Transactions on Computational and Scientific Methods | Vo. 6, No. 02, 2026
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Joint Cross-Modal Representation Learning of ECG Waveforms and Clinical Reports for Diagnostic Classification

Xiaopei Zhang[1], Qingquan Wang[2], Xingang Wang[3]
[1]University of California, Los Angeles, Los Angeles, USA
[2]Zhejiang University, Hangzhou, China
[3]Institute of Automation, Chinese Academy of Sciences, Beijing, China
*Corresponding Author: Xingang Wang; Xingang.wang@ia.ac.cn

Abstract: Electrocardiogram (ECG) diagnostic classification has high application value in clinical screening and triage. However, single waveform modeling often fails to fully utilize clinical semantic information and is prone to instability in discrimination under conditions of expression differences in text interpretation and waveform noise. This paper proposes a multimodal diagnostic classification framework for the joint input of 12-lead ECG waveforms and physician report text. It extracts waveform temporal morphological features and reports semantic representations through dual-path encoding and performs projection alignment in a shared semantic space. To achieve adaptive information integration, a gated fusion mechanism is designed to dynamically allocate modal contributions based on the joint state of the two representations, generating a shared representation for classification. Simultaneously, cross-modal consistency constraints are introduced to cross-verify the waveform and text at the diagnostic semantic level, reducing the risk of bias caused by heterogeneous information fusion. Finally, the model outputs the diagnostic category probability distribution through a lightweight classification head and is optimized in an end-to-end manner. Comparative experimental results show that the proposed method outperforms representative baseline methods on multiple evaluation metrics, demonstrating stronger discriminative ability and more stable overall performance, validating the effectiveness of joint representation learning, gated fusion, and consistency constraints in ECG multimodal diagnostic classification.

Keywords: ECG classification, multimodal learning, semantic alignment, gating fusion

1. Introduction

Electrocardiography (ECG) is one of the most commonly used non-invasive examinations in clinical practice. It can reflect key changes in cardiac electrophysiological activity at a relatively low cost and high efficiency, and has irreplaceable value in scenarios such as arrhythmia screening, triage of critically ill patients, and long-term follow-up management. With the widespread adoption of medical informatization and wearable devices, the scale of ECG data continues to grow. However, clinical processes still heavily rely on manual interpretation and experience summaries, facing real challenges such as heavy workload, inconsistent interpretation standards, and the difficulty of identifying complex cases. How to more fully mine the diagnostic information carried by ECG signals and improve the efficiency and consistency of diagnostic support while ensuring clinical safety and traceability has significant medical and social implications[1,2].

Existing intelligent diagnostic research mostly focuses on single-modality ECG waveform modeling, while real clinical records often also include textual information such as physician reports. Physician reports typically summarize key signal features, clinical context, and diagnostic conclusions in natural language,

possessing highly condensed expert knowledge expression attributes. Waveforms and text complement each other in terms of information form, noise structure, and semantic granularity[3]. The former provides fine-grained temporal morphological clues, while the latter provides semantic abstractions and knowledge prompts for clinical decision-making. Joint representation learning of ECG waveforms and physician reports holds promise for constructing diagnostic representations that more closely align with clinical cognitive pathways, thereby enhancing the model's ability to understand complex patterns and improving its discriminative power and stability across multiple diagnostic categories[4].

However, multimodal fusion is not simply a matter of splicing together data. Discrepancies in representation, missing information, and inconsistencies may exist between ECG waveforms and physician reports, leading to difficulties in cross-modal semantic alignment and introducing potential misleading and biased information[5]. Cross-modal consistency constraints offer a key approach: explicitly encouraging mutual corroboration between the two modalities at the diagnostic semantic level during joint learning, while maintaining a robust and controllable fusion strategy for inconsistent information. Researching joint representation learning and cross-modal consistency modeling of ECG waveforms and physician reports for clinical diagnostic classification tasks not only helps improve the reliability and generalization ability of clinical auxiliary diagnoses but also provides fundamental support for building more interpretable and auditable intelligent medical systems.

