658**	0.636	0.648	0.639	0.655	0.653	0.642	0.634	0.625	0.630	0.563	0.531	0.528	0.531	0.557	0.535	0.528
	80%	**0.656**	0.609	0.509	0.599	0.629	0.650	0.617	0.611	0.635	0.618	0.558	0.515	0.531	0.531	0.550	0.552	0.524
SCP1	30%	**0.939**	0.877	0.877	0.864	0.925	0.918	0.911	0.908	0.898	0.887	0.608	0.771	0.710	0.891	0.710	0.836	0.710
	40%	**0.939**	0.864	0.867	0.843	0.911	0.913	0.918	0.898	0.908	0.887	0.771	0.727	0.775	0.887	0.648	0.833	0.689
	50%	**0.935**	0.891	0.864	0.819	0.911	0.911	0.898	0.904	0.898	0.884	0.765	0.765	0.628	0.894	0.727	0.836	0.707
	60%	**0.932**	0.864	0.860	0.836	0.901	0.898	0.881	0.898	0.884	0.877	0.649	0.741	0.775	0.918	0.666	0.829	0.700
	70%	**0.932**	0.857	0.846	0.857	0.908	0.896	0.877	0.898	0.887	0.877	0.659	0.713	0.707	0.894	0.764	0.829	0.696
	80%	**0.932**	0.843	0.846	0.829	0.898	0.898	0.853	0.887	0.867	0.850	0.628	0.642	0.840	0.898	0.662	0.823	0.713
SAD	30%	**0.995**	0.991	0.988	0.991	0.973	0.980	0.969	0.989	0.986	0.985	0.967	0.976	0.948	0.979	0.909	0.954	0.672
	40%	**0.993**	0.985	0.991	0.990	0.968	0.978	0.959	0.989	0.988	0.985	0.973	0.981	0.956	0.974	0.921	0.949	0.663
	50%	**0.993**	0.989	0.989	0.990	0.964	0.971	0.956	0.984	0.985	0.976	0.970	0.975	0.931	0.969	0.856	0.946	0.659
	60%	**0.992**	0.986	0.988	0.983	0.965	0.970	0.944	0.985	0.983	0.977	0.970	0.971	0.919	0.960	0.842	0.952	0.666
	70%	0.983	0.974	0.978	0.978	0.930	0.970	0.924	0.976	**0.984**	0.960	0.911	0.960	0.875	0.954	0.802	0.920	0.641
	80%	**0.982**	0.966	0.961	0.960	0.880	0.960	0.904	0.952	0.966	0.942	0.850	0.947	0.796	0.926	0.685	0.878	0.600
SCP2	30%	**0.617**	0.533	0.550	0.539	0.533	0.567	0.578	0.556	0.550	0.539	0.578	0.572	0.550	0.544	0.528	0.583	0.572
	40%	**0.622**	0.572	0.567	0.583	0.539	0.528	0.544	0.550	0.533	0.511	0.589	0.606	0.528	0.533	0.483	0.589	0.556
	50%	**0.611**	0.578	0.578	0.556	0.511	0.539	0.533	0.522	0.561	0.528	0.561	0.578	0.578	0.567	0.472	0.567	0.556
	60%	**0.628**	0.578	0.556	0.589	0.522	0.561	0.522	0.550	0.533	0.522	0.550	0.561	0.533	0.544	0.533	0.572	0.544
	70%	**0.617**	0.533	0.550	0.556	0.533	0.544	0.539	0.550	0.572	0.556	0.561	0.567	0.528	0.556	0.556	0.578	0.550
	80%	**0.617**	0.572	0.550	0.517	0.517	0.550	0.556	0.528	0.528	0.528	0.567	0.600	0.578	0.589	0.522	0.567	0.583
JV	30%	**0.984**	0.957	0.976	0.954	0.957	0.960	0.949	0.946	0.962	0.935	0.946	0.938	0.935	0.943	0.430	0.784	0.324
	40%	**0.986**	0.949	0.973	0.968	0.951	0.954	0.938	0.951	0.951	0.914	0.922	0.927	0.924	0.911	0.397	0.749	0.335
	50%	**0.970**	0.968	0.954	0.930	0.935	0.935	0.903	0.932	0.949	0.859	0.900	0.897	0.889	0.857	0.238	0.711	0.305
	60%	**0.968**	0.908	0.949	0.927	0.897	0.900	0.851	0.897	0.908	0.776	0.887	0.870	0.881	0.787	0.219	0.714	0.308
	70%	**0.951**	0.857	0.892	0.846	0.876	0.884	0.830	0.881	0.884	0.759	0.851	0.854	0.835	0.741	0.205	0.765	0.305
