Structure-enhanced Multimodal Medical Concept Quantization for Medication Recommendation

1st Hang Lv
Fuzhou University
Fuzhou, China
lvhangkenn@gmail.com

2nd Xusheng Yu
Fuzhou University
Fuzhou, China
exungsh@foxmail.com

3rd Siying Xu
Fuzhou University
Fuzhou, China
siyingxu604@gmail.com

4th Zixuan Guo
Fuzhou University
Fuzhou, China
832304221@fzu.edu.cn

6th Yanchao Tan†
Fuzhou University
Fuzhou, China
yctan@fzu.edu.cn

5th Xing Chen†
Fuzhou University
Fuzhou, China
chenxing@fzu.edu.cn

7th Carl Yang
Emory University
Atlanta, GA, United States
j.carlyang@emory.edu

***Abstract*—Multimodal medication recommendation leverages diverse medical concepts in Electronic Health Records (EHRs), modeling patient health conditions to support treatment decision-making. However, most existing approaches assign each medical concept an independent continuous embedding, resulting in limited parameter and memory efficiency. Although recent advances in vector quantization are promising for medical concept representation learning, semantic-only quantization often fails to distinguish clinically distinct concepts with similar textual descriptions, leading to code collision. To this end, we propose `SCQMed`, a structure-enhanced multimodal medical concept quantization framework that integrates complementary structural information from medical knowledge graphs. By learning compact and clinically discriminative quantized representations, `SCQMed` enables effective aggregation of multimodal patient longitudinal histories for medication recommendation. Experiments on two real-world EHR datasets show that `SCQMed` significantly improves medication recommendation effectiveness while reducing memory overhead and alleviating code collision.**

***Index Terms*—Multimodal Medication Recommendation, Multimodal Medical Concepts, Residual Vector Quantization**

I. Introduction

Multimodal medication recommendation aims to assist clinicians in making appropriate treatment decisions based on a patient's longitudinal Electronic Health Records (EHRs) [1]–[3]. These EHRs typically consist of various multimodal medical concepts (e.g., diagnoses, procedures, and medications). Effectively modeling multimodal medical concepts across patient visits is essential for understanding patient health states and enabling accurate medication recommendations.

Conventional medication recommendation approaches represent medical concepts using continuous embedding vectors, assigning each concept an independent embedding [4], [5]. This representation design treats medical concepts as isolated entities and fails to explicitly exploit the inherent shared clinical information, which leads to redundant parameterization and memory usage [6], [7]. Recently, Vector Quantization (VQ) has emerged as a powerful technique for representation learning [8], [9]. By mapping continuous features to a fixed-size set of discrete codes, VQ introduces a compact bottleneck that constrains representation capacity and facilitates the modeling of shared semantic patterns. Building upon VQ, Residual-Quantized Variational AutoEncoder (RQ-VAE) [10] employs multi-level codebooks to hierarchically encode residual information, enabling more expressive discrete representations.

Despite their strong potential, directly applying RQ-VAE-based methods to the medication recommendation domain presents notable limitations. These methods primarily rely on semantic similarity derived from textual descriptions to optimize codebooks, which can cause clinically distinct concepts with similar textual descriptions to be assigned to the same discrete codes (i.e., "code collisions") [11], [12]. For example, the diagnoses *Bacterial Pneumonia, Unspecified* and *Viral Pneumonia, Unspecified* are described using closely overlapping terms, yet correspond to fundamentally different etiologies and treatment pathways. Relying solely on textual semantics for quantization fails to capture such fine-grained clinical distinctions, thereby degrading the discriminative power of learned representations and the accuracy of medication recommendations.

To address this challenge, we propose a **S**tructure-enhanced multimodal medical **C**oncept **Q**uantization framework for **Med**ication recommendation (`SCQMed`). Rather than relying purely on a semantic view for quantization, `SCQMed` incorporates a complementary structural view based on medical Knowledge Graphs (KGs) to learn compact yet clinically discriminative quantized medical concept representations. Based on the resulting quantized embeddings, `SCQMed` further aggregates multimodal patient longitudinal histories across visits to support medication recommendations.

