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

Anonymous ACL submission

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 a lightweight MoE 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., 2022a; 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.

In this study, rather than proposing another larger ICD coding architecture, we focus on a more funda-

mental 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** (**S**tructure-**A**ware Contrastive **R**epresentation **L**earning 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 encoders, we introduce a lightweight MoE-based fusion module to adaptively combine predictions from PLM-CA and PLM-ICD. Importantly, the MoE component is not the primary contribution of this work; instead, it serves as a practical mechanism for leveraging the improved representations learned through JMoCo and label-text interaction.

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 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

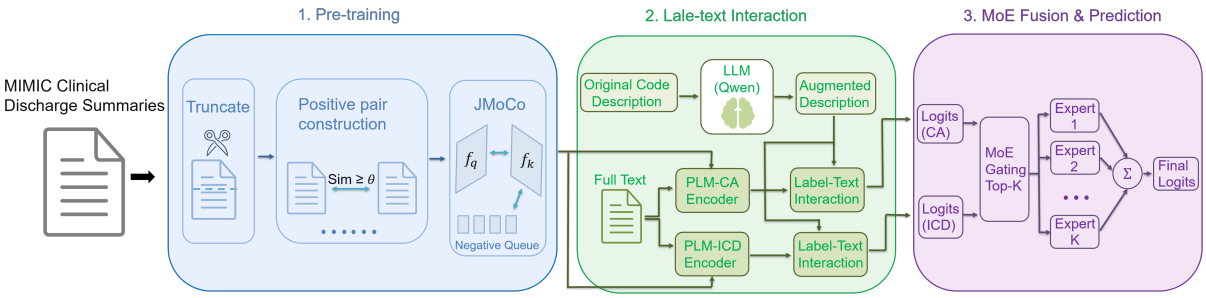

Figure 1: Overview of the proposed SARL-ICD framework. Our approach consists of three key components. (1) Pretraining, where clinically similar discharge summaries are identified and truncated. A MoCo-style contrastive objective is then applied to learn clinically discriminative representations. (2) 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.

neural networks (RNNs)—exemplified by 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. Li and Yu (2020) first 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 (typically 100–200) and does not scale well to extreme multi-label datasets such as MIMIC-III and MIMIC-IV, which contain thousands of ICD codes.

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, contrastive learning has also been leveraged. Yang et al. (2022b) pre-trained a Longformer via hierarchical self-alignment (HSAP) using the UMLS knowledge graph and a hierarchical contrastive loss. 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, thereby aligning similar clinical notes and improving discriminative power.

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". The set of all code descriptions is denoted as $\mathcal{C} = \{c_1, c_2, \ldots, c_{N_c}\}$. 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 contrastive 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 6 and Table 7 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` → `001` → `001-009` → `001-139`).

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

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

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:

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

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

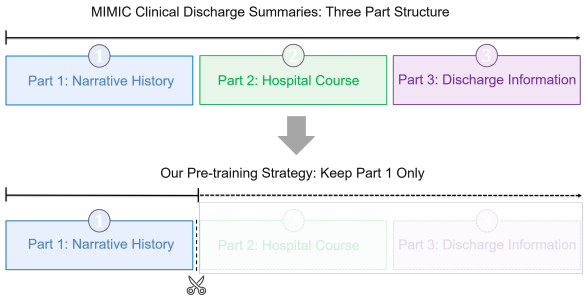

Figure 2: Clinical narrative extraction for JMoCo pre-training. We only keep part1:Narrative History for the pretraining process.

Step 1: Augmented Jaccard Similarity
Raw labels → Expand with ancestors → Compute Jaccard
Select pairs with sim $\geq \theta$
Step 2: Text Truncation
Keep clinical narrative section (chief complaint, HPI, PMH)
Step 3: MoCo Pre-training
Encode truncated texts with f_q and f_k
Contrastive loss weighted by similarity
Update queue and momentum encoder

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

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 7).

