

SARL-ICD: Structure-aware Contrastive Representation Learning with LLM-augmented Label Descriptions for Automatic ICD Coding

Jincheng Shi^{1,2}, Tianyang Huang², Shanfeng Zhu^{2*}

¹School of Data Science, Fudan University

²Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University
25210980093@m.fudan.edu.cn zhusf@fudan.edu.cn

Abstract

Automatic ICD coding is an extreme multi-label classification task that requires assigning standardized diagnostic codes to clinical notes. Recent pretrained language model (PLM) based methods achieve strong performance, but they often overfit to repetitive semi-structured patterns in discharge summaries rather than learning clinically meaningful semantics. Moreover, existing approaches insufficiently utilize the rich semantic information contained in ICD code descriptions. In this study, we propose SARL-ICD, a structure-aware framework for automatic ICD coding centered around two key ideas: contrastive representation learning and label-text semantic interaction. First, we introduce JMoCo, a Jaccard-guided MoCo-style pretraining strategy that constructs positive pairs using ICD-code similarity while restricting pretraining to clinically informative narrative sections. This design encourages the model to learn patient-specific clinical semantics. Second, we propose a lightweight label-text interaction module that injects semantic information from LLM-augmented ICD code descriptions into document representations. We further employ an adaptive fusion module to combine complementary strengths of different ICD coding architectures. Extensive experiments on MIMIC-III-full and MIMIC-IV demonstrate that our method achieves state-of-the-art performance across multiple evaluation metrics. Additional analyses further show that JMoCo improves representation similarity alignment and reduces structural shortcut learning in semi-structured clinical documents.

1 Introduction

Automatic International Classification of Diseases (ICD) coding aims to automatically assign standardized diagnostic and procedure codes to clinical notes. As a core component of hospital billing,

clinical research, and healthcare analytics, accurate ICD coding is essential for large-scale medical systems (Xie and Xing, 2018). However, manual coding remains labor-intensive, time-consuming, and error-prone, motivating extensive research on automatic ICD coding.

Existing ICD coding methods generally formulate the task as an extreme multi-label text classification (XML) problem. Early approaches relied on convolutional or recurrent architectures with label-wise attention mechanisms (Mullenbach et al., 2018; Vu et al., 2020), while recent state-of-the-art systems increasingly adopt pretrained language models (PLMs), such as PLM-ICD (Huang et al., 2022) and PLM-CA (Edin et al., 2024). Despite their strong performance, current approaches still suffer from three fundamental limitations.

First, clinical discharge summaries exhibit highly repetitive semi-structured patterns (Lu et al., 2023). Sections such as “hospital course”, “physical examination”, and “discharge diagnosis” often follow nearly identical templates across patients. While PLMs are effective at modeling contextual similarity, they can easily overfit to these superficial structural regularities instead of learning clinically meaningful patient-specific semantics. As a result, representations of clinically distinct cases may become overly similar simply because they share common document templates.

Second, most existing ICD coding models underutilize the semantic information contained in ICD code descriptions. Although recent studies attempt to inject label descriptions into prediction modules (Yang et al., 2022; Wu et al., 2025b), these approaches are often tightly coupled with specific architectures or rely on limited textual descriptions that fail to fully capture label semantics.

Third, different ICD coding architectures exhibit strong but complementary inductive biases, but existing methods rarely exploit this architectural complementarity in a principled way.

*Corresponding author.

In this study, rather than proposing another larger ICD coding architecture, we focus on a more fundamental question: **How can we learn clinically discriminative representations from highly structured medical documents while effectively leveraging label semantics ?**

To address this problem, we propose **SARL-ICD** (Structure-Aware Contrastive Representation Learning with LLM-augmented Label Descriptions for Automatic ICD Coding), a framework centered around two key ideas: structure-aware contrastive pretraining and label-text semantic interaction. The overall framework is illustrated in Figure 1.

First, we propose **JMoCo**, a Jaccard-guided MoCo-style contrastive pretraining strategy specifically designed for semi-structured clinical documents. We leverage augmented ICD-code Jaccard similarity to identify clinically related discharge summaries. JMoCo performs pretraining only on clinically informative narrative sections, such as the history of present illness and past medical history, while excluding highly templated hospitalization and discharge sections. Experimental analyses show that this strategy substantially improves representation similarity alignment and downstream ICD coding performance, suggesting that the learned embeddings better capture patient-specific clinical semantics rather than superficial document structure.

Second, we introduce a lightweight and architecture-agnostic **label-text interaction module**. We augment ICD code descriptions using a large language model and encode them into dense semantic representations. These label embeddings interact directly with document token representations through a similarity-based interaction mechanism, producing label-aware semantic features that complement the base encoder representations. Unlike prior methods that tightly integrate label descriptions into specific architectures, our module can be seamlessly plugged into different ICD coding frameworks and consistently improves both PLM-CA and PLM-ICD.

Finally, motivated by the complementary strengths of different attention mechanisms, we introduce a lightweight MoE-based fusion module to adaptively combine the predictions of PLM-CA and PLM-ICD. PLM-CA uses cross-attention to capture clinical evidence distributed across different sections of long clinical notes, while PLM-ICD employs label-wise attention with semantic ICD representations to better model the correspondence

between clinical evidence and label semantics. The fusion module adaptively leverages these complementary attention paradigms, combining their respective strengths for more robust ICD code prediction.

Our contributions are summarized as follows:

- **We propose JMoCo, a structure-aware contrastive pre-training framework for automatic ICD coding.** Through representation similarity analysis and controlled section-selection experiments, we identify that highly templated sections (e.g., hospital course) can dominate contrastive objectives and reduce representation discriminability. To address this, JMoCo combines augmented ICD-code similarity with narrative-focused truncation, restricting pre-training to clinically informative narrative sections. This prevents shortcut learning and enables the model to learn more discriminative patient-specific representations. To the best of our knowledge, this is the first work to explicitly address shortcut learning in contrastive pre-training for ICD coding.
- **We propose an architecture-agnostic label-text interaction module based on LLM-augmented code semantics.** The proposed module can be integrated into different ICD coding architectures and consistently improves downstream performance by enhancing semantic alignment between labels and clinical text.
- **Our framework achieves state-of-the-art performance on multiple benchmark datasets.** Extensive experiments on MIMIC-III-full and MIMIC-IV demonstrate that our method consistently outperforms existing ICD coding systems while providing improved representation quality.

2 Related Work

2.1 ICD Coding

ICD coding employs natural language processing (NLP) models to predict expert-annotated ICD codes from discharge summaries. With the increasing availability of large-scale electronic health record (EHR) datasets, the most widely used benchmarks are MIMIC-III (Johnson et al., 2016; Goldberger et al., 2000) and MIMIC-IV (Johnson et al., 2016; Goldberger et al., 2000). The predominant

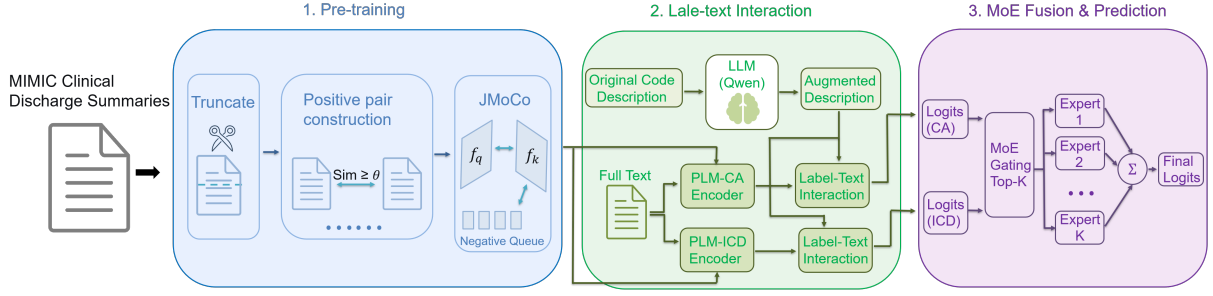


Figure 1: Overview of the proposed SARL-ICD framework. Our approach consists of three key components. (1) Pretraining, where clinically similar discharge summaries are identified and truncated. A MoCo-style contrastive objective is then applied to learn clinically discriminative representations. (2) Adding label-text semantic interaction, where ICD code descriptions are augmented and encoded. A lightweight label-text interaction module is integrated into both PLM-CA and PLM-ICD during downstream fine-tuning. (3) Lightweight MoE-based fusion, where the Logits from PLM-CA and PLM-ICD are adaptively combined using a mixture-of-experts module to leverage their complementary strengths across different ICD categories.

approach in current literature is to leverage a language model to encode clinical notes, followed by a label attention or cross-attention mechanism that aligns ICD codes with the input text for prediction (Mullenbach et al., 2018). The underlying language models have evolved from convolutional neural networks (CNNs) and recurrent neural networks (RNNs), such as LAAT (Vu et al., 2020) and MultiResCNN (Li and Yu, 2020), to transformer-based architectures, including pre-trained language models (PLMs) such as PLM-ICD and PLM-CA. Other PLM based models for XML like BERTMESH (You et al., 2021) can also be used on this task. However, most of these models operate solely on the raw text of clinical notes and ignore the rich semantic information contained in code descriptions.