Extensive experiments on two EHR datasets demonstrate that `SCQMed` consistently improves the effectiveness of med-

† Xing Chen and Yanchao Tan are the corresponding authors.

This work was supported in part by the National Natural Science Foundation of China under Grants (62302098), Fujian Provincial Natural Science Foundation of China under Grants (2025J01540), and Fujian Provincial Artificial Intelligence Industry Development Technology Project under Grant (2025H0042). Carl Yang was not supported by any funds from China.

ication recommendation (e.g., achieving an average performance gain of 3.48% in Jaccard over the best baseline), while significantly reducing code collision rate compared with semantic-only quantization (e.g., up to 49.09% reduction relative to the standard RQ-VAE). Qualitative case studies further highlight that SCQMed mitigates code collision by integrating structural information from KGs to distinguish clinically distinct medical concepts with similar textual descriptions.

II. Related Work

A. Multimodal Medication Recommendation

Multimodal medication recommendation utilizes heterogeneous medical concepts to provide accurate and safe prescriptions [1], [3], [13]. Based on their patient information modeling strategies, existing methods can be broadly divided into two categories: instance-based and longitudinal. Instance-based methods [14], [15] characterize a patient's clinical status using structured features from a single visit to generate medication recommendations. In contrast, longitudinal approaches [2], [4], [12] explicitly model temporal dependencies across multiple visits to capture long-term disease progression. Within such frameworks, SafeDrug [4] integrated molecular structure embeddings to reduce drug-drug interaction risks and improve recommendation safety. DEPOT [5] decomposed drug molecules into semantic motif-trees to capture collaborative interactions among sub-structures. Despite their effectiveness, most methods represent multimodal medical concepts using massive independent continuous embeddings, resulting in substantial memory overhead and limited parameter efficiency.

B. Vector Quantization-based Methods

VQ has emerged as an effective paradigm for learning compact discrete representations by mapping continuous latent features to a finite set of codes [9], [10]. For example, CF-aug [8] introduced a group-wise VQ to discretize latent features for counterfactual feature generation, alleviating overfitting and bias under data scarcity. To mitigate quantization error and limited representation precision in standard VQ, RQ-VAE is introduced to hierarchically encode residual information using multi-level codebooks [6], [11]. Based on RQ-VAE, MME-SID [7] leveraged multimodal residual quantization to mitigate embedding collapse across textual, visual, and collaborative signals. UDC [12] applied RQ-VAE with condition-aware and task-aware calibration to align textual knowledge with co-occurrence signals for healthcare prediction. However, these methods mainly rely on intrinsic semantic similarity during quantization, leading to code collision between clinically distinct concepts and limiting their effectiveness.

III. Methodology

A. Notation and Problem Definition

Given a patient's longitudinal EHR data consisting of a visit sequence, each associated with a set of multimodal medical concepts (e.g., diagnoses, procedures, and medications), the goal is to recommend an accurate and safe medication combination for the patient's current visit. We provide a detailed overview of the SCQMed framework in Fig. 1.

Formally, a patient's EHRs are represented as a sequence of visits $\mathcal{V} = \{v^1, v^2, \ldots, v^t\}$, where t denotes the total number of visits. Each visit $v^t = (\boldsymbol{d}^t, \boldsymbol{p}^t, \boldsymbol{m}^{t-1})$ contains various medical concepts, where $\boldsymbol{d}^t \in \{0,1\}^{|\mathcal{M}_d|}$, $\boldsymbol{p}^t \in \{0,1\}^{|\mathcal{M}_p|}$, and $\boldsymbol{m}^{t-1} \in \{0,1\}^{|\mathcal{M}_m|}$ are multi-hot vectors indicating the presence of diagnosis, procedure, and historical medication concepts in the visit, respectively. $|\mathcal{M}_d|$, $|\mathcal{M}_p|$, and $|\mathcal{M}_m|$ denote the sizes of the diagnosis, procedure, and medication sets. Since historical medication information in the first visit v^1 is unavailable, $\boldsymbol{m}^0$ is initialized as a zero vector.