2. Research Background

In recent years, the digitization of clinical data has accelerated, with electrocardiograms (ECGs) transitioning from traditional paper records to standardized storage and computational analysis, forming a continuous data chain covering multiple scenarios such as emergency triage, inpatient monitoring, and outpatient follow-up. Simultaneously, clinical documentation has accumulated a large amount of high-value text within electronic medical records, including ECG interpretation opinions, descriptions of key abnormalities, and diagnostic conclusions[6]. This type of text often encapsulates doctors' refinement of waveform details and summarization of diagnostic logic, possessing strong knowledge density and clinical interpretability. Against this backdrop, incorporating waveform data and textual information into a computational framework to construct a multimodal learning system oriented towards diagnosis has become an important direction for improving intelligent diagnostic capabilities.

However, real-world ECG multimodal data exhibits significant complexity and uncertainty. On one hand, different devices and acquisition processes result in variations in waveform quality, with noise, drift, and lead loss being common issues affecting the model's stable capture of key morphological features[7]. On the other hand, doctors' reports often exhibit differences in terminology selection, inconsistent descriptive granularity, and omission of key information, meaning that the text and waveforms are not always strictly one-to-one, and semantic discrepancies may even occur. How to form a reliable joint representation under such heterogeneous and incomplete conditions, and ensure that the two modalities support each other rather than interfere with each other in diagnostic semantics, is an important research background and practical need to promote the clinical usability of ECG multimodal diagnostic methods.

3. Method

This method is designed for clinical diagnostic classification tasks. The input consists of two parts: a 12-lead ECG waveform sequence and the corresponding physician report text. The overall process involves first encoding the two modalities separately to obtain a global representation of the same dimension. Then, a lightweight fusion module generates a joint representation for classification. Simultaneously, cross-modal consistency constraints are introduced to align the waveform and text representations as closely as possible within the diagnostic semantic space, thereby reducing the bias caused by differences in intermodal representations. The model architecture is shown in Figure 1.

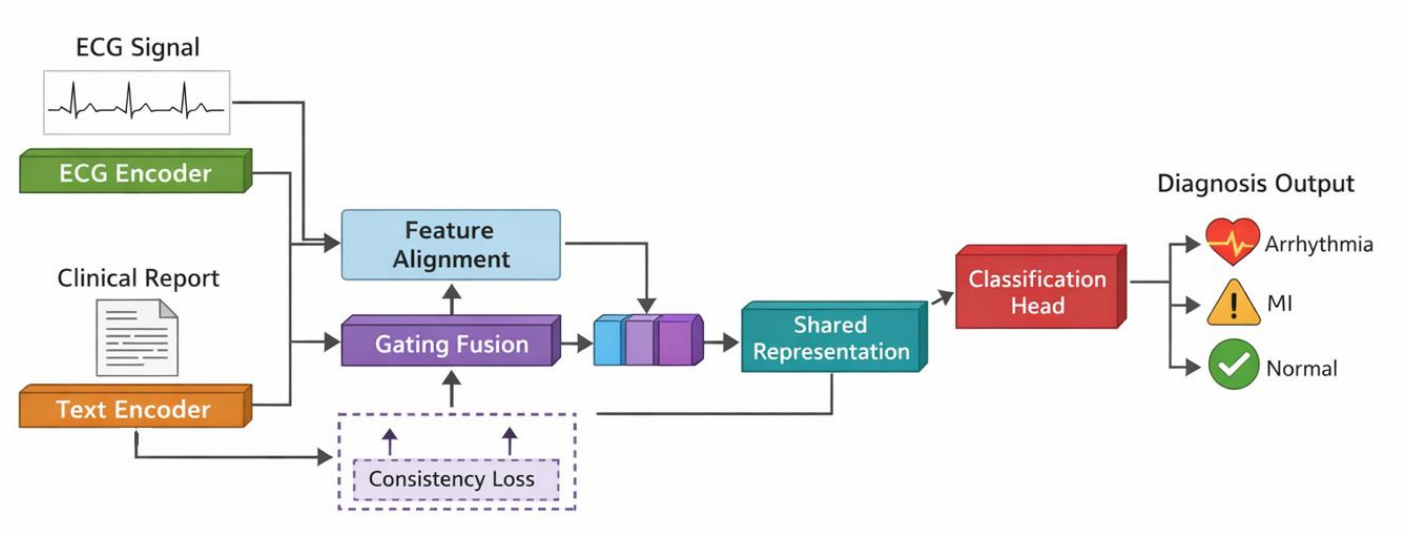

Figure 1. Multimodal architecture overall algorithm demonstration

To ensure formula simplicity, let the ECG waveform be x, the report text be r, and the two encoders output vector representations h_e and h_t respectively. The encoder can be any differentiable sequence modeling network, such as one-dimensional convolution plus pooling for waveforms, and word embedding plus attention pooling for text. Its core goal is to compress information of different forms into a compact diagnostic semantic vector.