To take advantage of label information in ICD coding, several studies have attempted to incorporate label description information into the prediction process. Early work by Xie and Xing (2018) encodes code descriptions using a tree-of-sequences LSTM to capture hierarchical structure. Li and Yu (2020) randomly initialized label representations, while Zhou et al. (2021) initialized them with code descriptions using Word2Vec. More recently, Wu et al. (2025b) enhanced code descriptions with LLM-generated synonyms and applies contrastive learning to align these synonyms with the original descriptions, followed by label attention to compute text-label similarity for prediction. Despite these advances, existing methods either fail to provide sufficiently rich label description information to effectively enhance model capacity, or lack generalizability, making them difficult to transfer to other model architectures.

In this paper, we propose a novel label-text interaction mechanism. Specifically, we leverage an LLM to enrich ICD code descriptions and use BERT to encode these enriched descriptions as label representations. Our label-text interaction module is designed to be easily pluggable into existing methods such as PLM-ICD and PLM-CA, consistently improving their performance.

2.2 Multi-Label Contrastive Learning

Contrastive learning has emerged as a powerful paradigm for representation learning. He et al. (2020) introduced an unsupervised contrastive learning method, while Khosla et al. (2020) first adapted contrastive loss from a self-supervised to a supervised setting by incorporating label information. Building on these foundations, Zhang and Wu (2024) proposed a multi-label supervised contrastive loss (MulSupCon) that treats each label independently. It constructed separate positive sets for each class to which an anchor belongs and aggregates gradients across all labels to push the anchor toward the mean representation of each of its positive classes. However, this multi-label contrastive learning method is primarily suitable for datasets with a modest number of labels and does not scale well to extreme multi-label datasets.

Recently, several studies have employed synonym-based strategies to conduct biomedical entity representation learning (Liu et al., 2021; Lai et al., 2021). For instance, Liu et al. (2021) introduces an additional pre-training scheme that self-aligns the representation space of biomedical entities derived from a pre-trained medical LM.

In the context of automatic ICD coding, con-

trastive learning has also been leveraged. Yang et al. (2022) pre-trained a Longformer via hierarchical self-alignment (HSAP) using the UMLS knowledge graph and a hierarchical contrastive loss. Kumar et al. (2023) proposed a transfer learning framework that pre-trains a hierarchical BERT-Bi-LSTM model on MIMIC data using a contrastive objective, and then fine-tunes it on internal clinical data for ICD coding. Wu et al. (2025b) applied contrastive learning to align them with original descriptions, thereby enhancing few-shot label prediction. However, most existing contrastive learning approaches in this domain focus on label representations. To complement these efforts, we propose a contrastive learning method that pulls together the contextual representations of clinically similar documents while distinguishing dissimilar ones.

3 Method

ICD coding is a multi-label classification task. Given an input medical note t consisting of thousands of words, the task is to assign a binary label $y_i \in \{0, 1\}$ for each ICD code in the label space $\mathcal{Y}$, where $y_i = 1$ indicates that the note is positive for a specific ICD disease or procedure, and $i \in [1, N_c]$. Here, $N_c = |\mathcal{Y}|$ denotes the total number of ICD codes. Each ICD code is associated with a short free-text description phrase c_i . For example, code 00.96 has the description “Infusion of 4-Factor Prothrombin Complex Concentrate”. For convenience, for each training document t_i , we denote its ground-truth ICD code set by $\mathcal{Y}_i \subseteq \mathcal{Y}$, where $\mathcal{Y}_i$ contains all ICD codes assigned to t_i . The set of all code descriptions is denoted as $\mathcal{C} = \{c_1, c_2, \dots, c_{N_c}\}$. The whole framework of our proposed method is shown in Figure 1.

3.1 Jaccard-guided Contrastive Pre-training

Clinical discharge summaries exhibit a semi-structured organization, typically comprising sections such as “chief complaint”, “history of present illness”, “laboratory findings”, “hospital course”, and “discharge diagnosis” (Lu et al., 2023). While this structured format facilitates information extraction, it also introduces a risk of overfitting: sections such as physical examination, hospital course, and surgical details follow highly similar linguistic patterns across different patients, causing contextual encoders to attend to these shared templates rather than clinically meaningful distinctions.

To address this challenge, we propose a con-

trastive pre-training framework that focuses on two key principles. First, we construct positive pairs based on augmented Jaccard similarity of ICD code sets, incorporating hierarchical label information. Second, we limit the model’s input to the clinical narrative sections, namely the chief complaint, history of present illness, and past medical history—while excluding the subsequent highly structured sections. This design forces the model to learn representations that capture true clinical similarity based on patient-specific narrative content, avoiding overfitting to redundant structural templates. Analyses in Table 1 and Table 4 further support this hypothesis by showing that narrative-focused pre-training improves both representation similarity alignment and downstream ICD coding performance.

3.1.1 Augmented Jaccard Similarity for Positive Pair Construction

A naive application of Jaccard similarity on raw ICD code sets fails to capture semantic relationships between related codes. For instance, codes 251.1 (Other specified hypoglycemia) and 251.2 (Hypoglycemia, unspecified) share no overlap in their literal code strings, yet are clinically related. To address this, we extend each document’s label set to include all ancestor codes in the ICD hierarchy. Specifically, we first build a complete ICD ancestry tree, where each node’s ancestors include the node itself, its parent chapter, and all higher-level categories (e.g., 001.0 $\rightarrow$ 001 $\rightarrow$ 001-009 $\rightarrow$ 001-139).

Given a document t_i (the i -th medical note in the training dataset $\mathcal{D}_{\text{train}} = \{t_1, t_2, \dots, t_N\}$) with ground-truth label set $\mathcal{Y}_i \subseteq \mathcal{Y}$, we compute its augmented label set by incorporating all ancestor codes:

$$\tilde{\mathcal{Y}}_i = \mathcal{Y}_i \cup \bigcup_{y \in \mathcal{Y}_i} \text{Ancestors}(y) \quad (1)$$

where y denotes an individual ICD code, and $\text{Ancestors}(y)$ returns the set of all ancestor codes of y in the ICD hierarchy (including y itself). The similarity between two documents t_i and t_j is then computed using the augmented Jaccard coefficient:

$$s_{ij} = \frac{|\tilde{\mathcal{Y}}_i \cap \tilde{\mathcal{Y}}_j|}{|\tilde{\mathcal{Y}}_i \cup \tilde{\mathcal{Y}}_j|}, \quad s_{ij} \in [0, 1] \quad (2)$$

where i and j are indices of two distinct documents in the dataset. We construct positive pairs

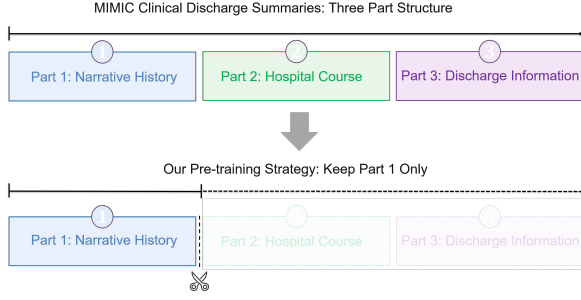


Figure 2: Clinical narrative extraction for JMoCo pre-training. We only keep part1:Narrative History for the pretraining process.

for contrastive learning by selecting document pairs with $s_{ij} \geq \theta$, where θ is a manually specified threshold and its optimal value is determined through experiments. Each pair is additionally weighted by its similarity score s_{ij} during training, allowing higher-weight pairs to contribute more to the contrastive objective.