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

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

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 $k_j = f_k(\text{Trunc}(x_j))$.

The contrastive loss for a single pair is:

$$\ell_{ij} = -\log \frac{\exp(q_i \cdot k_j/\tau)}{\exp(q_i \cdot k_j/\tau) + \sum_{n \in \mathcal{N}} \exp(q_i \cdot k_n/\tau)} \tag{4}$$

where τ is a temperature hyperparameter and $\mathcal{N}$ denotes the negative sample queue. The final pre-training loss is the similarity-weighted average over all positive pairs:

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

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 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 F. The original description and its augmented version are encoded via a Sentence-BERT model to obtain a fixed-dimensional embedding $\mathbf{e}_i \in \mathbb{R}^H$. 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 $\mathbf{H} \in \mathbb{R}^{L \times H}$ (where L is the sequence length after chunking), we compute a similarity matrix between label embeddings and text tokens:

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

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:

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

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:

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

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.

3.3 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.

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 favor PLM-CA for categories where it demonstrates clear superiority while relying more on PLM-ICD for others, 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 $(\ell_i^{\text{P}}, \ell_i^{\text{B}})$, 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:

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

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:

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

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. 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 E.Our code is available at https://anonymous.4open.science/r/SARL-D5FB.

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 attentionto 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 Design Rationale

We motivate two key design choices in our framework. First, for JMoCo pre-training, we hypothesize that restricting the input to clinically informative narrative sections encourages the model to learn more discriminative patient-specific representations. As validated in Appendix C, this strategy substantially improves the alignment between embedding similarity and ICD-code similarity, and yields better downstream fine-tuning performance compared to pre-training on full documents or other sections. Second, for MoE fusion, we observe complementary specialization patterns between PLM-ICD and PLM-CA: the former excels on chronic diseases with rich semantic descriptions, while the latter performs better on conditions with strong temporal or procedural patterns (see Appendix D). This architectural complementarity motivates the use of MoE to adaptively combine their predictions for improved overall performance.

4.5 Compare with SOTA Methods

To evaluate the effectiveness of our proposed SARL-ICD, we compare it with existing state-of-the-art models on the MIMIC-III-full and MIMIC-IV-ICD10 datasets. The baseline results are adopted from Edin et al. (2023). Table 1 presents the comparative results, averaged over multiple seeds.

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

4.6 Ablation Experiments

To investigate the effectiveness of each proposed component, we conduct ablation experiments on the MIMIC-IV ICD-10 dataset. Table 2 presents the results, demonstrating that all three components contribute to the final performance of SARL-ICD.

Table 1: 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)	0.211	0.585	0.668	0.522	0.911	0.985	0.088	0.539	0.709	0.561	0.895	0.986
DR-CAML (Mullenbach et al., 2018)	–	–	–	–	–	–	0.086	0.529	0.690	0.548	0.897	0.985
LAAT (Vu et al., 2020)	0.203	0.579	0.689	0.543	0.954	0.990	0.099	0.575	0.738	0.591	0.919	0.988
Joint-LAAT (Vu et al., 2020)	–	–	–	–	–	–	0.107	0.575	0.735	0.590	0.921	0.988
MSMN (Yuan et al., 2022)	–	–	–	–	–	–	0.103	0.584	0.752	0.599	0.950	**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

Effectiveness of Label-Text Interaction. Adding the label-text interaction module to the baseline PLM-CA improves macro F1 from 26.1% to 27.0%. For the PLM-ICD baseline, the improvement is even more substantial: macro F1 jumps from 21.1% to 26.3%, with P@8 increasing from 69.9% to 71.5%. Notably, using LLM-augmented descriptions consistently outperforms the version with non-augmented raw descriptions.