	80%	**0.897**	0.830	0.824	0.762	0.714	0.797	0.686	0.797	0.789	0.670	0.838	0.762	0.770	0.619	0.178	0.605	0.289
UW	30%	**0.922**	0.853	0.872	0.850	0.847	0.856	0.841	0.872	0.866	0.866	0.866	0.828	0.806	0.763	0.719	0.747	0.166
	40%	**0.916**	0.838	0.856	0.841	0.853	0.850	0.841	0.869	0.869	0.863	0.828	0.797	0.725	0.744	0.641	0.675	0.163
	50%	**0.922**	0.822	0.813	0.831	0.856	0.844	0.831	0.869	0.869	0.863	0.800	0.844	0.659	0.731	0.538	0.650	0.169
	60%	**0.916**	0.813	0.816	0.781	0.841	0.847	0.828	0.850	0.859	0.841	0.709	0.806	0.622	0.669	0.553	0.713	0.153
	70%	**0.916**	0.778	0.828	0.819	0.831	0.847	0.828	0.838	0.863	0.844	0.125	0.866	0.634	0.634	0.572	0.513	0.159
	80%	**0.913**	0.731	0.825	0.791	0.816	0.838	0.816	0.825	0.847	0.819	0.666	0.844	0.125	0.578	0.519	0.613	0.172
HW	30%	0.319	0.306	0.321	0.278	0.196	0.215	0.199	0.260	**0.324**	0.284	0.171	0.198	0.180	0.124	0.066	0.038	0.034
	40%	0.294	0.294	**0.318**	0.301	0.195	0.259	0.181	0.236	0.241	0.253	0.146	0.171	0.133	0.129	0.072	0.038	0.044
	50%	**0.272**	0.207	0.237	0.221	0.146	0.211	0.153	0.201	0.219	0.202	0.127	0.137	0.138	0.080	0.056	0.038	0.046
	60%	**0.252**	0.205	0.215	0.192	0.142	0.153	0.144	0.178	0.196	0.206	0.111	0.128	0.107	0.093	0.049	0.038	0.041
	70%	**0.211**	0.060	0.188	0.134	0.118	0.109	0.122	0.153	0.149	0.152	0.085	0.082	0.077	0.085	0.052	0.039	0.044
	80%	**0.161**	0.089	0.106	0.089	0.093	0.095	0.093	0.107	0.127	0.114	0.068	0.078	0.068	0.074	0.051	0.038	0.037
PM	30%	**0.913**	0.699	0.659	0.734	0.867	0.821	0.763	0.850	0.827	0.855	0.832	0.804	0.792	-	-	0.133	0.173
	40%	**0.908**	0.769	0.647	0.786	0.855	0.798	0.711	0.890	0.815	0.861	0.780	0.832	0.705	-	-	0.133	0.173
	50%	**0.896**	0.711	0.665	0.624	0.827	0.775	0.607	0.850	0.855	0.855	0.688	0.838	0.751	-	-	0.133	0.173
	60%	**0.896**	0.653	0.624	0.705	0.832	0.769	0.665	0.844	0.736	0.832	0.740	0.757	0.636	-	-	0.133	0.173
	70%	**0.855**	0.633	0.520	0.642	0.821	0.751	0.636	**0.855**	0.786	0.850	0.746	0.815	0.642	-	-	0.133	0.173
	80%	**0.838**	0.653	0.572	0.653	0.775	0.694	0.595	0.809	0.584	0.780	0.682	0.486	0.665	-	-	0.127	0.173
HB	30%	0.785	0.761	0.766	**0.790**	0.732	0.727	0.732	0.737	0.741	0.746	0.732	0.756	0.712	0.742	0.678	0.766	0.722
	40%	0.780	0.776	**0.781**	0.776	0.722	0.746	0.741	0.722	0.737	0.737	0.722	0.751	0.698	0.742	0.624	0.761	0.727
	50%	**0.785**	0.776	0.742	0.771	0.722	0.741	0.727	0.737	0.741	0.727	0.722	0.712	0.717	0.742	0.673	0.761	0.727
	60%	0.776	0.761	0.737	**0.785**	0.737	0.732	0.732	0.746	0.741	0.746	0.746	0.751	0.712	0.742	0.605	0.756	0.727
	70%	**0.771**	0.766	0.737	0.766	0.741	0.732	0.732	0.732	0.727	0.741	0.737	0.722	0.722	0.737	0.620	0.756	0.732
	80%	**0.771**	0.761	0.732	0.756	0.732	0.732	0.722	0.727	0.756	0.737	0.717	0.727	0.712	0.722	0.717	0.761	0.712
Average		**0.735**	0.674	0.673	0.674	0.676	0.680	0.656	0.687	0.683	0.676	0.619	0.645	0.603	0.623	0.494	0.548	0.415