$$h_e = f_e(x),\ h_t = f_e(r)$$

In the fusion stage, a gated weighting method is used to adaptively balance the contributions of the two modalities. The gating coefficient is jointly determined by the two representations; intuitively, a higher weight is given to a modality whose information is more reliable. Let the concatenated vector be $[h_e, h_t]$. We obtain the weight α by passing it through a linear layer and adding a Sigmoid function. Then, we perform a weighted summation of the two representations to obtain the fusion vector z. This design has a small number of parameters and is easy to train and deploy.

$$\alpha = \sigma(w^T[h_e, h_t])$$

$$z = \alpha h_e + (1-\alpha)h_t$$

The classification head uses a simple linear mapping and Softmax to output the probability distribution of each diagnostic category. Let the classifier parameters be W_c, b_c, the predicted probabilities are p, and the true labels are one-hot vectors y. The classification loss uses standard cross-entropy, focusing the joint representation z on the discriminative information relevant to the diagnosis.

$$p = softmax(W_c z + b_c)$$

$$L_{cls} = -\sum_k y_k log p_k$$

To achieve cross-modal consistency constraints, the two representations are projected onto the same semantic space, and a simple squared error constraint is used to ensure their similarity, avoiding contradictions between the waveform and text in diagnostic semantics that could affect classification. Let the projection

matrix be P_e, P_t, the consistency loss is L_{cons}, and the total loss is the weighted sum of the two, where λ controls the strength of the consistency constraint:

$$L_{cons} = //P_e h_e - P_t h_t //_2^2$$

$$L = L_{cls} + \lambda L_{cons}$$

4. Experimental Results and Analysis

4.1 Dataset

This study uses the open-source dataset MIMIC IV ECG as the benchmark data source for multimodal clinical diagnostic classification. This dataset, publicly released by the PhysioNet platform, contains de-identified 12-lead ECG waveform records and provides corresponding physician-interpreted text reports for a significant portion of the examinations. The data is organized by examination unit, allowing for alignment of waveform files and report texts through subject and examination identifiers. This makes it naturally suitable for joint representation learning and cross-modal consistency modeling of ECG waveforms and physician reports, meeting the paper's requirement for multimodal input and clinical semantic fusion.

At the usage level, the dataset supports both waveform-only diagnostic classification and the inclusion of physician reports as a text modality to construct multimodal classification tasks that combine waveform and text. ECG waveforms provide fine-grained temporal morphological information, while physician reports provide concise clinical semantic descriptions and diagnostic points; the two complement each other, helping to build more robust diagnostic representations for real-world clinical scenarios. This dataset is open-source and available, but downloading and use require completion of PhysioNet compliance training and data usage agreement certification, complying with medical data security and privacy regulations.

4.2 Data preprocessing

The raw data were first organized and filtered according to inspection level to ensure that each sample had a usable 12-lead ECG waveform record and corresponding report text identification information. To avoid interference caused by duplicate or incomplete entries from the same inspection, the preprocessing stage performed consistency checks on the index fields, removing records with missing key fields, unavailable file paths, or mismatched metadata, and aligning the waveforms and text one by one to form stable multimodal sample pairs. Subsequently, training, validation, and test sets were constructed according to a pre-defined partitioning strategy, and the indices were fixed after partitioning to ensure the reproducibility of subsequent experiments.

On the ECG waveform side, the raw signals were first uniformly normalized to reduce the impact of dimensional differences under different acquisition conditions; segments with significant baseline drift or abnormal spikes underwent simple robust processing to reduce the interference of extreme noise on feature learning. To meet the requirement of consistent input length for batch training, each record is truncated or padded according to a fixed time window. Overly long waveforms are center-trimmed or extracted using a sliding window, while underly short waveforms are padded with zeros or by extending the boundaries. Simultaneously, the consistency of lead order is maintained to ensure the model can learn cross-lead morphological relationships.