3.1.2 Focusing on Clinical Narrative Sections

Inspired by Lu et al. (2023), we adopt a keyword-based section parsing strategy to segment clinical discharge summaries into three thematic blocks: Narrative History, Hospital Course, and Discharge Information. As illustrated in Figure 2, our pre-training strategy focuses exclusively on the Narrative History section, which contains the most informative signals for ICD code assignment, while excluding the templated hospital course and discharge information. This truncation is applied only during pre-training. While during SARL-ICD fine-tuning, we use full-length documents. The complete algorithmic details, including keyword definitions and the rationale for this design choice, are provided in Appendix A. Empirical results confirm that this focused strategy consistently outperforms processing full documents (see Table 4).

3.1.3 MoCo Pre-training with Similarity-weighted Loss

We adopt a MoCo-style (He et al., 2020) contrastive pre-training framework with momentum-updated encoders and a dynamic negative sample queue. Positive pairs are constructed using augmented Jaccard similarity of ICD code sets. The architecture consists of a query encoder f_q and a momentum-updated key encoder f_k , along with a dynamic queue storing negative samples.

For each positive pair (i, j) with augmented Jac-

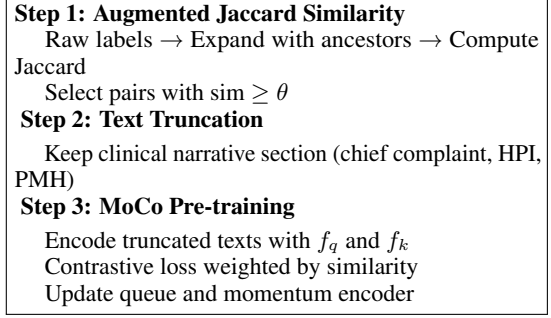


Figure 3: Overview of the proposed Jaccard-guided contrastive pre-training framework.

card similarity $s_{ij} \geq \theta$, we define our pretraining dataset:

$$\mathcal{P} = \{(i, j) \mid s_{ij} \geq \theta\}. \quad (3)$$

Then we let x_i denotes the raw text of the i -th medical note, and $\text{Trunc}(\cdot)$ extracts the clinical narrative sections. The query and key representations are then computed as $q_i = f_q(\text{Trunc}(x_i))$ and $k_j = f_k(\text{Trunc}(x_j))$. Following MoCo, we maintain a first-in-first-out (FIFO) queue of key representations produced by the momentum encoder f_k .

The contrastive loss for a single pair is:

$$\ell_{ij} = \log \left(1 + \sum_{n \in \mathcal{N}} \exp \left(\frac{q_i^\top (k_n - k_j)}{\tau} \right) \right). \quad (4)$$

where τ is a temperature hyperparameter and $\mathcal{N}$ denotes the set of key representations stored in the FIFO memory queue, excluding the current positive key k_j . The final pre-training loss is the similarity-weighted average over all positive pairs:

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} s_{ij} \cdot \ell_{ij} \quad (5)$$

This weighting scheme encourages the model to focus more on pairs with higher clinical similarity. The complete pre-training pipeline is illustrated in Figure 3.

3.2 Analysis of JMoCo Pre-training

To further understand how JMoCo shapes clinical representations, we analyze the alignment between document embeddings and ICD-code semantics. Specifically, we measure the Pearson and Spearman correlations between pairwise embedding similarity and the corresponding augmented ICD-code similarity.

Table 1: Correlation between embedding similarity and augmented ICD-code similarity on MIMIC-III-full.

Method	Pearson r	Spearman ρ
JMoCo w/o truncation	0.0115	-0.0802
JMoCo w/o weighting	0.1021	0.1846
JMoCo (ours)	0.1463	0.2209

As shown in Table 1, JMoCo achieves the strongest alignment between representation similarity and clinical label similarity. Removing narrative truncation leads to particularly weak alignment, with a negative Spearman correlation, while removing similarity weighting also reduces the alignment. These results provide representation-level evidence that narrative-focused pre-training and similarity-aware weighting help JMoCo capture clinically meaningful label semantics rather than superficial document similarity. To further examine the improvement, we conduct a controlled zero-shot retrieval experiment reported in Appendix C.

3.3 Label-Text Interaction Enhanced Encoder

After the pretraining, to enrich the semantic representation of medical codes, we introduce a lightweight label-text interaction module that can be seamlessly integrated into existing ICD coding architectures. Our approach leverages large language models (LLMs) to augment ICD code descriptions, which are then encoded into dense representations using a pre-trained sentence encoder.

Specifically, for each ICD code description c_i , we prompt Qwen-7B-Chat to generate an augmented description. The prompt template and generation hyperparameters are detailed in Appendix G. The original description and its augmented version are encoded via a Sentence-BERT model to obtain a fixed-dimensional embedding $\mathbf{e}_i \in \mathbb{R}^d$, where d denotes the hidden embedding dimension. These label embeddings are precomputed and kept frozen during fine-tuning to preserve semantic integrity.

The proposed label-text interaction module operates in parallel with the base encoder’s existing attention mechanism. Given a document representation $\mathbf{H} \in \mathbb{R}^{L \times d}$ (where L is the sequence length after chunking), we compute a similarity matrix between label embeddings and text tokens:

$$\mathbf{S}_{ij} = \frac{\mathbf{e}_i \cdot \mathbf{h}_j}{\|\mathbf{e}_i\| \|\mathbf{h}_j\|}, \quad \mathbf{S} \in \mathbb{R}^{N_c \times L} \quad (6)$$

where $N_c = |\mathcal{Y}|$ is the total number of ICD codes.

This matrix captures token-level relevance for each code. To obtain a fixed-dimensional interaction feature regardless of input length, we apply adaptive average pooling along the sequence dimension followed by a linear projection:

$$\mathbf{f}_i = \text{Linear}(\text{AdaptiveAvgPool}(\mathbf{S}_{i,:})) \in \mathbb{R}^{d_{\text{interact}}} \quad (7)$$

The resulting interaction features $\mathbf{F} \in \mathbb{R}^{N_c \times d_{\text{interact}}}$ are then concatenated with the base model’s label-aware contextual representations before the final classification head. For example, in our enhanced PLM-CA framework (PLM-CA-L), given the cross-attention output $\mathbf{Y} \in \mathbb{R}^{N_c \times H}$ and interaction features $\mathbf{F} \in \mathbb{R}^{N_c \times d_{\text{interact}}}$, the final representation is:

$$\tilde{\mathbf{Y}} = [\mathbf{Y} \parallel \mathbf{F}] \in \mathbb{R}^{N_c \times (d + d_{\text{interact}})} \quad (8)$$

which is then passed through a linear classifier to obtain prediction logits $\hat{\mathbf{y}} \in \mathbb{R}^{N_c}$.

This module offers two key advantages. First, it injects rich semantic knowledge from LLM-augmented descriptions without increasing the complexity of the base architecture. Second, it is architecture-agnostic and can be readily incorporated into various existing models. In this work, we integrate it into both PLM-CA and PLM-ICD frameworks, demonstrating its generalizability.

3.4 Mixture-of-Experts Fusion Framework

Our empirical analysis reveals that PLM-CA and PLM-ICD exhibit complementary specialization patterns across different ICD categories. PLM-CA excels on conditions with strong temporal or procedural patterns, such as perinatal conditions and congenital anomalies, while PLM-ICD performs better on chronic diseases with rich semantic descriptions, including circulatory and endocrine disorders. These complementary patterns are consistently observed across MIMIC-III-full and MIMIC-IV ICD-9, as detailed in Appendix D.

To harness their synergistic effects, we are inspired by Wu et al. (2025a) and propose a Mixture-of-Experts (MoE) fusion framework that adaptively combines the predictions from both encoders. The MoE module learns to dynamically favor PLM-CA or PLM-ICD according to the input instance and ICD label, effectively leveraging their complementary strengths for improved overall performance.

As illustrated in Figure 1, our fusion framework takes the output logits from both models as input.

For each document and each ICD code, we have a pair of logits (ℓ_i^P, ℓ_i^B) , where P represents PLM-CA (Edin et al., 2024), B represents PLM-ICD (Huang et al., 2022). A lightweight MoE module learns to dynamically weight these two sources:

$$\tilde{\ell}_i = \sum_{k=1}^K g_k(\ell_i^P, \ell_i^B) \cdot \text{Expert}_k(\ell_i^P, \ell_i^B) \quad (9)$$

where K is the number of experts, Expert_k is a two-layer MLP that maps the pair of logits to a scalar, and g_k is a learnable gating function implemented as:

$$\mathbf{g} = \text{TopK}(\text{Softmax}(\mathbf{W}_{\text{gate}} \cdot [\ell^P, \ell^B])) \quad (10)$$

Following common MoE practice, we apply top- k sparsification to the gate outputs, activating only the most relevant experts per input. The entire framework proceeds in two stages: (1) JMoCo pre-training and downstream fine-tuning of PLM-CA and PLM-ICD; (2) MoE fusion training based on the learned representations of both models. Empirical results demonstrate that this hierarchical fusion strategy consistently outperforms simple averaging and fixed weighting baselines.