For any medical concept $c \in \mathcal{C} = \mathcal{M}_{\{d,p,m\}}$, we associate it with two complementary view information $c = (c^{Sem}, c^{Str})$, where c^{Sem} is the textual description of concept, and $c^{Str} = (\mathcal{E}, \mathcal{R})$ is a medical KG derived from the Unified Medical Language System (UMLS) [16]. $\mathcal{E}$ denotes the set of medical entities in the KG and $\mathcal{R}$ denotes the relations between them.

B. Structure-enhanced Medical Concept Quantization

In Fig. 1, Stage 1 of SCQMed designs a Structure-Enhanced Residual-Quantized Variational AutoEncoder (SE-RQVAE), which integrates semantic and structural quantization to yield compact and discriminative medical concept embeddings.

1) Standard RQ-VAE: For each medical concept c, we utilize a pre-trained BioBERT [17] to encode its textual description c^{Sem} (e.g., *Viral Pneumonia, Unspecified*), generating the semantic embedding $\boldsymbol{c}^{Sem}$. This embedding is then fed into a semantic encoder $\mathcal{E}^{Sem}(\cdot)$ to produce a latent representation $\boldsymbol{z}^{Sem} \in \mathbb{R}^{1\times dim}$. Next, the RQ-VAE discretizes $\boldsymbol{z}^{Sem}$ into a sequence of codes via L-level residual quantization codebooks, where L denotes the code sequence length.

Specifically, at the l-th level, given the previous-level semantic residual $\boldsymbol{r}_{l-1}^{Sem}$, the quantizer selects the closest code index q_l from the l-th codebook $\mathcal{B}_l^{Sem} = \{\boldsymbol{e}_{l,k}^{Sem}\}_{k=1}^{N}$. The residual is then updated by subtracting the selected code embedding $\boldsymbol{e}_{q_l}^{Sem}$ as follows:

$$q_l = \arg\min_k \|\boldsymbol{r}_{l-1}^{Sem} - \boldsymbol{e}_{l,k}^{Sem}\|^2, \boldsymbol{r}_l^{Sem} = \boldsymbol{r}_{l-1}^{Sem} - \boldsymbol{e}_{q_l}^{Sem}, \quad (1)$$

where N denotes the codebook size, $\boldsymbol{e}_{l,k}^{Sem} \in \mathbb{R}^{1\times dim}$ is a learnable code embedding, and the initial residual is set as $\boldsymbol{r}_0^{Sem} = \boldsymbol{z}^{Sem}$. The final quantized semantic representation is obtained by summing all selected code embeddings, i.e., $\boldsymbol{q}_c^{Sem} = \sum_{l=1}^{L} \boldsymbol{e}_{q_l}^{Sem}$. The semantic decoder $\mathcal{D}^{Sem}(\cdot)$ subsequently reconstructs the original semantic embedding $\boldsymbol{c}^{Sem}$ from $\boldsymbol{q}_c^{Sem}$, yielding $\tilde{\boldsymbol{c}}^{Sem}$. The semantic RQ-VAE is trained by minimizing a reconstruction loss together with a standard vector quantization commitment loss as follows:

$$\begin{aligned}
&\mathcal{L}_{RQ-VAE}^{Sem} = \mathcal{L}_{Recon}^{Sem} + \mathcal{L}_{Commit}^{Sem},\\
&\mathcal{L}_{Recon}^{Sem} = \|\boldsymbol{c}^{Sem} - \tilde{\boldsymbol{c}}^{Sem}\|^2,\\
&\mathcal{L}_{Commit}^{Sem} = \sum_{l=1}^{L} \|\mathrm{sg}[\boldsymbol{r}_{l-1}^{Sem}] - \boldsymbol{e}_{q_l}^{Sem}\|^2 + \alpha\|\boldsymbol{r}_{l-1}^{Sem} - \mathrm{sg}[\boldsymbol{e}_{q_l}^{Sem}]\|^2,
\end{aligned} \quad (2)$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient operator, and α controls the balance between optimizing code embeddings and the encoder.