Effectiveness of Pretrained Weights. Loading JMoCo pretrained weights brings consistent improvements. PLM-CA (PT) achieves a macro F1 of 27.2%, outperforming the baseline by 1.1%. PLM-ICD (PT) also shows notable gains, reaching 26.0% macro F1 compared to 21.1% of the baseline, demonstrating that contrastive pre-training effectively learns meaningful clinical representations.

Effectiveness of MoE Fusion. The final SARL-ICD model achieves the best overall performance with a macro F1 of 28.6% and micro F1 of 61.1%, when compared to a linear weighted baseline, the MoE structure brings additional improvements of 0.3% in macro F1 and 0.2% in micro F1, with more notable gains in P@8 and P@15. This demonstrates that MoE-based adaptive fusion is more effective than static combination.

Table 2: 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	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
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	0.263	0.596	0.715	0.564
+ PT	0.260	0.596	0.714	0.563
+ Label-Text + PT	0.270	0.598	0.716	0.566
Linear Weighted (PLM-CA + PLM-ICD)	0.283	0.609	0.715	0.570
SARL-ICD	**0.286**	**0.611**	**0.723**	**0.573**

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) a lightweight MoE-based fusion framework that adaptively combines 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, causing models to overfit. 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 MoE fusion design and confirm 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 experiments were conducted on MIMIC-III and MIMIC-IV, which are derived from a single institution. Future work should evaluate SARL-ICD on multi-institutional datasets to assess generalizability across different clinical settings.

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*, pages 2572–2582.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob D. Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. *Preprint*, arXiv:2406.08958.

Ary L Goldberger, Luis AN 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: components of a new research resource for complex physiologic signals. *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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

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 EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei 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. *Advances in neural information processing systems*, 33:18661–18673.

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.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, 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.

Chang Lu, Chandan K. Reddy, Ping Wang, and Yue Ning. 2023. Towards semi-structured automatic icd coding via tree-based contrastive learning. *Preprint*, arXiv:2310.09672.

James Mullenbach, Sarah Wiegreffe, 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-PRICAI-2020, page 3335–3341. International Joint Conferences on Artificial Intelligence Organization.

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. 2022a. 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.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022b. 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.

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.

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 3, 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 4 shows how a clinical discharge summary is segmented according to the keywords defined in Table 3.

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 standard practice in representation learning, where pre-training and fine-tuning distributions often differ. As shown in our ablation study (Table 2), 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 3: 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

Table 4: 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)

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

We present additional experimental results on the MIMIC-IV ICD-9 dataset. Table 5 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 Analysis of JMoCo Pre-training

C.1 Representation Similarity Analysis

To evaluate whether JMoCo learns clinically meaningful representations rather than relying on semi-structured shortcut patterns, we measure the alignment between embedding similarity and augmented ICD-code similarity.

We observe from Table 6 that using similarity weighting alone without narrative truncation leads to poor alignment, suggesting that highly templated discharge sections can dominate the contrastive objective and introduce shortcut signals.

In contrast, truncation substantially improves both Pearson and Spearman correlations, indicating that restricting pre-training to clinically informative narrative sections encourages the model to learn more semantically discriminative representations.

Combining narrative truncation with similarity-aware pair weighting achieves the strongest alignment, demonstrating that the two components are complementary.

The stronger alignment achieved by JMoCo suggests that the learned representations better capture clinical semantic similarity rather than superficial document structure.

C.2 Effect of Section Selection During Pre-training

To investigate whether the performance gains of JMoCo arise from suppressing semi-structured shortcut signals rather than simple denoising, we conduct a controlled ablation study on the MIMIC-III-full dataset using different section-selection strategies during contrastive pre-training.

Specifically, we evaluate both PLM-CA and PLM-ICD under four different pre-training settings: (1) full discharge summaries, (2) narrative sections only (chief complaint, history of present illness, and past medical history), (3) hospital-course and procedure sections only, and (4) discharge diagnosis sections only.

All downstream fine-tuning settings are kept identical, allowing us to isolate the effect of pre-training input structure.

