Supplementary of DERI: Cross-Modal ECG Representation Learning with Deep ECG-Report Interaction

1 Implementation Details for Classification.

We provide more details about ECG classification datasets used in our experiments here. We adopt the same data split strategy as MERL ([Liu *et al.*, 2024]). The details are shown in Table 1. Notably, the PTBXL dataset with different tasks is split according to the official strategy ([Wagner *et al.*, 2020]) while the CPSC2018 and CSN datasets are split as 7:1:2.

Table 1: More details about the ECG classification datasets.

Dataset	Number of Categories	Samples	Train	Valid	Test
PTBXL-Super	5	21,388	17,084	2,146	2,158
PTBXL-Sub	23	21,388	17,084	2,146	2,158
PTBXL-Form	19	8,978	7,197	901	880
PTBXL-Rhythm	12	21,030	16,832	2,100	2,098
CPSC2018	9	6,877	4,950	551	1,376
CSN	38	23,026	16,546	1,860	4,620

Furthermore, we provided more details about the hyperparameters used for linear probing on ECG classification tasks in Table 2.

Table 2: Hyperparameter settings for Linear Probing

Learning rate	0.01
Batch size	16
Epochs	100
Optimizer	AdamW
Learning rate scheduler	Cosine annealing
Warump steps	5

2 Baselines for ECG Classification.

- **SimCLR [Chen *et al.*, 2020].** This method aims to maximize consistency between differently augmented views of the same data example including random cropping, random distortion, and random Gaussian blur. Contrastive loss obtained by the representation of augmented views is used to optimize the base encoder.
- **BYOL [Grill *et al.*, 2020].** This method uses an online network to predict the presentation of other augmented views of the same sample obtained by a target network without negative pairs.
- **BarlowTwins [Zbontar *et al.*, 2021].** This method proposes to measure the cross-correlation matrix between the outputs obtained by two identical networks with distorted versions of a sample. The cross-correlation matrix is made as close to the identity matrix as possible since it can make the representation of distorted versions from the same simple to be similar.
- **MoCo-v3 [Chen *et al.*, 2021].** This method proposes to train the transformer for self-supervised learning based on the investigation of the fundamental components during training. In this way, they aim to overcome the instability of the transformer for a better representation.
- **SimSiam [Chen and He, 2021].** This method uses an encoder to process two augmented views of one sample and then a prediction MLP is applied on one side while stop-gradient is applied on the other side. Neither negative pairs nor momentum are used to learn meaningful representations.
- **TS-TCC [Eldele *et al.*, 2021].** This method first augments the sample using weak and strong augmentation and then learns robust temporal representation with a cross-view prediction task. Finally, the similarity among contexts from the same sample is minimized as contextual contrasting learning.
- **CLOCS [Kiyasseh *et al.*, 2021].** This method learns patient-specific representations of ECG signals via contrastive learning with consideration of spatial-temporal information.
- **ASTCL [Wang *et al.*, 2023].** This method proposes a novel ECG augmentation method based on the noise attributes and then combines an adversarial module and a spatial-temporal contrastive module to learn the spatial-temporal and semantic representations of ECG signals.
- **CRT [Zhang *et al.*, 2023].** This method proposes to model temporal-spectral correlations of temporal time series by a cross-reconstruction transformer. Through cross-domain dropping reconstruction, the model can adequately capture the correlations between temporal and spectral information.
- **STMEM [Na *et al.*, 2024].** This method is proposed to learn the spatial-temporal relations of ECG signals for masked modeling. ECG signals are divided into patches on the temporal and spatial dimensions and then the