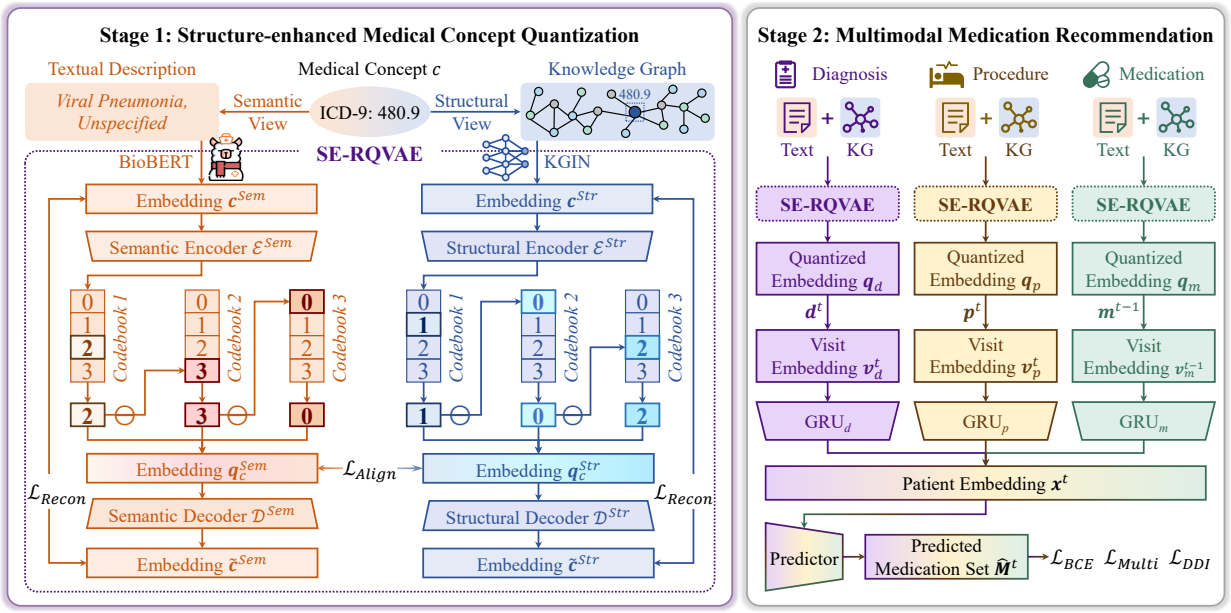


Fig. 1. Overview of our proposed SCQMed framework. In Stage 1, SCQMed performs structure-enhanced medical concept quantization using an SE-RQVAE, where semantic and structural views are discretized separately and then aligned. In Stage 2, the resulting quantized multimodal medical concept embeddings are used for medication recommendation.

2) Structure-enhanced RQ-VAE: As discussed in Section I, semantic-only quantization using a standard RQ-VAE may lead to code collisions among medical concepts. To alleviate this issue, we propose an SE-RQVAE that explicitly incorporates an additional structural view derived from medical KGs into the quantization process. Particularly, we obtain the structural embedding of each medical concept $\boldsymbol{c}^{Str}$ through a KG encoder KGIN [18], which can effectively filter out irrelevant signals and capture informative structural semantics of medical concepts in the UMLS graph $c^{Str} = (\mathcal{E}, \mathcal{R})$. Following a similar quantization procedure as in the standard RQ-VAE, the structural embedding $\boldsymbol{c}^{Str}$ is encoded and mapped to the corresponding quantized structural representation $\boldsymbol{q}_c^{Str}$. The structural RQ-VAE is optimized by minimizing the objective $\mathcal{L}_{RQ\text{-}VAE}^{Str} = \mathcal{L}_{Recon}^{Str} + \mathcal{L}_{Commit}^{Str}$.