4 Experiments

4.1 Datasets

We conduct our experiments on the publicly available benchmark MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) dataset. There are 52,722 discharge summaries and 8,692 unique ICD-9 codes in the MIMIC-III dataset. The MIMIC-IV dataset is meticulously annotated with both ICD-9 and ICD-10 codes. Edin et al. (2023) constructed the MIMIC-IV ICD-9 and MIMIC-IV ICD-10 datasets using stratified sampling techniques. The MIMIC-IV ICD-10 dataset contains 122,278 discharge summaries and 7942 unique ICD-10 codes while MIMIC-IV ICD-9 dataset contains 209,326 discharge summaries and 6150 unique ICD-9 codes. For our analysis, we select three datasets: MIMIC-III full, MIMIC-IV ICD-9 and ICD-10 datasets.

4.2 Implementation Details

We implement our model in PyTorch on a single NVIDIA 4090 24G GPU. We set the initial learning rate as $5e-5$ with batch size 16. To improve

the prediction performance of ICD coding, we utilize a domain specific PLM, called RoBERTa-PM (Lewis et al., 2020). For evaluating the performance of our proposed model, we employ four commonly used metrics: AUC-ROC (Micro and Macro), F1-score (Micro and Macro), and precision at K (P@8 and P@15). Complete implementation details and hyperparameter configurations for all datasets are provided in Appendix G. Our code is available at <https://github.com/ZhuLab-Fudan/SARL-ICD>.

4.3 Baselines

This section introduces the baselines we used in the evaluation.

CAML (Mullenbach et al., 2018) is a classic benchmark of ICD coding using a convolutional neural network and attention mechanism.

LAAT (Vu et al., 2020) adds a hierarchical joint learning mechanism on label attention to enhance the classifications of rare-occurred codes.

PLM-ICD (Huang et al., 2022) proposes a PLM framework to solve the problem of imbalanced labels and long input texts. Edin et al. (2023) report that PLM-ICD outperforms all baseline models.

MSMN (Yuan et al., 2022) utilizes the importance of synonyms for improved code representation learning and proposes a synonyms matching network for code classification.

PLM-CA (Edin et al., 2024) is a variant of PLM-ICD, it employs a RoBERTa-PM encoder with a label-wise cross attention layer.

PLM-LLM (Wu et al., 2025b) uses GPT-4 to generate synonymous code descriptions and applies contrastive learning to align them with original descriptions, enhancing few-shot label prediction.

4.4 Compare with SOTA Methods

To evaluate the effectiveness of SARL-ICD, we compare it with existing state-of-the-art models on the MIMIC-III-full and MIMIC-IV-ICD10 datasets. To ensure a fair comparison, we re-implement all baseline methods based on their publicly available source code and evaluate them under a unified experimental pipeline. Following prior work, primarily Edin et al. (Edin et al., 2023), we maintain consistent data splits, text and ICD preprocessing, evaluation protocols, and backbone configurations across all methods. Table 2 presents the comparative results, averaged over multiple seeds.

Results on MIMIC-III-full. SARL-ICD achieves a balanced and superior performance

Table 2: Performance comparison on MIMIC-IV ICD-10 and MIMIC-III-full datasets.

Models	MIMIC-IV ICD-10						MIMIC-III-full					
	F1		P@K		AUC		F1		P@K		AUC	
	Macro	Micro	P@8	P@15	Macro	Micro	Macro	Micro	P@8	P@15	Macro	Micro
CAML (Mullenbach et al., 2018)	0.210	0.584	0.668	0.522	0.910	0.982	0.087	0.534	0.709	0.560	0.895	0.986
DR-CAML (Mullenbach et al., 2018)	—	—	—	—	—	—	0.086	0.528	0.690	0.548	0.897	0.984
LAAT (Vu et al., 2020)	0.222	0.574	0.684	0.539	0.944	0.989	0.090	0.574	0.738	0.588	0.895	0.985
Joint-LAAT (Vu et al., 2020)	—	—	—	—	—	—	0.101	0.575	0.735	0.587	0.900	0.985
MSMN (Yuan et al., 2022)	—	—	—	—	—	—	0.103	0.584	0.751	0.598	0.948	0.992
PLM-ICD (Huang et al., 2022)	0.211	0.585	0.699	0.550	0.966	0.992	0.104	0.598	0.771	0.613	0.926	0.989
PLM-CA (Edin et al., 2024)	0.261	0.596	0.709	0.561	0.962	0.992	0.119	0.600	0.766	0.614	0.913	0.987
PLM-LLM (Wu et al., 2025b)	—	—	—	—	—	—	0.103	0.613	0.779	0.615	0.952	0.988
SARL-ICD	0.286	0.611	0.723	0.573	0.974	0.992	0.123	0.618	0.786	0.631	0.954	0.990

across both macro and micro metrics. Notably, existing SOTA models exhibit a trade-off: PLM-CA achieves relatively high macro-F1 (11.9%) but lower micro-F1 (60.0%), while PLM-LLM achieves high micro-F1 (61.3%) but much lower macro-F1 (10.3%). In contrast, SARL-ICD simultaneously achieves the highest macro-F1 (12.3%, outperforming PLM-CA by 3.4% relatively) and the highest micro-F1 (61.8%, outperforming PLM-LLM by 0.8% relatively), demonstrating its ability to excel on both frequent and rare codes. For ranking-based metrics, SARL-ICD attains 78.6% in P@8 and 63.1% in P@15, consistently surpassing all baselines.

Results on MIMIC-IV ICD-10. On the larger and more challenging MIMIC-IV ICD-10 dataset, SARL-ICD demonstrates even more significant improvements, particularly in macro-F1. SARL-ICD achieves a macro-F1 of 28.6%, representing a **9.6% relative improvement** over PLM-CA (26.1%) and a **35.5% relative improvement** over PLM-ICD (21.1%). For micro-F1, SARL-ICD reaches 61.1%, outperforming PLM-CA (59.6%) and PLM-ICD (58.5%). In P@8 and P@15, SARL-ICD achieves 72.3% and 57.3%, outperforming the best baseline by 1.4% and 1.2% absolute, respectively.

Overall, SARL-ICD consistently outperforms existing SOTA methods across both datasets, with particularly notable gains in macro-F1, a critical metric for imbalanced multi-label classification. Additional results on the MIMIC-IV ICD-9 dataset are provided in Appendix B. We additionally analyze the computational cost and inference overhead of SARL-ICD in Appendix E.

4.5 Ablation Experiments

4.5.1 Overall Ablation

To investigate the effectiveness of each proposed component, we conduct ablation experiments on the MIMIC-IV ICD-10 dataset. Table 3 shows

Table 3: Ablation study on MIMIC-IV ICD-10 dataset. “Label-Text” denotes label-text interaction module, “PT” denotes pretrained weights.

Models	F1 Macro	F1 Micro	P@8	P@15
PLM-CA (baseline)	0.261	0.596	0.709	0.561
+ Label-Text (not augmented)	0.265	0.596	0.707	0.558
+ Label-Text (des augmented)	0.270	0.597	0.709	0.561
+ PT	0.272	0.596	0.709	0.561
+ Label-Text + PT	0.277	0.599	0.713	0.563
+ Label-Text + PT(not augmented)	0.273	0.598	0.711	0.561
PLM-ICD (baseline)	0.211	0.585	0.699	0.550
+ Label-Text (not augmented)	0.255	0.588	0.705	0.555
+ Label-Text (des augmented)	0.263	0.596	0.715	0.564
+ PT	0.260	0.596	0.714	0.563
+ Label-Text + PT (des augmented)	0.270	0.598	0.716	0.566
+ Label-Text + PT(not augmented)	0.233	0.587	0.706	0.543
Linear Weighted (PLM-CA + PLM-ICD)	0.283	0.609	0.715	0.570
SARL-ICD	0.286	0.611	0.723	0.573

that label-text interaction, JMoCo pre-training, and adaptive fusion each contribute to the performance.