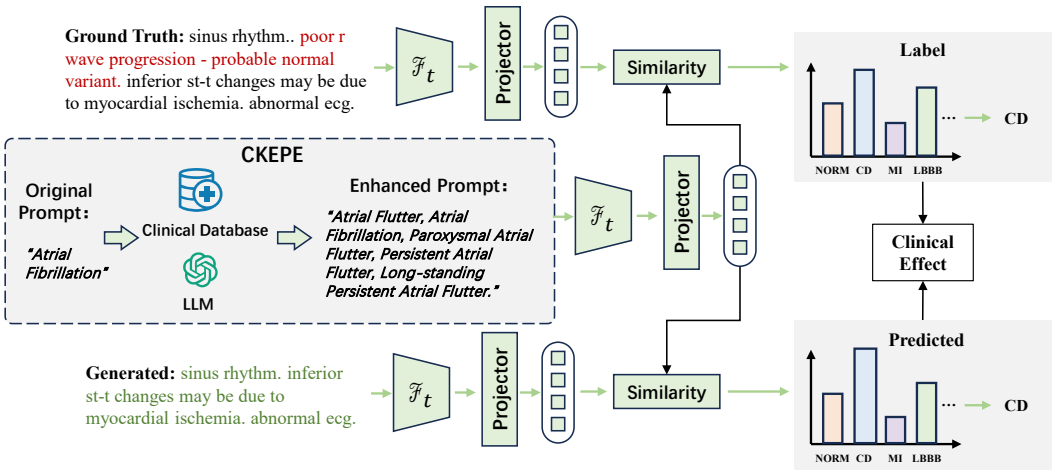

Figure 1: CE metrics calculation inspired by zero-shot classification.

model reconstructs the masked patches to learn spatial-temporal features of 12-lead ECG signals.

- **MERL [Liu *et al.*, 2024].** This method conducts multi-modal ECG self-supervised learning by directly aligning the ECG signal encoding and the text encoding to learn ECG representation. However, shallow interaction between ECG signals and reports can not provide effective clinical semantics for the representation.

3 Report CE metrics inspired by zero-shot classification.

Inspired by ECG zero-shot classification, we conduct report classification as Fig. 1. Specifically, we use the pre-trained text decoder to obtain prompt embeddings of all categories in the CKEPE. Considering there is no annotated label on this dataset, we feed the pre-trained text decoder with the ground clinical report and calculate the similarity as the classification probability for the category. We adopt the category with the highest classification probability as the ground truth label. Then we use the generated report to obtain the predicted label in the same way, which is then used to compute precision, recall, and F1 scores against ground truths as the CE metrics. We conduct the experiments on 2 NVIDIA GeForce RTX 4090 GPUs.

4 Report Generation on PTB-XL.

To conduct more comprehensive experiments about report generation, we conduct report generation tasks on PTB-XL, which has not been used in our pre-training. We compare our proposed method with methods that depend on LLMs, such as MEIT and ECG-Chat. Experimental results are shown in Supplementary.

Specifically, MEIT aims to use an instruction prompt to generate reports based on the ECG signal input with LLMs. ECG-Chat combines the ECG encoder and classification results to construct instructions for LLM and also uses the electrical health record and other information. However, our proposed DERI achieves the best performance. In addition, the experiments of our DERI are conducted on 4090 GPUs with fewer computing resources, which shows a great advantage.

Table 3: REPORT GENERATION RESULTS on PTBXL dataset.

Method	NLG				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
MERL-DisGPT2	48.9	44.4	41.2	37.4	55.4
DERI-DisGPT2	**58.6**	**54.8**	**51.9**	**48.6**	**64.2**
DERI-Align-DisGPT2	54.1	49.8	46.6	43.2	60.1
GPT2-Medium	32.9	27.8	25.4	23.2	39.1
GPT2-Large	43.7	39.5	35.5	32.0	48.1
GPT-Neo	47.4	44.9	39.8	37.3	48.6
GPT-NeoX	46.9	45.3	41.7	39.9	55.3
GPT-J	48.5	45.2	42.8	40.5	55.0
BLOOM	49.1	46.2	42.7	41.5	58.0
OPT	50.2	47.7	43.1	41.8	56.8
LLaMA-1	51.4	**48.5**	46.5	43.0	58.8
Mistral	48.6	47.5	44.6	42.1	59.1
LLaMA-2†	**51.5**	48.4	**46.9**	**43.9**	**59.4**
Mistral-Instruct†	50.1	48.1	45.7	42.5	59.2
PTB-XL	6.5	-	-	0.9	25.6
ECG-Chat	15.9	-	-	2.3	23.9
ECG-Chat-DDP	**32.3**	-	-	**11.2**	**29.9**

For report classification, we conduct two different tasks. On the one hand, we regard each report with a single label, which is the same as we conduct on the MIMIC-ECG dataset. On the other hand, we use the learned representation of generated reports to calculate the similarity with the prompt embedding of targeted categories for zero-shot classification. After conducting an optimal classification threshold search, all categories above this threshold are considered to be predicted, so our method can effectively solve the problem of multi-label ECG zero-shot classification. The experimental results are shown in Table 4.