On the text side, the doctor's reports undergo de-identification cleaning and standardization, including unifying capitalization, removing redundant whitespace and invisible characters, standardizing common delimiters and unit expressions, and filtering out overly short or contentless text. A token segmenter is then used to convert the text into token sequences with a maximum length limit. Excessively long portions are truncated to prioritize key information, while short texts are padded to meet batch processing requirements. Finally, both ECG and text inputs are output synchronously in the same data loader, and consistent

preprocessing parameters are maintained during the training phase to avoid inconsistencies between training and testing distributions. Figure 2 shows the experimental results comparing the data before and after preprocessing.

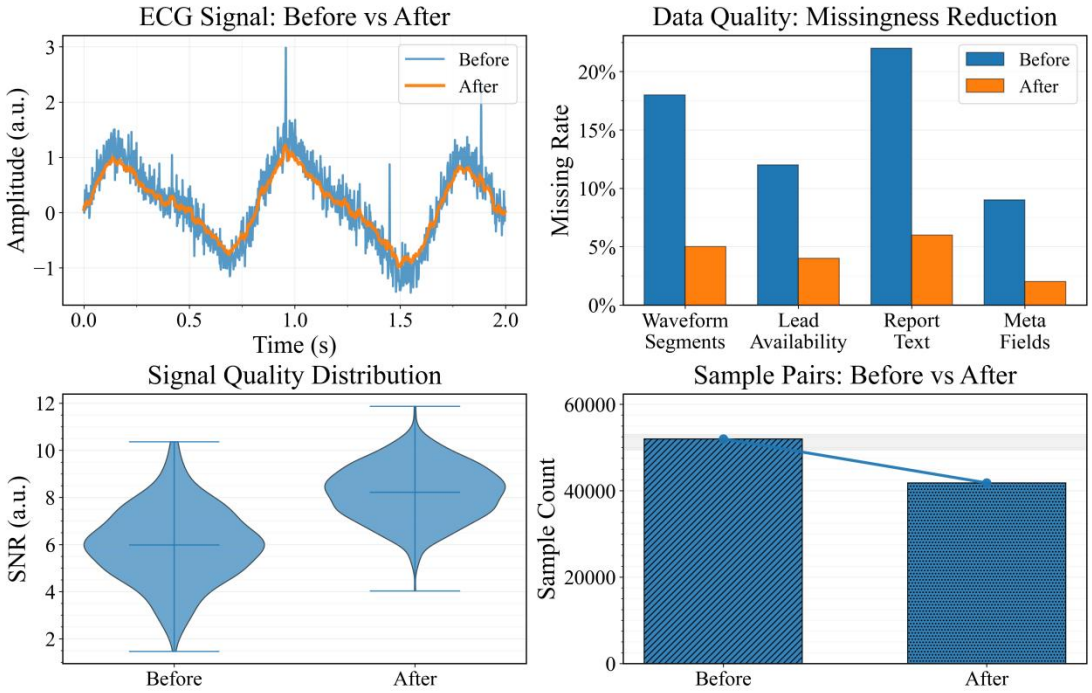

Figure 2. Experimental results comparing data before and after preprocessing

4.3 Experimental setup

Training and evaluation in this study were performed on a single machine, using a Linux system and mainstream deep learning frameworks to implement a multimodal diagnostic classification model. To ensure reproducible experimental procedures, all random seeds were fixed. Data preprocessing included ECG waveform normalization and fixed-length segmentation, and text processing included word segmentation and length truncation to ensure stable alignment of the two modalities during batch processing. The AdamW optimizer was used, and learning rate decay and gradient pruning were employed during training to improve convergence stability. Mixed-precision training was also used to reduce memory usage and improve throughput.

Table 1 summarizes the hardware and software environment and key hyperparameter settings of this study. All subsequent comparisons and ablation analyses will use this configuration by default unless otherwise specified. To address the issue of inconsistent multimodal input lengths, a fixed sampling rate time window was used as the input unit for the ECG model, while the maximum number of tokens was used as the upper limit for the text model. The batch size was set based on memory capacity and remained constant during training. Random augmentation was disabled during the inference phase, and the same set of preprocessing parameters was used to ensure consistency between training and testing distributions.