To encourage consistency between semantic and structural quantized representations while preserving their complementary information, we introduce a cross-view alignment loss:

$$\mathcal{L}_{Align} = \mathcal{L}_{Sem\to Str} + \mathcal{L}_{Str\to Sem},$$
$$\mathcal{L}_{Sem\to Str} = -\frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\log\frac{\exp(\text{sim}(\boldsymbol{q}_c^{Sem}, \boldsymbol{q}_c^{Str})/\tau)}{\sum_{k\in\mathcal{C}_b}\exp(\text{sim}(\boldsymbol{q}_c^{Sem}, \boldsymbol{q}_k^{Str})/\tau)},$$
$$\mathcal{L}_{Str\to Sem} = -\frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\log\frac{\exp(\text{sim}(\boldsymbol{q}_c^{Str}, \boldsymbol{q}_c^{Sem})/\tau)}{\sum_{k\in\mathcal{C}_b}\exp(\text{sim}(\boldsymbol{q}_c^{Str}, \boldsymbol{q}_k^{Sem})/\tau)}, \tag{3}$$

where $\mathcal{C}_b$ is the medical concept set in a mini-batch, $\text{sim}(\cdot,\cdot)$ is cosine similarity, and τ is a temperature hyperparameter. Finally, the overall training objective of SE-RQVAE is:

$$\mathcal{L}_{SE-RQVAE} = \mathcal{L}_{RQ-VAE}^{Sem} + \mathcal{L}_{RQ-VAE}^{Str} + \beta\mathcal{L}_{Align}, \tag{4}$$

where β balances the contribution of cross-view alignment.

C. Multimodal Medication Recommendation

With compact and discriminative quantized representations of medical concepts learned in Stage 1, Stage 2 enables accurate and efficient multimodal medication recommendation.

1) Quantized Multimodal Medical Concept Representation: We retrieve the quantized embedding of each medical concept c from both semantic and structural views using the trained SE-RQVAE. The final quantized embedding of concept c is constructed by concatenating the two view-specific representations $\boldsymbol{q}_c = [\boldsymbol{q}_c^{Sem} \| \boldsymbol{q}_c^{Str}] \in \mathbb{R}^{1\times 2dim}$. Accordingly, we obtain the quantized embeddings $\boldsymbol{q}_d$, $\boldsymbol{q}_p$, and $\boldsymbol{q}_m$ for diagnosis, procedure, and medication concepts, respectively.

2) Medication Recommendation: Given a patient's current visit $v^t = (\boldsymbol{d}^t, \boldsymbol{p}^t, \boldsymbol{m}^{t-1})$, we first construct visit-level representations by aggregating the quantized multimodal embeddings of diagnoses, procedures, and historical medications:

$$\boldsymbol{v}_d^t = \boldsymbol{d}^t\boldsymbol{q}_d, \boldsymbol{v}_p^t = \boldsymbol{p}^t\boldsymbol{q}_p, \boldsymbol{v}_m^{t-1} = \boldsymbol{m}^{t-1}\boldsymbol{q}_m. \tag{5}$$

To capture temporal dependencies across visits, we adopt three modality-specific GRUs to encode longitudinal patient histories, i.e., $\boldsymbol{h}_d^t = \text{GRU}_d(\{\boldsymbol{v}_d^k\}_{k=1}^t)$, $\boldsymbol{h}_p^t = \text{GRU}_p(\{\boldsymbol{v}_p^k\}_{k=1}^t)$, and $\boldsymbol{h}_m^{t-1} = \text{GRU}_m(\{\boldsymbol{v}_m^k\}_{k=1}^{t-1})$. The obtained modality-aware hidden states are concatenated to form the comprehensive patient representation $\boldsymbol{x}^t = [\boldsymbol{h}_d^t \| \boldsymbol{h}_p^t \| \boldsymbol{h}_m^{t-1}] \in \mathbb{R}^{1\times 3dim}$. Furthermore, medication prediction for the current visit is performed as: $\widehat{\boldsymbol{M}}^t = \text{MLP}(\boldsymbol{x}^t)$, where $\text{MLP} : \mathbb{R}^{1\times 3dim} \to \mathbb{R}^{1\times|\mathcal{M}_m|}$ is a multi-layer perceptron. Following prior studies [2], [4], [5], we train our SCQMed with three loss functions:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{|\mathcal{M}_m|}[\boldsymbol{M}_i^t \log\widehat{\boldsymbol{M}}_i^t + (1-\boldsymbol{M}_i^t)\log(1-\widehat{\boldsymbol{M}}_i^t)],$$
$$\mathcal{L}_{Multi} = \sum_{\{i|\boldsymbol{M}_i^t=1\}}\sum_{\{j|\boldsymbol{M}_j^t=0\}}\frac{\max\{1-(\widehat{\boldsymbol{M}}_i^t - \widehat{\boldsymbol{M}}_j^t), 0\}}{|\mathcal{M}_m|}, \tag{6}$$
$$\mathcal{L}_{DDI} = \sum_{i=1}^{|\mathcal{M}_m|}\sum_{j=1}^{|\mathcal{M}_m|}\boldsymbol{A}_{ij}\cdot\widehat{\boldsymbol{M}}_i^t\left(\widehat{\boldsymbol{M}}_j^t\right)^\top,$$