We can see that for all report classification tasks on the PTBXL dataset, our proposed DERI with DisGPT2 achieves the best performance. Even using the Aligned ECG Encoding instead (DERI-Align) of the Mix Encoding in the DERI framework, our method still shows an obvious improvement on MERL. These results demonstrate that our proposed DERI framework can effectively learn the high-level semantics from the report.

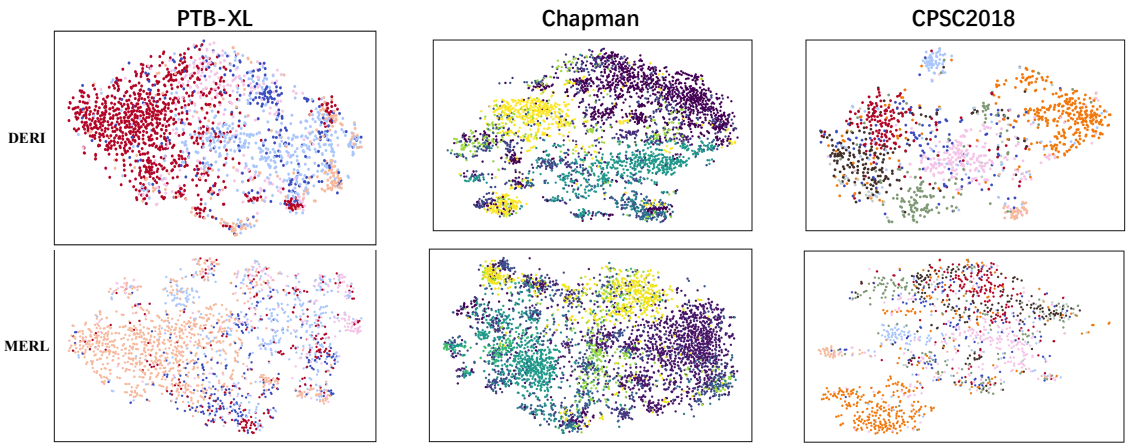

Figure 2: T-SNE visualization of DERI and MERL on three classification datasets.

Table 4: Generated Report Classification on PTBXL dataset.

Single Label	Method	CE		
		F1	PRE	REC
Prompt	MERL-DisGPT2	22.6	27.9	21.5
	DERI-DisGPT2	**40.1**	**46.5**	**38.7**
	DERI-Align-DisGPT2	32.0	38.1	30.0
Multi-label	**Method**	**F1**	**Acc**	**AUC**
Super	MERL-DisGPT2	53.3	66.0	75.7
	DERI-DisGPT2	**56.1**	**72.9**	**76.9**
	DERI-Align-DisGPT2	55.5	72.3	76.1
Sub	MERL-DisGPT2	19.3	85.0	71.1
	DERI-DisGPT2	**21.1**	**86.5**	**72.7**
	DERI-Align-DisGPT2	19.7	85.3	72.2
Form	MERL-DisGPT2	20.8	78.3	62.9
	DERI-DisGPT2	**26.5**	**89.0**	**68.0**
	DERI-Align-DisGPT2	24.8	84.0	66.0
Rhythm	MERL-DisGPT2	18.5	93.3	71.1
	DERI-DisGPT2	**24.1**	**95.2**	**74.5**
	DERI-Align-DisGPT2	23.1	94.0	73.7

5 Representation Visualization

To investigate the learned ECG representation further, we visualize the learned representation of DERI and MERL. We use t-SNE to visualize the representation as Fig. 2. For PTBXL, we adopt the PTBXL-Super setting to obtain the category. For Chapman, we keep the nine categories with the largest sample sizes for better visualization. It can be observed that after t-SNE, the representations learned by our DERI can be more clustered, with greater differentiation between different categories, especially on the CPSC2018 dataset. These results indicate that our method can learn discriminative ECG representations more efficiently than the simple utilization of diagnostic reports by MERL, containing more cardiac clinical information for better classification.

6 Limitations and Future Work.

One potential limitation of our work is that the report used is closer to a clinical semantic description of the signal and remains structurally different from a real diagnostic report. Additionally, we plan to expand our DERI into a more comprehensive cross-modal representation learning model, which can learn from other modal data, such as electronic medical records, further enhancing its relevance in clinical medicine.

References

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[Kiyasseh *et al.*, 2021] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[Liu *et al.*, 2024] Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024.

[Na *et al.*, 2024] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

[Wagner *et al.*, 2020] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[Wang *et al.*, 2023] Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Zbontar *et al.*, 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[Zhang *et al.*, 2023] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.