Adding the label-text interaction module improves macro-F1 from 26.1% to 27.0% for PLM-CA, and from 21.1% to 26.3% for PLM-ICD, with P@8 increasing from 69.9% to 71.5% for the latter. Notably, using LLM-augmented descriptions consistently outperforms the version with non-augmented raw descriptions. To further assess the quality and reliability of the generated descriptions, we conduct an automatic LLM-based evaluation in Appendix F. The results indicate that the generated descriptions generally preserve the original ICD semantics while providing additional information that may facilitate label disambiguation. Furthermore, the results in Appendix H show that the interaction module alone achieves substantially lower F1 performance, suggesting that it is more effective as a complementary semantic enhancement to the backbone representation.

Loading JMoCo pretrained weights also yields consistent improvements. PLM-CA with pretrained weights reaches 27.2% macro-F1, compared with 26.1% for the baseline, while PLM-ICD improves from 21.1% to 26.0%. These results demonstrate the effectiveness of JMoCo pre-

training in learning meaningful representations.

Finally, the complete SARL-ICD achieves the best performance, reaching 28.6% macro-F1 and 61.1% micro-F1. Compared with linear weighted fusion, the MoE-based adaptive fusion further improves macro-F1 and micro-F1, with additional gains in P@8 and P@15. This consistent improvement indicates that adaptive fusion can effectively leverage the complementary strengths of PLM-CA and PLM-ICD.

4.5.2 JMoCo Component Ablation

To further isolate the contributions of JMoCo, we conduct two complementary ablation analyses on MIMIC-III-full. First, we independently remove ICD hierarchy expansion, similarity weighting, and narrative truncation while keeping all other pre-training and fine-tuning settings unchanged. Second, we examine whether the choice of input sections itself affects pre-training quality by comparing full discharge summaries with narrative-only, hospital-course-only, and discharge-diagnosis-only inputs.

Table 4: Ablation study of JMoCo components and pre-training section selection on MIMIC-III-full.

Pre-training Strategy	PLM-CA		PLM-ICD	
	F1 Micro	F1 Macro	F1 Micro	F1 Macro
Full JMoCo	0.606	0.122	0.601	0.118
w/o Hierarchy Expansion	0.602	0.118	0.600	0.112
w/o Similarity Weighting	0.602	0.120	0.601	0.109
w/o Narrative Truncation	0.603	0.120	0.600	0.112
<i>Section Selection During Pre-training</i>				
Full discharge summary	0.603	0.120	0.600	0.112
Narrative sections only	0.606	0.122	0.601	0.118
Hospital-course only	0.603	0.117	0.596	0.103
Discharge diagnosis only	0.602	0.120	0.598	0.110

As shown in Table 4, removing any individual JMoCo component leads to performance degradation on at least one backbone. Removing hierarchy expansion consistently reduces Macro-F1 for both PLM-CA and PLM-ICD, while removing similarity weighting causes the largest Macro-F1 drop for PLM-ICD. Removing narrative truncation also degrades performance, supporting its role in suppressing less informative structural signals during contrastive pre-training. Although the individual improvements are modest, the consistent trends across both backbones indicate that the three components provide complementary benefits.

The lower part of Table 4 further shows that the choice of pre-training sections has a clear impact on downstream performance. Narrative-only pre-training achieves the best results on both PLM-CA and PLM-ICD, matching the performance of full

JMoCo. In contrast, using only hospital-course sections substantially degrades Macro-F1, particularly for PLM-ICD (10.3%), while discharge-diagnosis-only pre-training also underperforms the narrative-only setting. These results suggest that simply increasing the amount of clinical text is not sufficient; restricting pre-training to clinically informative narrative sections helps the model learn more discriminative representations and supports the narrative-focused design of JMoCo.

5 Conclusion

In this paper, we propose SARL-ICD, a structure-aware representation learning framework for automatic ICD coding. Our approach combines three complementary components: (1) JMoCo, a narrative-aware contrastive pre-training strategy that constructs positive pairs using augmented ICD-code similarity while restricting learning to clinically informative narrative sections; (2) a lightweight label-text interaction module that injects semantic knowledge from LLM-augmented ICD descriptions into downstream ICD coding models; and (3) an adaptive fusion mechanism that leverages the complementary strengths of PLM-CA and PLM-ICD..

Experiments on MIMIC-III-full and MIMIC-IV benchmarks demonstrate that SARL-ICD consistently outperforms previous methods across multiple evaluation metrics. In addition, our analyses provide empirical evidence supporting the design of JMoCo. We show that highly templated sections in discharge summaries can introduce shortcut signals during contrastive pre-training. In contrast, restricting pre-training to clinically informative narrative sections consistently improves downstream ICD coding performance. Furthermore, our complementary analysis reveals that PLM-CA and PLM-ICD exhibit distinct specialization patterns, which provides empirical justification for our fusion module design and confirms the generalizability of the observed complementarity.

Overall, our findings highlight the importance of structure-aware contrastive representation learning and semantic label interaction for extreme multi-label medical coding tasks, and suggest that mitigating structural shortcut learning is an important direction for future research.

Limitations

Despite the strong performance of SARL-ICD, several limitations remain to be addressed in future work. First, our Jaccard-guided pre-training relies on the quality and completeness of the ICD hierarchy. While ICD-9 and ICD-10 hierarchies are well-established, certain codes may have ambiguous parent-child relationships. Future work could incorporate alternative medical ontologies. Second, the label-text interaction module requires LLM-generated augmented descriptions, and the quality of these descriptions may vary. This approach may not generalize well to code sets without existing descriptions or to low-resource language settings. Third, our current experiments do not provide a dedicated analysis of rare ICD codes. Although MIMIC-III and MIMIC-IV contain relatively few low-frequency codes, the limited number of rare-code instances makes it difficult to draw reliable conclusions about the model's performance on these labels. Future work could evaluate rare-code performance using larger datasets or dedicated long-tail evaluation protocols.

Finally, our experiments were conducted on MIMIC-III and MIMIC-IV, which are derived from a single institution. Future work could evaluate SARL-ICD on multi-institutional datasets to assess its generalizability across different clinical settings.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (Grant Nos 62272105 and U24A20257).

References

- Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. [Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2572–2582, New York, NY, USA. Association for Computing Machinery.
- Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. [An unsupervised approach to achieve supervised-level explainability in healthcare records](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA. Association for Computational Linguistics.
- Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. [Physiobank, physiotoolkit, and physionet](#). *Circulation*, 101(23):e215–e220.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.
- Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. [PLM-ICD: Automatic ICD coding with pre-trained language models](#). In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.
- Alistair Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei Lehman, Leo Celi, and Roger G. Mark. 2023. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific Data*, 10:1.
- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3(1):160035.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. [Supervised contrastive learning](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Amit Kumar, Souparna Das, and Suman Roy. 2023. [Transfer learning improves unsupervised assignment of icd codes with clinical notes](#). In *2023 IEEE International Conference on Digital Health (ICDH)*, pages 278–287.
- Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. [BERT might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1631–1639, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. [Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art](#). In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.
- Fei Li and Hong Yu. 2020. [ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional](#)

- Neural Network**. In *AAAI Conference on Artificial Intelligence*, pages 8180–8187.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. **Self-alignment pretraining for biomedical entity representations**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.
- Chang Lu, Chandan K. Reddy, Ping Wang, and Yue Ning. 2023. Towards semi-structured automatic icd coding via tree-based contrastive learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- James Mullenbach, Sarah Wiegrefe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. **Explainable prediction of medical codes from clinical text**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.
- Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. **A label attention model for icd coding from clinical text**. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Yifan Wu, Weiyan Qiu, Min Zeng, Xi Chen, Min Li, and Hongtao Zhu. 2025a. **Gom-icd: Automatic icd coding with gap schemes and mixture of experts**. *Big Data Mining and Analytics*, 8(6):1211–1224.
- Yuzhou Wu, Jin Zhang, Xuechen Chen, Xin Yao, and Zhigang Chen. 2025b. **Contrastive learning with large language models for medical code prediction**. *Expert Systems with Applications*, 277:127241.
- Pengtao Xie and Eric Xing. 2018. **A neural architecture for automated ICD coding**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, Melbourne, Australia. Association for Computational Linguistics.
- Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. **Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1767–1781, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. 2021. **BERTMeSH: deep contextual representation learning for large-scale high-performance mesh indexing with full text**. *Bioinformatics*, 37(5):684–692.
- Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. **Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.
- Pingyue Zhang and Mengyue Wu. 2024. Multi-label supervised contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16786–16793.
- Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. **Automatic ICD coding via interactive shared representation networks with self-distillation mechanism**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957, Online. Association for Computational Linguistics.