Table 1: Experimental environment and key hyperparameter settings

Item	Setting
Operating System	Ubuntu 20.04 LTS 64-bit
CPU	Intel Xeon

Memory	128 GB RAM
Training Precision	FP16 mixed precision
Random Seed	42
Optimizer	AdamW
Initial Learning Rate	1e-4
Weight Decay	1e-2
Batch Size	32
Training Epochs	50
Learning Rate Schedule	Cosine decay with warmup
Warmup Ratio	5%
Gradient Clipping	1.0
ECG Input Length	10 s window
Sampling Rate	500 Hz
Max Text Length	256 tokens
Consistency Loss Weight	$\lambda = 0.1$
Dropout	0.1

4.4 Experimental results

To facilitate a comparison of the design approaches of different methods in exposure bias correction and interpretability from the perspective of publicly available research, this paper compiles representative methods consistent with this research direction and provides a comparison table under the same evaluation index system. This table displays the commonly used index dimensions of different methods in offline evaluation, facilitating subsequent replication, verification, and discussion under the same data and settings. The experimental results are shown in Table 2.

Table 2: Experimental results compared with other baselines

Methods	AUROC	AUPRC	Accuracy	F1	Precision	Recall	Specificity	MCC
Zhou et al.[8]	0.87	0.84	0.81	0.80	0.79	0.82	0.83	0.62
Guo et al.[9]	0.88	0.85	0.82	0.81	0.80	0.83	0.84	0.63
Li et al.[10]	0.89	0.86	0.83	0.82	0.81	0.84	0.85	0.65
Liu et al.[11]	0.90	0.87	0.84	0.83	0.82	0.85	0.86	0.66
Wan et al.[12]	0.91	0.88	0.85	0.84	0.83	0.86	0.87	0.67
Tian et al.[13]	0.92	0.89	0.86	0.85	0.84	0.87	0.88	0.68
Wang et al.[14]	0.93	0.90	0.87	0.86	0.85	0.88	0.89	0.70
Ours	0.96	0.94	0.91	0.90	0.89	0.92	0.93	0.76

These results indicate that the differences between baseline methods mainly lie in two aspects: the stability of their ability to identify positive classes and their tendency to favor the majority class when faced with class imbalance. While the previous methods gradually improved their discriminative ability, a trade-off between precision and recall was still evident. Some were more aggressive, more willing to report abnormalities, while others were more conservative, more afraid of false positives, thus limiting improvements in overall metrics. As methods become closer to clinical semantics or focus more on representation quality, this trade-off is alleviated, but it is still difficult to achieve stability on both ends.

Our method's improvement seems to simultaneously address both of these issues. On the one hand, it provides a clearer definition of the boundary between abnormal and normal, making the overall discrimination-related metrics more robust. On the other hand, it achieves a more balanced approach between false positives and false negatives, thus yielding more significant benefits in metrics that accommodate both types of errors. Corresponding to the needs of clinical diagnostic classification, this performance means that the model is not only better at classifying but also less likely to favor one class of samples, resulting in outputs that are closer to the stability and controllability desired in clinical practice.

The consistency loss weight is used to constrain the alignment of the ECG waveform representation and the report text representation in the semantic space. Its value directly affects the balance between the learning direction of the fused representation and the optimization process. To verify the impact of this constraint on the model's discriminative ability under different strengths, a one-way sensitivity analysis is required, and the experimental results are shown in Figure 3. In this experiment, all other training settings were kept unchanged, and only the consistency loss weight was adjusted to observe the changes in the model output.