B.4 Performance of Representation Learning

In this section, we validate the effectiveness of the information bottleneck for learning classification feature representations by performing classification tasks on complete multivariate datasets, evaluated using average rank and top-2 count metrics, and comparing against SOTA deep learning and traditional multivariate classification methods.As shown in Table 10, our method achieves the best performance in both average rank (2.22) and top-2 count (8). Compared to three Transformer-based methods (TST, ConvTran, SVP-T), our approach leads by 4.11, 5.11, and 3.89 in average rank, respectively. Additionally, it outperforms ShapeFormer and WHEN in terms of top-2 count. These results demonstrate the effectiveness of leveraging the information bottleneck for learning discriminative feature representations in classification tasks.

Table 9: Complete Performance comparison results of all baselines and HCIB on univariate datasets.

Method		Our	G-MS	AJRNN	RainDrop	GRU-D	BRITS
CBF	20%	0.992	**0.998**	0.993	0.827	0.373	0.598
	40%	0.983	**0.993**	0.979	0.766	0.377	0.582
	60%	0.962	**0.978**	0.934	0.511	0.376	0.441
	80%	**0.917**	0.916	0.814	0.331	0.354	0.398
CinC_ECG_torso	20%	**0.953**	0.700	0.450	0.891	0.273	0.298
	40%	**0.901**	0.748	0.425	0.828	0.241	0.269
	60%	**0.930**	0.673	0.395	0.684	0.249	0.320
	80%	**0.776**	0.680	0.371	0.425	0.247	0.248
FaceAll	20%	**0.831**	0.784	0.764	0.797	0.511	0.801
	40%	**0.796**	0.780	0.749	0.706	0.583	0.726
	60%	**0.783**	0.745	0.651	0.472	0.383	0.653
	80%	**0.745**	0.637	0.446	0.299	0.248	0.451
ProxPhxAgeGp	20%	**0.883**	0.868	0.876	0.829	0.805	0.849
	40%	**0.883**	0.873	0.870	0.790	0.805	0.815
	60%	**0.873**	**0.873**	0.865	0.678	0.834	0.849
	80%	**0.854**	**0.854**	0.850	0.581	0.649	0.795
ScreenType	20%	**0.483**	0.464	0.469	0.405	0.360	0.400
	40%	**0.485**	0.475	0.459	0.413	0.307	0.408
	60%	**0.472**	0.456	0.453	0.411	0.333	0.403
	80%	**0.469**	0.448	0.444	0.389	0.368	0.405
SonyRobot Sur	20%	**0.844**	0.717	0.836	0.429	0.429	0.429
	40%	0.789	0.659	**0.796**	0.429	0.429	0.429
	60%	**0.752**	0.669	0.748	0.429	0.429	0.429
	80%	**0.677**	0.599	0.666	0.429	0.429	0.429
SonyRobot Sur2	20%	**0.850**	0.789	0.818	0.830	0.652	0.662
	40%	**0.813**	0.758	0.791	0.796	0.659	0.745
	60%	**0.773**	0.738	0.769	0.765	0.633	0.626
	80%	**0.739**	0.720	0.713	0.617	0.620	0.617
UWavGestAll	20%	**0.960**	0.922	0.934	0.930	0.179	0.383
	40%	**0.957**	0.918	0.926	0.876	0.315	0.286
	60%	**0.951**	0.913	0.918	0.883	0.235	0.388
	80%	**0.934**	0.918	0.902	0.790	0.248	0.348
Average		**0.813**	0.758	0.721	0.632	0.435	0.515