Table 5: Example of clinical document section parsing

Section	Extracted Content (Excerpt)
Narrative History	admission date , discharge date , date of birth , sex , service , allergies , attending , chief complaint (occipital bleed, fever, septicemia), major surgical or invasive procedure (extra ventricular drain placement and removal), history of present illness (endocarditis, aortic valve replacement with repair of mitral valve and repair of several aortic root abscesses, discharged to rehab, post-op complications including garbled speech, facial droop, left-sided weakness, occipital hemorrhage with intraventricular extension, torsades de pointes requiring defibrillation, ventricular shunt placement), past medical history (endocarditis, aortic valve replacement with mechanical valve, mitral valve repair, history of etoh abuse, prolonged QT, torsades de pointes), social history (lives with parents, cats, dog, fish, works handling shipping products, drinks heavily, smokes 1ppd, denies IV drug use), family history (mother with diabetes and hypertension, father with hypertension), physical exam (vitals, gen, heent, cv, pul, abd, ext, neuro findings including anisocoria, left-sided weakness, hyperreflexia, confabulation), pertinent results (labs: CBC, PT/INR, blood cultures positive for coagulase-negative staphylococcus; imaging: CXR, CT head, MRI brain, TEE showing no vegetation on mitral valve, mechanical aortic valve leaflets not fully seen, echolucent areas consistent with post-operative changes), medications on admission (docusate, aspirin, metoprolol, ceftriaxone, warfarin, tramadol)
Hospital Course	brief hospital course (year old gentleman with pmh significant for endocarditis, occipital hemorrhage, torsades de pointes, VP shunt placement and removal, neurological monitoring with q8h exams, anticoagulation management with heparin bridging to coumadin, treatment of coagulase-negative staph line infection with vancomycin, completion of ceftriaxone course for endocarditis, physical therapy and rehabilitation)
Discharge Information	discharge instructions , discharge plan (coumadin 7mg, heparin drip, goal INR 2.5-3.5), discharge medications (levetiracetam, warfarin, citalopram, thiamine, multivitamin, nicotine patch, acetaminophen, heparin flush), discharge disposition (extended care facility), discharge diagnosis (fever, parieto-occipital hemorrhagic stroke, prosthetic AV heart valve, endocarditis, prolonged QT, torsades de pointes, memory deficit), discharge condition (mental status confused at baseline, ambulatory requires assistance, alert and interactive), followup instructions (cardiac surgery, radiology, neurosurgery appointments)

A Pre-training Truncation vs. Fine-tuning Full Documents

A.1 Section Parsing Strategy

Our pre-training phase truncates clinical documents to the narrative sections, while SARL-ICD fine-tuning uses full-length documents without truncation. To ensure a scalable, reproducible, and fully automated preprocessing pipeline without any manual curation, the extraction of these narrative sections is performed via an algorithmic rule-based keyword matching strategy.

Based on the findings of (Lu et al., 2023) and as outlined in Table 6, we categorize clinical document sections into three major thematic blocks based on standard medical note headers: **Narrative History**, **Hospital Course** and **Discharge Information**

During the JMoCo pre-training phase, our document parser automatically scans the raw clinical notes to match section headers using the predefined keywords for *Narrative History* (e.g., “chief complaint”, “history of present illness”, “past medical history”). The text content falling under these automatically matched headers is then extracted, concatenated, and truncated to the same length. This automated string-matching mechanism operates seamlessly across large-scale EHR datasets, ensuring high efficiency and minimizing the risk of parsing failures due to variations in formatting.

A.2 Case Study: Section Parsing Example

To illustrate the effectiveness of our keyword-based section parsing strategy, we present a concrete example from the MIMIC-III dataset. Table 5 shows how a clinical discharge summary is segmented according to the keywords defined in Table 6.

As shown in the example, the document naturally follows the three-part structure described in Lu et al. (2023). The **Narrative History** section is notably the richest and most diverse, containing patient demographics, admission and discharge dates, allergies, attending physician, chief complaint, major procedures, history of present illness, past medical history (encompassing cardiac, neurological, and substance use history), social history (living situation, occupation, alcohol and tobacco use), family history, comprehensive physical examination findings (including detailed neurological assessment), pertinent laboratory and imaging results, and medications on admission. This section provides the most clinically informative and

patient-specific signals for ICD code assignment, as it describes the patient’s condition in narrative form with rich contextual details.

The **Hospital Course** section, typically marked by keywords such as “brief hospital course” or “hospital course”, summarizes the patient’s in-hospital management, including surgical procedures, medication adjustments, consultations with specialists, and responses to treatment. This section tends to follow more templated language patterns across different patients and is more concise compared to the narrative history.

The **Discharge Information** section encompasses discharge diagnosis, discharge medications, discharge disposition, discharge condition, discharge instructions, and followup plans. While this section contains clinically valuable information, it often reiterates findings from earlier sections and follows highly structured formatting with standardized categories.

In our JMoCo pre-training phase, we retain only the **Narrative History** section, as it contains the richest patient-specific clinical information with minimal templated language. The **Hospital Course** and **Discharge Information** sections are discarded during pre-training to prevent the model from learning spurious correlations based on document structure rather than clinical content. During SARL-ICD fine-tuning, however, we use the complete documents to allow the model to leverage all available information for downstream ICD coding.

A.3 Rationale for Distribution Shift

This deliberate distribution shift between pre-training and fine-tuning serves two primary purposes:

First, pre-training on automated, truncated, and information-dense narrative sections allows the model to learn patient-specific clinical representations. By filtering out repetitive, semi-structured templated language patterns (e.g., hospital course, standard discharge instructions) at this stage, we effectively prevent the model from falling into shortcut learning.

Second, fine-tuning on full documents enables the model to leverage additional contextual information from later sections (e.g., discharge diagnoses, follow-up plans) that may reinforce or complement the semantic cues extracted from earlier narrative sections.

This two-stage paradigm is consistent with stan-

dard practice in representation learning, where pre-training and fine-tuning distributions often differ. As shown in our ablation study (Table 3), models pre-trained with this automated truncation strategy consistently outperform those trained from scratch, demonstrating that the benefits of pre-training transfer effectively despite the distribution shift.

Table 6: Document section classification based on section headers

Section	Keywords
Narrative History	chief complaint, history of present illness, past medical history, social history, family history, physical exam, pertinent results, medications on admission, admission date, date of birth, sex, service
Hospital Course	hospital course, brief hospital course, surgical course, major surgical or invasive procedure
Discharge Information	discharge medications, discharge medication, discharge disposition, discharge condition, discharge instructions, followup instructions, attending, discharge plan

B Additional Results of SARL-ICD on MIMIC-IV ICD-9

We present additional experimental results on the MIMIC-IV ICD-9 dataset. Table 7 shows the performance comparison. Consistent with the findings on MIMIC-III-full and MIMIC-IV ICD-10, our proposed SARL-ICD and its variants achieve superior performance across all metrics, further demonstrating the effectiveness and generalizability of our approach.

C Additional Shortcut-Learning Analysis

The main paper provides representation-level evidence that narrative-focused pre-training improves the alignment between document representations and ICD-label semantics. To further examine whether this improvement is related to reduced reliance on structural shortcuts, we conduct two additional controlled analyses comparing three model variants: (1) the original RoBERTa-PM encoder without JMoCo pre-training, (2) JMoCo pre-trained using narrative sections only, and (3) JMoCo pre-trained using complete discharge summaries.

We first evaluate whether the learned document representations encode clinically meaningful ICD-label semantics without any task-specific ICD coding fine-tuning. Specifically, we perform zero-shot

ICD label retrieval by comparing a document embedding with the embeddings of all ICD label descriptions. For each document, ICD labels are ranked according to the cosine similarity between the document representation and the corresponding label-description representation. We then measure Precision@5 (P@5) by checking whether the ground-truth ICD codes appear among the top five retrieved labels. We report results separately for rare, frequent, and all labels to examine whether the learned representations capture label semantics across different levels of code frequency.

As shown in Table 8, narrative-only JMoCo substantially improves zero-shot alignment between document and ICD-label representations, particularly for frequent codes. In contrast, full-text pre-training provides only marginal improvements over the original encoder.