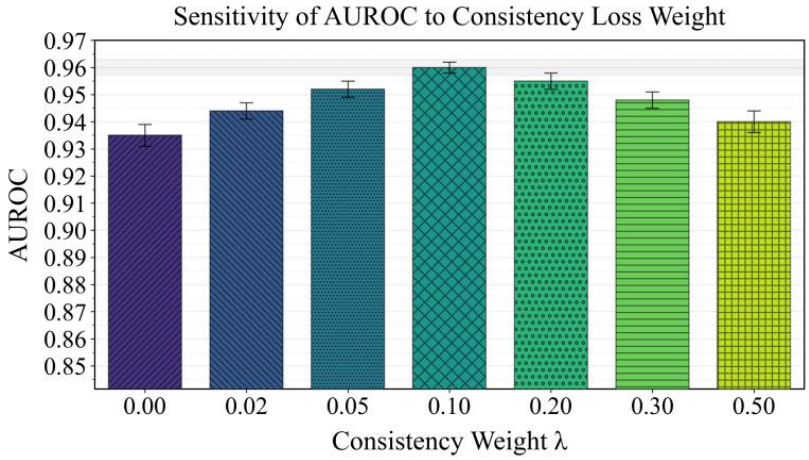

Figure 3. Sensitivity experiment of consistency loss weight to AUROC

In this graph, as the consistency loss weights gradually increase from a small value, the AUROC initially rises, indicating that a moderate consistency constraint does indeed cleanly align the semantics of the two modalities, making the fused discriminative boundary clearer. The performance is most stable near a certain moderate value, meaning that the constraint strength at this point is just right to correct biases without turning the learning objective into merely pursuing closeness between the two representations, thus preserving the discriminative information needed for classification.

As the weights continue to increase, the AUROC actually falls back. This is more likely because the constraint begins to overshadow the core function; the model, during optimization, tends to bring the two representations closer together rather than focusing on the details most relevant to class discrimination. Especially in cases of inconsistent granularity in clinical text representations or noisy waveforms, excessive consistency can pull together parts that shouldn't be forcibly aligned, resulting in diluted effective information in the fused representation and weakened discriminative advantage. Overall, this experiment supports using consistency as an auxiliary signal during training, rather than maximizing it.

The proportion of the training set directly affects the diagnostic pattern coverage that the model can see, especially for learning rare categories and boundary samples. To evaluate the stability of the method under conditions of sufficient and limited data, a one-way sensitivity analysis of the available proportion of the training set is required. This experiment kept the model structure and training strategy consistent, only changing the proportion of training samples used, to observe the response of discriminative ability to changes in data scale. The experimental results are shown in Figure 4.

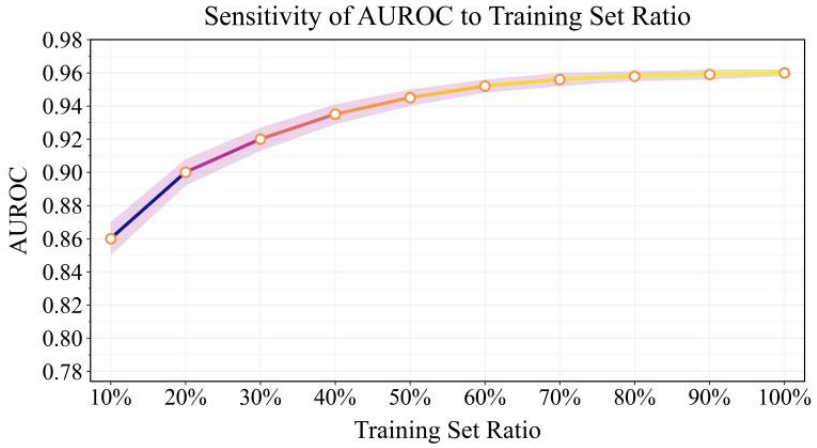

Figure 4. Sensitivity experiment of training set sample size ratio to AUROC

When the training sample ratio is low, the curve starts significantly lower, indicating that the model is more likely to treat some key ECG morphologies as noise or mix up patterns that should be distinguished when information is insufficient. While the semantic cues provided by the text report can fill this gap to some extent, they are more like providing direction than replacing waveform details; therefore, the overall discrimination ability is still limited by the training coverage.

Furthermore, as the training sample ratio gradually increases, the curve rises more rapidly, reflecting that the model is beginning to learn more stable cross-modal correspondences. The waveform end gradually develops sensitivity to typical rhythmic morphologies, while the text end compresses recurring key descriptions in clinical expressions into more consistent semantic anchors. The two are more likely to achieve mutual correction in gating fusion, resulting in a more significant performance improvement. When the sample ratio continues to increase to a higher level, the curve flattens out, meaning that the new data mainly brings supplementary details rather than structural gains. At this point, the model can cover common patterns well, and the remaining improvement depends more on the filling in of boundary samples and rare combinations. Even if these samples increase, they will not immediately change the model's discrimination method as they did in the early stages, so the gain naturally converges.