B.5 Visualization Analysis

To better understand the mechanisms behind HCIB, we analyzed the features in the final hidden layer, which serves as input to the classifier. The JV dataset was selected as an example to investigate the effectiveness of HCIB under varying missing data rates. High-dimensional features were mapped to a 2D space using t-SNE (Figure 6).At low missing rates (30% 50%), the features formed well-separated clusters, indicating that HCIB maintains strong classification performance. As the missing rate increases, the learned features still exhibit distinct cluster structures even at an 80% missing rate. This demonstrates that HCIB effectively mitigates the negative impact of high missing rates by focusing on residual temporal patterns and features while preserving enriched information.

C Theoretical Analysis and Proof

C.1 Discussion: Differences from standard information bottlenecks:

We employed Venn diagrams to comparatively analyze the differences in information retention mechanisms between the unsupervised Information Bottleneck (IB) principle and the proposed Conditional Information Bottleneck (CIB) principle. For a theoretical foundation on information diagrams, further details can be found in the work titled *A New Outlook on Shannon's Information Measures*. Unlike the traditional IB principle, the innovation of CIB lies in avoiding the simultaneous minimization and maximization operations on A_2 information, thereby ensuring the complete preservation of A_2 information (i.e., residual temporal patterns and

Table 10: Performance of Our Proposed Method Compared to 13 Other Methods on 10 Datasets in the UEA Archive.

Method	EDI	DTWD	WEASEL +MUSE	Mini Rocket	LCEM	MLSTM-FCNs	Tapnet	Shapenet	WHEN	TST	Conv Tran	SVPT	Shape-Former	Our
EC	0.293	0.323	0.430	**0.468**	0.372	0.373	0.323	0.312	0.422	0.337	0.361	0.331	0.378	0.414
FD	0.519	0.529	0.545	0.620	0.614	0.545	0.556	0.602	0.658	0.681	0.672	0.512	0.658	**0.705**
HW	0.200	0.286	**0.605**	0.507	0.287	0.286	0.357	0.452	0.561	0.305	0.375	0.433	0.507	0.379
HB	0.619	0.717	0.727	0.771	0.761	0.663	0.751	0.756	0.780	0.712	0.785	0.790	0.800	**0.805**
JV	0.924	0.949	0.973	0.989	0.978	0.976	0.965	0.984	0.995	0.994	0.989	0.978	**0.997**	0.995
PEMS	**0.973**	0.711	N/A	0.522	0.942	0.699	0.751	0.751	0.925	0.919	0.828	0.867	0.925	0.942
SCP1	0.771	0.775	0.710	0.925	0.839	0.874	0.652	0.782	0.908	0.925	0.918	0.884	0.911	**0.939**
SCP2	0.483	0.539	0.460	0.522	0.550	0.472	0.550	0.578	0.589	0.589	0.583	0.600	**0.633**	0.611
SAD	0.967	0.963	0.982	0.620	0.973	0.990	0.983	0.975	0.997	0.993	N/A	0.986	0.997	**0.998**
UWGL	0.881	0.903	0.916	0.938	0.897	0.891	0.894	0.906	0.919	0.903	0.891	**0.941**	0.922	0.925
Ave rank	11.89	11.00	8.00	6.33	7.22	9.78	9.00	7.78	3.33	6.33	7.33	6.11	2.78	**2.22**
Num of top-2	1	0	2	2	1	0	0	0	3	1	0	1	4	**8**