From Table 7, we observe that restricting pre-training to clinically informative narrative sections yields the best downstream performance for both PLM-CA (Edin et al., 2024) and PLM-ICD (Huang et al., 2022). In contrast, pre-training on highly templated hospital-course sections leads to substantially worse results despite containing abundant clinical text.

These findings support our hypothesis that semi-structured sections introduce shortcut patterns that are less useful for representation learning, whereas narrative sections encourage the model to capture more patient-specific clinical semantics.

D Complementary Analysis of Base Models

D.1 Analysis on MIMIC-III-full

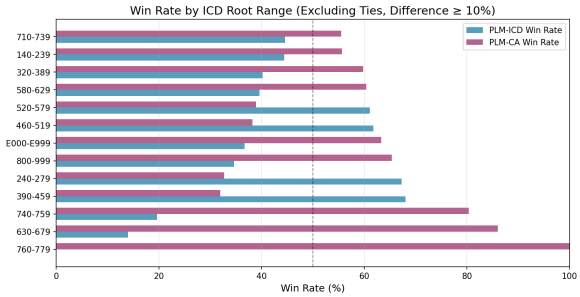

Figure 4: Win rate comparison between PLM-ICD (Huang et al., 2022) and PLM-CA (Edin et al., 2024) across ICD-9 root categories on MIMIC-III-full. Root categories with win rate difference $\geq 10\%$ are shown.

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:

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

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:

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

Table 5: 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	0.193	0.586	0.663	0.503
LAAT (Vu et al., 2020)	0.960	0.993	0.264	0.617	0.689	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 6: 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**

Table 7: Effect of different section-selection strategies during JMoCo pre-training on MIMIC-III-full. Results are evaluated using micro-F1 and macro-F1 on both PLM-CA and PLM-ICD backbones.

Pre-training Input	PLM-CA		PLM-ICD	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Full discharge summary	60.3	12.0	60.0	11.2
Narrative sections only	**60.6**	**12.2**	**60.1**	**11.8**
Hospital-course only	60.3	11.7	59.6	10.3
Discharge diagnosis only	60.2	12.0	59.8	11.0

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.

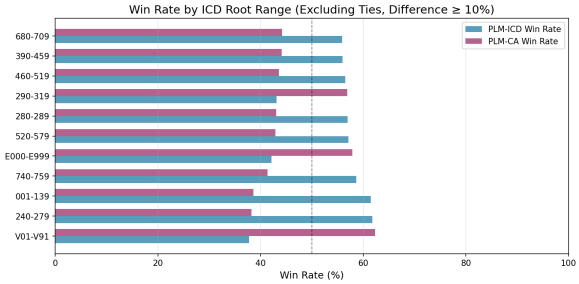

Figure 5: Win rate comparison between PLM-ICD (Huang et al., 2022) and PLM-CA (Edin et al., 2024) across ICD-9 root categories on the MIMIC-IV ICD-9 dataset. Root categories with win rate difference $\geq$ 10% and at least 5 labels are shown.

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 8 compares the predictions of PLM-CA, PLM-ICD, and our LITE-MoE against the ground truth labels.

Table 8: Case study comparison of PLM-CA, PLM-ICD and LITE-MoE predictions. ✓ marks correct predictions.

Ground Truth	**PLM-CA**	**PLM-ICD**	**LITE-MoE**
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, LITE-MoE 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, LITE-MoE 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 coding, 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 9: 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 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 9 and the MoE fusion training hyperparameters are shown in Table 10.

Table 10: 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

Table 11: 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

F LLM Generation Details

We use Qwen-7B-Chat (Qwen-Turbo) via the DashScope API to generate augmented descriptions for ICD codes. The specific model version and generation hyperparameters are detailed below.

F.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.

F.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.
```

F.3 Generation Hyperparameters

The generation hyperparameters are shown in Table 11.