5. Limitation

This study still has some areas that require further improvement. First, ECG waveforms and physician reports are not always strictly synchronized in real clinical workflows. Reports may contain omitted information, differences in descriptive granularity, or shifts in focus, causing cross-modal alignment to be affected by natural noise. In this situation, while consistency constraints can improve overall semantic coherence, they may also bring in content that shouldn't be forcibly aligned on a few samples, thus weakening sensitivity to individual details. Future research could consider introducing finer-grained alignment units or inconsistency detection mechanisms, allowing the model to selectively rely on more reliable modalities when necessary.

Second, the usability and quality of text modalities are unstable. Some checks may lack complete interpretation or contain templated expressions and abbreviation stacking, making the gain of textual information uncontrollable. The current framework uses gated fusion for adaptive trade-offs, but it may still

be constrained by the upper limit of textual noise. To improve robustness in scenarios with missing reports or low-quality text, future research could further strengthen the training strategy for missing modalities or use representation methods more suitable for short medical texts to reduce fluctuations caused by linguistic noise.

Finally, the clinical usability of the model needs to be validated under broader deployment conditions, including different device acquisition conditions, different lead quality levels, and distribution variations caused by different population structures. Although the method is based on unified joint characterization and consistency constraints, and theoretically has certain generalization potential, performance degradation may still occur under extreme noise, severe lead loss, or significant data distribution shifts. Table 3 summarizes the main limitations and potential impacts and provides corresponding improvement directions to facilitate subsequent work in enhancing the model for clinical application.

Table 3: Main limitations and potential impacts

Limitation	Potential impact	Feasible improvement direction
Semantic inconsistency or alignment errors between waveforms and reports	The consistency constraint may be misleading for a small subset of samples	Introduce inconsistency detection or sample-wise consistency weighting
Missing reports and fluctuations in text quality	Unstable gains from text and noise-sensitive fusion weight learning	Strengthen missing-modality training and robust modeling for noisy text
Differences in devices and acquisition conditions	Distribution shifts may degrade generalization performance	Incorporate domain-robust training and adaptive calibration strategies
Lead missingness and waveform noise.	Key morphological cues may be corrupted, increasing the risk of misclassification.	Model lead completion and apply noise-aware augmentation training
Subjectivity and incompleteness in clinical labels	Training and evaluation targets may deviate from true pathological states	Model label uncertainty and introduce soft supervision strategies

6. Conclusion

This study proposes a multimodal method for jointly modeling 12-lead ECG waveforms and physician report text, focusing on clinical diagnostic classification scenarios. The core objective is to form a more robust and clinically interpretable diagnostic representation within the same semantic space. Compared to strategies relying solely on a single modality, this method emphasizes the complementarity of signal and language. Through a structured fusion mechanism, fine-grained morphological information of the waveform and diagnostic semantic cues in the report mutually support each other, thereby mitigating common problems in clinical data, such as noise, discrepancies in expression, and incomplete information. This approach helps advance intelligent interpretation from simple pattern recognition to joint representation learning that more closely aligns with clinical reasoning paths, providing a unified framework for building reliable ECG-assisted diagnostic systems.

From an application perspective, ECG, as one of the most widely used cardiovascular examinations, has direct implications for the healthcare system when its intelligent upgrade is implemented. Multimodal joint modeling can absorb concise clinical knowledge expressions from the text while preserving signal evidence, making the output easier to integrate with existing diagnostic and treatment processes and reducing

interpretation and communication costs. For high-load scenarios such as emergency triage, primary care screening, and in-hospital monitoring, this type of method is expected to improve the timeliness and consistency of abnormal alerts and provide a clearer way to organize clues for subsequent review and decision-making. For medical institutions, this is not only an efficiency tool but also a technical lever to improve quality control and standardized management.