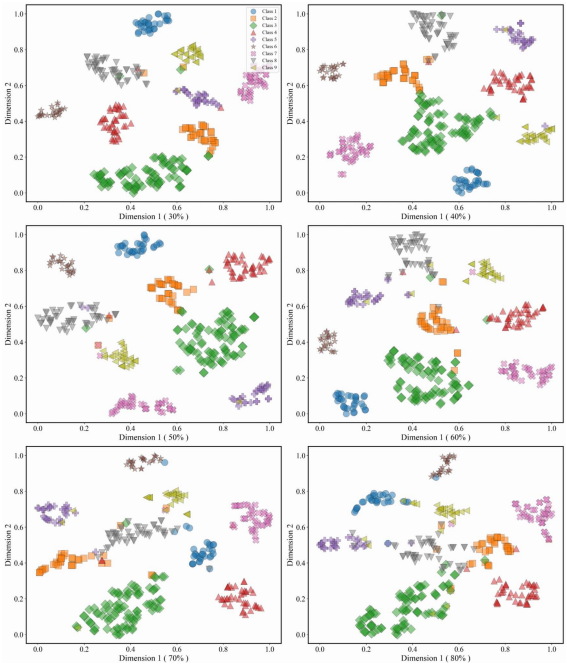

Figure 6: The t-SNE visualization of JV dateset .

dynamic features). Specifically, by more comprehensively learning A_2 information, CIB provides robust support for recovering missing A_1 information (i.e., lost temporal patterns and dynamic features). Notably, through the maximization of A_3 information (i.e., the suppression of information interference), our method enables the imputed latent representations to better retain classification gain information arising from data missingness, thereby enhancing overall performance.

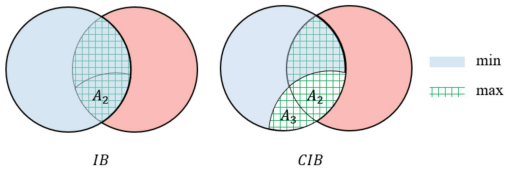

Figure 7: The difference of optimization between standard information bottleneck and our conditional information bottleneck.

C.2 Variational approximation of Imputation Information Term

Here we derive the upper bound of $-I(X;Z^1)$through variational inference. According to the definition of variational information bottleneck:

$$\begin{aligned} -I(X;Z^1) &= -\mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\log \frac{p(\mathbf{x},\mathbf{z}^1)}{p(\mathbf{x})p(\mathbf{z}^1)}\right] \\ &= -\mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\log \frac{p(\mathbf{x}|\mathbf{z}^1)}{p(\mathbf{x})}\right] \\ &= \mathbb{E}_{p(\mathbf{x})}\left[\log p(\mathbf{x})\right] - \mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\log p(\mathbf{x}|\mathbf{z}^1)\right] \\ &= -H(X) - \mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\log p(\mathbf{x}|\mathbf{z}^1)\right], \end{aligned} \tag{17}$$

among them:

$$\begin{aligned} &-\mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\log p(\mathbf{x}|\mathbf{z}^1)\right] = -\mathbb{E}_{p(\mathbf{x},\mathbf{x}^0)}\left[\mathbb{E}_{p(\mathbf{z}^1|\mathbf{x}^0)}\left[\log p(\mathbf{x}|\mathbf{z}^1)\right]\right] \\ &= -\mathbb{E}_{p(\mathbf{x},\mathbf{x}^0)}\left[\mathbb{E}_{p(\mathbf{z}^1|\mathbf{x}^0)}\left[\log p(\mathbf{x}|\mathbf{z}^1)\frac{q(\mathbf{x}|\mathbf{z}^1)}{q(\mathbf{x}|\mathbf{z}^1)}\right]\right] \\ &= -\mathbb{E}_{p(\mathbf{x}^0)}\left[\mathbb{E}_{p(\mathbf{z}^1|\mathbf{x}^0)}\left[\log q(\mathbf{x}|\mathbf{z}^1)\right]\right] - \mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\frac{p(\mathbf{x}|\mathbf{z}^1)}{q(\mathbf{x}|\mathbf{z}^1)}\right]. \end{aligned} \tag{18}$$