TABLE I
STATISTICS OF THE DATASETS USED IN OUR EXPERIMENTS.

Dataset	MIMIC-III	MIMIC-IV
# patients / # visits	6350 / 15031	61264 / 163877
# diag. / # proc. / # med.	1903 / 1409 / 131	2000 / 11056 / 131
Avg. / Max # visits	2.3671 / 29	2.6749 / 70
Avg. / Max # diag. per visit	10.2266 / 127	8.2343 / 270
Avg. / Max # proc. per visit	3.8244 / 50	2.3579 / 95
Avg. / Max # med. per visit	11.4361 / 65	6.5055 / 72
DDI Rate	0.0815	0.0793

where $\boldsymbol{M}^t \in \{0,1\}^{|\mathcal{M}_m|}$ denotes the ground-truth medication set for the patient at the t-th visit, and $\boldsymbol{A}$ is a symmetric binary Drug-Drug Interaction (DDI) adjacency matrix. Notably, the multi-label margin loss $\mathcal{L}_{Multi}$ ensures that true labels have at least 1 margin larger than others, leading to more stable predictions. To further balance effectiveness and safety, we employ a dynamic weighting strategy inspired by [2], [5] and the final training objective is formulated as:

$$\mathcal{L}_{Rec} = \gamma \mathcal{L}_{Pred} + (1-\gamma)\mathcal{L}_{DDI}, \tag{7}$$

where the prediction loss $\mathcal{L}_{Pred} = \lambda \mathcal{L}_{BCE} + (1-\lambda)\mathcal{L}_{Multi}$. The weight γ is adaptively adjusted based on the current DDI rate. When the DDI rate exceeds a predefined safety threshold ϵ, $\gamma = \min\{\tanh(\varepsilon \frac{\epsilon}{\text{DDI rate}-\epsilon}), 1\}$. Otherwise, $\gamma = 1$ prioritizes prediction accuracy. λ and ε are tunable hyperparameters.

IV. EXPERIMENTS

In this section, we evaluate our SCQMed on two EHR datasets by addressing the following four core research questions: **RQ1:** How does SCQMed compare with state-of-the-art baselines on medication recommendation? **RQ2:** What are the respective contributions of SCQMed's components to efficiency and effectiveness? **RQ3:** How do the key hyperparameters affect recommendation performance, and how should they be selected? **RQ4:** Does SE-RQVAE mitigate code collisions in semantic-only RQ-VAE for medication recommendation?

A. Experimental Settings

1) Datasets and Evaluation Protocols: To verify the effectiveness of SCQMed, we utilize two real-world EHR datasets: **MIMIC-III** [20] and **MIMIC-IV** [21]. Both datasets are fully anonymized and carefully sanitized before our access. Following [2], [4], [5], we chose patients who made at least two visits for both datasets and the ATC third-level code[1] as the target label. The statistics are summarized in Table I. For evaluation metrics, we use Jaccard Similarity Score (Jaccard), Average F1 Score (F1), Precision Recall AUC (PRAUC), Drug-Drug Interaction Rate (DDI), and Average Number of Medications (# Med.), highlighting how well the model aligns with real-world prescribing patterns, which are consistent with [2], [5].