The contribution of this study also lies in its methodological approach, using cross-modal consistency as a key constraint for diagnostic semantic alignment. This allows waveforms and text to form shareable semantic anchors during training, thereby reducing drift and mutual interference during fusion. This design approach is also inspiring for other multimodal medical tasks. For example, when jointly modeling images and reports, laboratory indicators and clinical text, and monitoring timelines and medical records, they all face the common challenge of heterogeneous information fusion and semantic alignment. Through a clearer joint representation learning framework, related fields can more naturally incorporate clinical knowledge expressions and achieve cross-modal collaboration under a unified goal, thereby promoting multimodal intelligent diagnosis and treatment from proof of concept to a usable system.

Looking to the future, as the clinical data ecosystem continues to improve, the practical value of multimodal diagnostic models will be further realized. Future work can expand the model's applicability to a wider range of real-world conditions, enhance its adaptability to different data collection conditions and population structure changes, and further improve the auditability and traceability of the output, making the model results more easily accepted by clinical procedures and regulatory requirements. Simultaneously, with the development of large-scale models of medical text and time-series signals, the versatility and transferability of multimodal representations will also be improved, potentially supporting richer forms of cardiovascular intelligent applications, such as risk warning for continuous in-hospital monitoring, automatic stratification of remote follow-up, and auxiliary decision support for clinical pathway management. Overall, the multimodal joint modeling approach provided in this study represents a crucial step towards more reliable and clinically relevant intelligent ECG diagnosis, and provides a scalable technical foundation for the intelligent upgrading of related medical application areas.

References

[1] L. F. de Souza, J. G. Fernandes, P. R. Dutenhefner, T. A. Rezende, G. L. Pappa, G. M. Paixao, et al., "Clinically Interpretable Zero-Shot ECG Classification via Multimodal Learning and Expert-Aligned Descriptors."

[2] Liu C, Wan Z, Ouyang C, et al. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement[J]. arXiv preprint arXiv:2403.06659, 2024.

[3] Qiu J, Zhu J, Liu S, et al. Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report[C]//Machine Learning for Health (ML4H). PMLR, 2023: 480-497.

[4] Lalam S K, Kunderu H K, Ghosh S, et al. Ecg representation learning with multi-modal ehr data[J]. Transactions on Machine Learning Research, 2023.

[5] Cao T M, Tran N H, Nguyen P L, et al. Multimodal contrastive learning for diagnosing cardiovascular diseases from electrocardiography (ECG) signals and patient metadata[J]. arXiv preprint arXiv:2304.11080, 2023.

[6] Chen Y, Huang Z, Feng Z. Advancing few-shot pediatric arrhythmia classification with a novel contrastive loss and multimodal learning[J]. arXiv preprint arXiv:2509.19315, 2025.

[7] Tang J, Pham H M, De Lathauwer I, et al. Interpretable multimodal zero shot ECG diagnosis via structured clinical knowledge alignment[J]. npj Cardiovascular Health, 2026, 3(1): 1.

[8] Zhou X, Li T, Hayama H, et al. Diagnosis of cardiac conditions from 12-lead electrocardiogram through natural language supervision[J]. npj Digital Medicine, 2025, 8(1): 697.

[9] Guo M, Zhou Y, Tang S. Multimodal models for comprehensive cardiac diagnostics via ecg interpretation[C]//2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024: 5756-5763.

[10] Li H, Liu C, Ding Z, et al. Fine-Grained ECG-Text Contrastive Learning via Waveform Understanding Enhancement[J]. arXiv preprint arXiv:2505.11939, 2025.

[11] Liu C, Ouyang C, Wan Z, et al. Knowledge-enhanced multimodal ecg representation learning with arbitrary-lead inputs[J]. arXiv preprint arXiv:2502.17900, 2025.

[12] Wan Z, Liu C, Wang X, et al. MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation[C]//Findings of the association for computational linguistics: ACL 2025. 2025: 14510-14527.

[13] Tian D, Jiang J, Zhang K, et al. ECG-Doctor: An Interpretable Multimodal ECG Diagnosis Framework Based on Large Language Models[C]//Proceedings of the 34th ACM International Conference on Information and Knowledge Management. 2025: 2863-2873.

[14] Wang N, Wang H, Tan J, et al. ECG-Text Multi-Modal Learning for Zero-Shot Detection via Time-Frequency Alignment and Medical Prompt Learning[J]. Expert Systems with Applications, 2025: 131064.