Here we introduce the variational distribution $q(\mathbf{x}|\mathbf{z}^1)$to overcome the difficult to compute $p(\mathbf{x}|\mathbf{z}^1)$:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\frac{p(\mathbf{x}|\mathbf{z}^1)}{q(\mathbf{x}|\mathbf{z}^1)}\right] &= \mathbb{E}_{p(\mathbf{z}^1)}\left[\mathbb{E}_{p(\mathbf{x}|\mathbf{z}^1)}\left[\frac{p(\mathbf{x}|\mathbf{z}^1)}{q(\mathbf{x}|\mathbf{z}^1)}\right]\right] \\ &= D_{KL}(p(\mathbf{x}|\mathbf{z}^1) \parallel q(\mathbf{x}|\mathbf{z}^1)),\end{aligned} \tag{19}$$

because KL divergence is non negative:

$$-\mathbb{E}_{p(\mathbf{x},\mathbf{z}^1)}\left[\log p(\mathbf{x}|\mathbf{z}^1)\right] \le -\mathbb{E}_{p(\mathbf{x},\mathbf{x}^0)}\left[\mathbb{E}_{p(\mathbf{z}^1|\mathbf{x}^0)}\left[\log q(\mathbf{x}|\mathbf{z}^1)\right]\right]. \tag{20}$$

Since $H(X)$is not related to our optimization, therefore:

$$-I(X;Z^1) \le -\mathbb{E}_{p(\mathbf{x}^0)}\left[\mathbb{E}_{p(\mathbf{z}^1|\mathbf{x}^0)}\left[\log q(\mathbf{x}|\mathbf{z}^1)\right]\right] \tag{21}$$

C.3 Variational approximation of Classification Information Term

We can derive the upper bound of $-I(Y;Z^2)$using a method similar to C.2:

$$\begin{aligned}&-I(Y;Z^2) \\ &= -\mathbb{E}_{p(\mathbf{y},\mathbf{z}^2)}\left[\log\frac{p(\mathbf{y},\mathbf{z}^2)}{p(\mathbf{y})p(\mathbf{z}^2)}\right] \\ &= -H(Y) - \mathbb{E}_{p(\mathbf{y},\mathbf{z}^2)}\left[\log p(\mathbf{y}|\mathbf{z}^2)\right] \\ &= -H(Y) - \mathbb{E}_{p(\mathbf{y},\mathbf{x})}\left[\mathbb{E}_{p(\mathbf{z}^2|\mathbf{x})}\left[\log p(\mathbf{y}|\mathbf{z}^2)\right]\right] \\ &= -H(Y) - \mathbb{E}_{p(\mathbf{y},\mathbf{x})}\left[\mathbb{E}_{p(\mathbf{z}^2|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{z}^2)\right]\right] - \mathbb{E}_{p(\mathbf{y},\mathbf{z}^2)}\left[\frac{p(\mathbf{y}|\mathbf{z}^2)}{q(\mathbf{y}|\mathbf{z}^2)}\right] \\ &= -H(Y) - \mathbb{E}_{p(\mathbf{y},\mathbf{x})}\left[\mathbb{E}_{p(\mathbf{z}^2|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{z}^2)\right]\right] - D_{KL}(p(\mathbf{y}|\mathbf{z}^2) \parallel q(\mathbf{y}|\mathbf{z}^2)).\end{aligned} \tag{22}$$

Due to the non negative nature of KL divergence and the independence of $H(Y)$from our optimization, ultimately:

$$-I(Y;Z^2) \le -\mathbb{E}_{p(\mathbf{y},\mathbf{x})}\left[\mathbb{E}_{p(\mathbf{z}^2|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{z}^2)\right]\right]. \tag{23}$$