Importantly, this comparison provides a control against the alternative explanation that narrative sections simply provide more informative supervision. The full-text model has access to the same narrative sections as well as additional information, yet performs substantially worse in zero-shot label retrieval. This result is consistent with the hypothesis that including highly structured sections can introduce competing shortcut signals during representation learning.

D Complementary Analysis of Base Models

D.1 Analysis on MIMIC-III-full

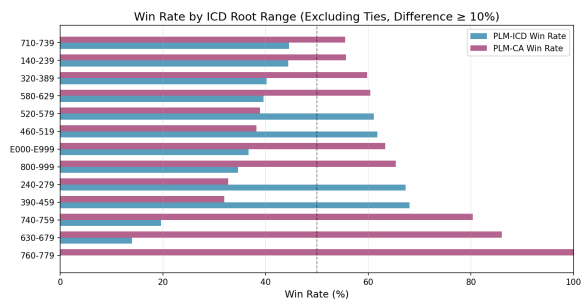


Figure 4: Win rate comparison between PLM-ICD (Huang et al., 2022) and PLM-CA (Edin et al., 2024) across ICD-9 root categories on MIMIC-III-full. Root categories with win rate difference $\geq 10\%$ are shown.

To demonstrate the complementary strengths of PLM-ICD (Huang et al., 2022) and PLM-CA (Edin et al., 2024), we analyze their per-label performance on the MIMIC-III-full dataset. For each label, we compute the F1 score:

Table 7: Performance comparison on MIMIC-IV ICD-9 dataset.

Models	MIMIC-IV ICD-9					
	AUC		F1		P@K	
	Macro	Micro	Macro	Micro	P@8	P@15
CAML (Mullenbach et al., 2018)	0.907	0.988	0.191	0.584	0.661	0.502
LAAT (Vu et al., 2020)	0.958	0.991	0.293	0.612	0.688	0.527
PLM-ICD (Huang et al., 2022)	0.972	0.994	0.298	0.626	0.700	0.535
PLM-CA (Edin et al., 2024)	0.971	0.994	0.324	0.633	0.707	0.540
SARL-ICD	0.974	0.995	0.356	0.644	0.719	0.550

Table 8: Zero-shot ICD label retrieval using document representations. No ICD coding fine-tuning is performed.

Model	P@5 (Rare)	P@5 (Frequent)	P@5 (All)
Original RoBERTa-PM	0.0000	0.0000	0.0020
JMoCo (Narrative only)	0.0010	0.0700	0.0820
JMoCo (Full discharge summary)	0.0010	0.0020	0.0040

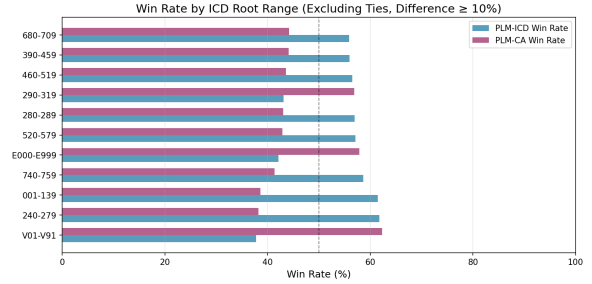
$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN},$$

where TP, FP, FN denote true positives, false positives, and false negatives, respectively. Labels are grouped according to the ICD-9 hierarchical structure. For each root category, we calculate the win rate for each model as:

$$R_m = \frac{w_m}{w_m + w_o} \times 100\%,$$

where w_m and w_o are the numbers of labels where the model wins and the other wins, respectively (ties excluded). A model is considered to win if its F1 exceeds the other's by more than 0.01.

Figure 4 presents the root categories where the win rate difference exceeds 10% (with at least 5 labels per category). The results reveal clear specialization patterns. PLM-CA significantly outperforms PLM-ICD on perinatal conditions (760-779, 100% win rate), pregnancy complications (630-679, 86.0%), and congenital anomalies (740-759, 80.3%). Conversely, PLM-ICD excels on circulatory system diseases (390-459, 68.1%), endocrine disorders (240-279, 67.3%), and respiratory diseases (460-519, 61.8%). Categories such as injury and poisoning (800-999) and external causes (E000-E999) favor PLM-CA, while digestive (520-579) and genitourinary (580-629) diseases slightly favor PLM-ICD.

Figure 5: Win rate comparison between PLM-ICD (Huang et al., 2022) and PLM-CA (Edin et al., 2024) across ICD-9 root categories on the MIMIC-IV ICD-9 dataset. Root categories with win rate difference $\geq 10\%$ and at least 5 labels are shown.

D.2 Analysis on MIMIC-IV ICD-9

Figure 5 presents the root categories on the MIMIC-IV ICD-9 dataset where the win rate difference exceeds 10%. The results are largely consistent with our findings on MIMIC-III-full. PLM-ICD (Huang et al., 2022) continues to demonstrate strong performance on endocrine, nutritional and metabolic diseases (240-279, 61.8% win rate), infectious and parasitic diseases (001-139, 61.4%), as well as circulatory (390-459, 55.9%) and respiratory system diseases (460-519, 56.5%). Conversely, PLM-CA (Edin et al., 2024) shows advantage on supplementary classification codes (V01-V91, 62.3% win rate) and external causes (E000-E999, 57.9%). Notably, while the exact win rate magnitudes differ slightly between datasets, the overall specialization patterns remain coherent: PLM-ICD tends to excel on chronic physiological diseases, whereas PLM-CA shows relative strength on external and supplementary code categories.

D.3 Case Study

To qualitatively illustrate the complementary behavior of the two base models and the effectiveness

of MoE fusion, we present a representative case from the MIMIC-III-full test set. Table 9 compares the predictions of PLM-CA, PLM-ICD, and our SARL-ICD against the ground truth labels.

Table 9: Case study comparison of PLM-CA, PLM-ICD and SARL-ICD predictions. ✓ marks correct predictions.

Ground Truth	PLM-CA	PLM-ICD	SARL-ICD
38.93 Venous catheterization	38.93 (0.840) ✓	38.93 (0.872) ✓	38.93 (0.802) ✓
38.91 Arterial catheterization	38.91 (0.863) ✓	38.91 (0.818) ✓	38.91 (0.759) ✓
401.9 Essential hypertension	401.9 (0.948) ✓	401.9 (0.954) ✓	401.9 (0.788) ✓
276.2 Acidosis	276.2 (0.897) ✓	276.2 (0.918) ✓	276.2 (0.729) ✓
389.9 Unspecified hearing loss	389.9 (0.637) ✓	389.9 (0.774) ✓	389.9 (0.626) ✓
427.31 Atrial fibrillation	427.31 (0.986) ✓	427.31 (0.977) ✓	427.31 (0.922) ✓
584.9 Acute kidney failure	584.9 (0.683) ✓	584.9 (0.894) ✓	584.9 (0.644) ✓
507.0 Pneumonitis	507.0 (0.636) ✓	507.0 (0.461) ✓	507.0 (0.552) ✓
458.9 Hypotension	458.9 (0.285) ✓	458.9 (0.270) ✓	458.9 (0.345) ✓
112.0 Candidiasis of mouth	–	–	112.0 (0.828) ✓
157.9 Pancreatic neoplasm	–	–	157.9 (0.268) ✓
389.8 Hearing loss	–	389.8 (0.648) ✓	–
038.9 Septicemia	038.9 (0.399) ✓	038.9 (0.526) ✓	038.9 (0.358) ✓
157.8 Pancreatic neoplasm	157.8 (0.487) ✓	157.8 (0.331) ✓	157.8 (0.578) ✓
E933.1 Drug adverse effect	E933.1 (0.546) ✓	–	E933.1 (0.468) ✓
Correct Count	12 / 15	12 / 15	14 / 15

Several observations can be made from this case. First, both PLM-CA and PLM-ICD correctly identify nine ground truth labels within their predictions, but their correct sets differ. PLM-CA successfully captures 112.0 (Candidiasis of mouth) while PLM-ICD captures 389.8 (Other specified hearing loss), demonstrating complementary predictive patterns. Notably, SARL-ICD successfully combines the strengths of both models, correctly predicting 112.0 (learned from PLM-CA) while also capturing 157.9 (Pancreatic neoplasm) that neither base model identified. Overall, SARL-ICD correctly predicts 14 out of 15 ground truth labels, substantially outperforming both individual models. This case demonstrates that MoE fusion effectively integrates the complementary predictions of the two base models, achieving superior recall by dynamically selecting the most reliable predictions from each.