C.4 Variational approximation of Classification Compression Term

Here we derive the upper bound of $I(Z^2;X)$:

$$\begin{aligned}I(Z^2;X) &= \mathbb{E}_{p(\mathbf{z}^2,\mathbf{x})}\left[\log\frac{p(\mathbf{z}^2,\mathbf{x})}{p(\mathbf{z}^2)p(\mathbf{x})}\right] \\ &= -\mathbb{E}_{p(\mathbf{z}^2)}\left[\log p(\mathbf{z}^2)\right] + \mathbb{E}_{p(\mathbf{z}^2,\mathbf{x})}\left[\log p(\mathbf{z}^2|\mathbf{x})\right].\end{aligned} \tag{24}$$

However, the entropy $-\mathbb{E}_{p(\mathbf{z}^2)}\left[\log p(\mathbf{z}^2)\right]$ requires the computation of the marginal distribution $p(\mathbf{z}^2) = \mathbb{E}_{p(\mathbf{x})}[p(\mathbf{z}^2|\mathbf{x})]$, which is computationally expensive in practice. Therefore, we use a distribution $q(z)$ to perform a variational approximation of $p(z)$:

$$\begin{aligned}I(Z^2;X) &= \mathbb{E}_{p(\mathbf{z}^2,\mathbf{x})}\left[\log\frac{p(\mathbf{z}^2|\mathbf{x})}{p(\mathbf{z}^2)}\frac{q(\mathbf{z}^2)}{q(\mathbf{z}^2)}\right] \\ &= \mathbb{E}_{p(\mathbf{x})}\left[D_{KL}(p(\mathbf{z}^2|\mathbf{x}) \parallel q(\mathbf{z}^2))\right] - D_{KL}(p(\mathbf{z}^2) \parallel q(\mathbf{z}^2)),\end{aligned} \tag{25}$$

Due to the fixed value of the second term and the non-negativity of the KL divergence, we finally have:

$$I(Z^2;X) \le \mathbb{E}_{p(\mathbf{x})}\left[D_{KL}(p(\mathbf{z}^2|\mathbf{x}) \parallel q(\mathbf{z}^2))\right]. \tag{26}$$

C.5 Variational approximation of imputation compression term

Here, we derive the upper bound of $I(Z^1;X^o,Y)$:

$$\begin{aligned}I(Z^1;X^o,Y) &= \int d\mathbf{x}\,d\mathbf{y}\,d\mathbf{z}\,p(\mathbf{z}^1,\mathbf{x}^o,\mathbf{y})\log\frac{p(\mathbf{z}^1,\mathbf{x}^o,\mathbf{y})}{p(\mathbf{x}^o,\mathbf{y})p(\mathbf{z}^1)} \\ &= \int d\mathbf{x}\,d\mathbf{y}\,d\mathbf{z}\,p(\mathbf{z}^1,\mathbf{x}^o,\mathbf{y})\log\frac{p(\mathbf{z}^1|\mathbf{x}^o,\mathbf{y})}{p(\mathbf{z}^1)}.\end{aligned} \tag{27}$$

Due to the Markov chain assumption: $Z^1 \leftarrow X^o \leftrightarrow Y$ (satisfying the conditional independence $Z^1 \perp Y|X^o$), it follows that $p(\mathbf{z}^1,\mathbf{x}^o,\mathbf{y}) = p(\mathbf{z}^1,\mathbf{x}^o)$:

$$\begin{aligned}I(Z^1;X^o,Y) &= \int d\mathbf{x}\,d\mathbf{z}\,p(\mathbf{z}^1,\mathbf{x}^o)\log\frac{p(\mathbf{z}^1|\mathbf{x}^o)}{p(\mathbf{z}^1)} \\ &\le \mathbb{E}_{p(\mathbf{x}^o)}\left[D_{KL}(p(\mathbf{z}^1|\mathbf{x}^o) \parallel q(\mathbf{z}^1))\right]\end{aligned} \tag{28}$$

C.6 Contrastive approximation of Conditional Term

We optimize $I(Z^1;Y)$ using a contrastive learning approach similar to [18, 24, 34].

$$\begin{aligned}I(Z^1;Y) &= -\mathbb{E}_{Z^1}\log\left[\frac{p(y)}{p(y|z^1)}\right] \\ &= -\mathbb{E}_{Z^1}\log\left[\frac{p(y)}{p(y|z^1)}N\right] + \log(N) \\ &\ge -\mathbb{E}_{Z^1}\log\left[\frac{p(y)}{p(y|z^1)}N\right] \\ &\ge -\mathbb{E}_{Z^1}\log\left[1+\frac{p(y)}{p(y|z^1)}(N-1)\right] \\ &= -\mathbb{E}_{Z^1}\log\left[1+\frac{p(y)}{p(y|z^1)}(N-1)\mathbb{E}_{Y'}\frac{p(y'|z^1)}{p(y')}\right] \\ &\approx -\mathbb{E}_{Z^1}\log\left[1+\frac{p(y)}{p(y|z^1)}\sum_{y'\in neg}\frac{p(y'|z^1)}{p(y')}\right] \\ &= \mathbb{E}_{Z^1}\log\left[\frac{\frac{p(y|z^1)}{p(y)}}{\frac{p(y|z^1)}{p(y)}+\sum_{y'\in neg}\frac{p(y'|z^1)}{p(y')}}\right] \\ &= \mathbb{E}_{Z^1}\log\left[\frac{f(z^1,y)}{f(z^1,y)+\sum_{y'\in neg}f(z^1,y')}\right].\end{aligned} \tag{29}$$

For the construction of the function $f(z^1,y)$, we replace y with the representation of a sample (augmented sample) that has the same label as z^1, and use cosine similarity for similarity calculation:

$$f(z^1,y) = \exp\left(\Phi(\mathbf{z}_i\cdot\mathbf{z}_j)/\tau\right), (\mathbf{y}_i = \mathbf{y}_j). \tag{30}$$

For the negative samples, we select the representations of samples (and their augmentations) with different labels within the same

batch. Ultimately:

$$\log\left[\frac{f(z^1,y)}{f(z^1,y)+\sum_{y'\in neg} f(z^1,y')}\right] \approx \frac{1}{N_{\mathrm{y}_i}-1}\sum_{j=1}^{N}\mathbf{1}_{i\neq j}\mathbf{1}_{\mathrm{y}_i=\mathrm{y}_j}\log\frac{Sim^1_{i,j}}{SumSim}+\frac{1}{N_{\mathrm{y}_i}}\sum_{j=1}^{N}\mathbf{1}_{\mathrm{y}_i=\mathrm{y}_j}\log\frac{Sim^2_{i,j}}{SumSim}, \tag{31}$$

where,$Sim^1_{ij}=\exp\left(\Phi(\mathbf{z}^o_i\cdot\mathbf{z}^o_j)/\tau\right)$ represents the similarity computation for the first type of positive samples; $Sim^2_{i,j}=\exp\left(\Phi(\mathbf{z}^o_i\cdot\mathbf{z}^{o'}_j)/\tau\right)$ represents the similarity for the second type of positive samples; $SumSim=\sum_{k=1}^{N}\mathbf{1}_{i\neq k}\exp\left(\Phi(\mathbf{z}^o_i\cdot\mathbf{z}^o_k)/\tau\right)+\sum_{k=1}^{N}\exp\left(\Phi(\mathbf{z}^o_i\cdot\mathbf{z}^{o'}_k)/\tau\right)$ represents the total sum of similarities for all positive and negative samples.

Finally:

$$I(Z^1;Y) \geq \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N_{\mathrm{y}_i}-1}\sum_{j=1}^{N}\mathbf{1}_{i\neq j}\mathbf{1}_{\mathrm{y}_i=\mathrm{y}_j}\log\frac{Sim^1_{i,j}}{SumSim}+\frac{1}{N_{\mathrm{y}_i}}\sum_{j=1}^{N}\mathbf{1}_{\mathrm{y}_i=\mathrm{y}_j}\log\frac{Sim^2_{i,j}}{SumSim}\right] \tag{32}$$

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009