D.4 Discussion

These complementary patterns suggest that the two models capture different aspects of medical cod-

ing, which can be attributed to their distinct architectural designs. PLM-CA employs a cross-attention mechanism with learnable label queries, where each query attends to the entire document representation to extract label-specific information. This design is particularly effective for conditions with strong temporal or procedural patterns, such as pregnancy complications and congenital anomalies, where the relevant clinical signals are distributed across specific sections of the note. In contrast, PLM-ICD utilizes a label-wise attention mechanism that computes similarity between token representations and pre-computed label embeddings derived from structured code descriptions. This architecture excels at capturing semantic alignment between chronic disease descriptions and clinical text, making it better suited for conditions like circulatory, endocrine, and respiratory diseases, which have rich semantic descriptions in the label space. Consequently, PLM-CA appears more sensitive to conditions related to pregnancy, infancy, and congenital defects, while PLM-ICD better handles chronic diseases of the circulatory, endocrine, and respiratory systems. The consistent observations across both MIMIC-III-full and MIMIC-IV ICD-9 datasets confirm the generalizability of these findings, further motivating the use of MoE fusion to dynamically combine their respective strengths for improved overall performance.

Table 10: JMoCo pretraining hyperparameters used across all datasets.

Parameter	Value
Backbone encoder	RoBERTa-base-PM-M3-Voc
Batch size	4
Optimizer	AdamW
Learning rate	2e-5
Weight decay	0.01
Gradient clipping	1.0
Projection head layers	2-layer MLP
Projection hidden dimension	768
Projection output dimension	256
Activation function	ReLU
Temperature τ	0.07
Momentum coefficient m	0.999
Queue size K	4096
Epochs	3
Pooling strategy	Mean pooling over chunks
Positive pair weighting	Similarity-weighted InfoNCE loss
Similarity threshold θ (MIMIC-III-full)	0.86
Similarity threshold θ (MIMIC-IV ICD-9)	0.97
Similarity threshold θ (MIMIC-IV ICD-10)	0.82

E computational cost

Table 11 summarizes the training time, inference latency, and GPU memory usage of the major components of SARL-ICD. JMoCo pre-training introduces a one-time offline cost of 43.5 minutes and

Table 11: Computational cost breakdown of SARL-ICD components. All measurements are conducted on a single NVIDIA RTX 4090. Training and inference times are reported on the MIMIC-III-full dataset.

Component	Stage	Training Time	Inference Latency	Training GPU Memory
JMoCo Pretraining	Pre-training	43.5 min	—	4,488 MiB
PLM-ICD (baseline)	Fine-tuning	7.2 h	24.7 ms	6,633 MiB
PLM-CA (baseline)	Fine-tuning	5.5 h	28.2 ms	6,772 MiB
PLM-ICD + JMoCo + Label-Text	Fine-tuning	5.6 h	25.9 ms	6,824 MiB
PLM-CA + JMoCo + Label-Text	Fine-tuning	4.7 h	29.4 ms	6,992 MiB
Full SARL-ICD	Fine-tuning	10 h	56.1 ms	12,918 MiB

requires 4,488 MiB of GPU memory. After pre-training, the additional cost of the proposed components during downstream fine-tuning remains moderate. Compared with the corresponding baselines, PLM-ICD and PLM-CA incur only small increases in inference latency and GPU memory when equipped with JMoCo and label-text interaction. In addition, their fine-tuning time is reduced from 7.2 to 5.6 hours and from 5.5 to 4.7 hours, respectively. The complete SARL-ICD requires higher inference cost (56.1 ms) because it jointly executes PLM-ICD and PLM-CA before adaptive fusion. Overall, the computational overhead is mainly associated with maintaining two backbone models, while JMoCo itself is an offline pre-training cost.

F Analysis of LLM-Augmented ICD Descriptions

To assess the quality and reliability of the LLM-augmented ICD descriptions, we conduct an automatic evaluation using DeepSeek-Chat as an LLM-based judge. We randomly sample 150 ICD codes across the three datasets and evaluate their generated descriptions along three dimensions: clinical accuracy, usefulness for label disambiguation, and faithfulness to the original ICD semantics. Each dimension is scored on a 3-point scale.

Importantly, this evaluation is intended as an automatic quality assessment rather than a human-annotated clinical evaluation. The results are summarized in Table 12.

Table 12: Automatic evaluation of LLM-augmented ICD descriptions using DeepSeek-Chat as an LLM-based judge. Scores range from 1 to 3, with higher scores indicating better quality.

Dataset	Clinical Accuracy	Usefulness	Faithfulness
MIMIC-III-full	2.95	2.89	2.95
MIMIC-IV ICD-9	3.00	2.94	3.00
MIMIC-IV ICD-10	3.00	3.00	3.00
Overall	2.98	2.94	2.98

The generated descriptions receive high scores

across all three dimensions, with overall averages of 2.98 for clinical accuracy, 2.94 for usefulness, and 2.98 for faithfulness. In particular, 98% of the sampled descriptions receive the maximum score for both clinical accuracy and faithfulness, while 94% receive the maximum usefulness score. All sampled descriptions receive a score of at least 2 for clinical accuracy and faithfulness, and the automatic evaluation does not identify hallucinated clinical content in the sampled cases.

These results provide automatic evidence that the LLM-generated descriptions generally preserve the semantics of the original ICD codes while introducing additional information that may help distinguish related labels. However, since the evaluation is LLM-based rather than human-annotated, these results should be interpreted as an automatic quality assessment rather than definitive clinical validation.

G hyperparameter

We adopt a staged optimization strategy, where PLM-CA (Edin et al., 2024) and PLM-ICD (Huang et al., 2022) are first pretrained independently, followed by MoE fusion training using their learned representations. The pretraining hyperparameters are shown in Table 10 and the MoE fusion training hyperparameters are shown in Table 13.

Table 13: Fine-tuning and MoE hyperparameters.

Parameter	MIMIC-III-full	MIMIC-IV ICD-9	MIMIC-IV ICD-10
Max document length	4096	4096	4096
Batch size	8	8	8
Gradient accumulation	4	4	4
Learning rate	5e-5	5e-5	5e-5
Epochs	20	20	20
Warmup ratio	0.1	0.1	0.1
Dropout	0.1	0.1	0.1
Number of experts K	4	6	6
Top-k routing	3	4	4
Expert hidden dimension	128	128	128
Random seed	42	42	42

G.1 Model Version

We use qwen-turbo from Alibaba Cloud’s DashScope API, which is equivalent to Qwen-7B-Chat. The API is accessed via dashscope Python library.

Table 14: Performance comparison of the interaction feature alone (F-only) with PLM-ICD and PLM-CA on the MIMIC-IV ICD-10 dataset.

Models	Macro-AUC	Micro-AUC	Macro-F1	Micro-F1	P@8	P@15
PLM-ICD	0.972	0.994	0.298	0.626	0.700	0.535
PLM-CA	0.971	0.994	0.324	0.633	0.707	0.540
F-only	0.913	0.972	0.089	0.531	0.704	0.552

vide sufficient class-specific discriminative power for precise multi-label ICD assignment. Therefore, the label-text interaction module is better viewed as a complementary semantic enhancement to the backbone representation rather than a standalone predictor.

G.2 Prompt Template

The following prompt template is used for each ICD code:

You are a clinical expert. Provide a detailed English description for the ICD-10 procedure code '{code}'.
Original description: '{original_desc}'.
Expand this into a comprehensive, professional explanation (2-3 sentences) that includes:

- 1) What the procedure involves technically
- 2) Its primary clinical indications or purpose
- 3) Relevant anatomical sites or potential complications.

Keep it factual, concise, and in English only. Do not include disclaimers.

G.3 Generation Hyperparameters

The generation hyperparameters are shown in Table 15.

Table 15: Qwen generation hyperparameters.

Parameter	Value
Model	qwen-turbo (Qwen-7B-Chat)
Temperature	0.3
Top-p	1.0 (default)
Max Tokens	Auto (max new tokens up to 2048)
Seed	42
Result Format	text
Timeout	30 seconds
Max Retries	3 per request

H F only

Table 14 compares the interaction feature alone (F-only) with the PLM-ICD and PLM-CA backbones on the MIMIC-IV ICD-10 dataset. The interaction feature alone achieves reasonably competitive AUC and P@K scores, indicating that it captures useful semantic signals for ICD label ranking. However, its F1 performance is substantially lower than that of both PLM-ICD and PLM-CA, with a particularly pronounced degradation in Macro-F1. This suggests that the interaction feature alone does not pro-