2) Methods for Comparison: To comprehensively assess our SCQMed, we employ 10 representative state-of-the-art baselines across two categories: (1) instance-based methods: **LR** [15] and **ECC** [14]; (2) longitudinal-based methods: **RETAIN** [22], **LEAP** [13], **GAMENet** [3], **MICRON** [1], **SafeDrug** [4], **MoleRec** [2], **DEPOT** [5], and **UDC** [12].

[1]https://www.who.int/tools/atc-ddd-toolkit/atc-classification

3) Implementation Details: We split training, validation, and test sets by 4:1:1, consistent with [2], [4], [5]. For Stage 1, the semantic and structural embeddings of medical concepts are obtained using BioBERT[2] and KGIN, with embedding dimensions of 768 and 256, respectively. We employ a 3-layer MLP with GELU activations as both the encoder and decoder in SE-RQVAE for the two views. We use $L = 4$ residual quantization levels with a code embedding dimension of $dim = 64$ for both semantic and structural views. The codebook size is set to $N = 256$ for diagnoses and procedures, and $N = 32$ for medications. All codebooks are initialized using K-Means clustering. Following [11], [23], the key loss hyperparameters are set to $\alpha = 1$ and $\beta = 0.01$.

For Stage 2, the SE-RQVAE is frozen and used to generate discrete code indices for medical concepts. These codes are mapped to learnable embedding tables, which are optimized jointly with the downstream medication recommendation model. We employ three modality-specific single-layer GRUs, each with a hidden dimension of 128. The key loss hyperparameters are set to $\lambda = 0.95$, $\varepsilon = 0.08$, and $\epsilon = 0.06$. All experiments are performed with two NVIDIA GTX 3090 Ti GPUs. The full code for this work is available[3].

B. Overall Performance Comparison (RQ1)

We present a comprehensive evaluation of the proposed SCQMed against 10 representative baselines on MIMIC-III and MIMIC-IV. As shown in Table II, SCQMed consistently outperforms all baselines across all accuracy metrics (i.e., Jaccard, F1, and PRAUC), achieving improvement ranging from 2.10% in F1 on MIMIC-III to 3.89% in Jaccard on MIMIC-IV. These results demonstrate that SCQMed effectively learns compact and clinically discriminative medical concept representations with structure-enhanced multimodal quantization, enabling more accurate aggregation of longitudinal patient information for medication recommendation.

Compared with instance-based methods. Instance-based approaches, such as LR and ECC, characterize patient status using information from a single visit and fail to capture temporal dependencies. Compared with these methods, SCQMed achieves improvements of up to 18.23% in Jaccard on MIMIC-III. These results highlight the importance of leveraging longitudinal patient histories and multimodal clinical information for accurate medication recommendations.

Compared with longitudinal methods. Among longitudinal approaches, DEPOT and UDC represent the most competitive baselines in terms of accuracy. Compared with DEPOT, which relies on continuous medical concept embeddings, SCQMed achieves an average accuracy improvement of 3.15% across the two datasets, indicating the advantage of discrete medical concept representations for modeling longitudinal EHR data. Additionally, unlike UDC, which relies solely on textual semantics, SCQMed incorporates structural knowledge into discrete representation learning, leading to more effective and reliable quantization of medical concepts.

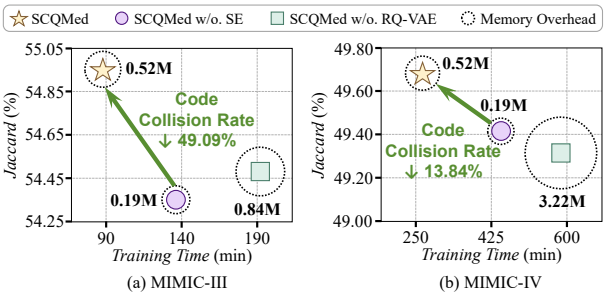
[2]https://huggingface.co/dmis-lab/biobert-base-cased-v1.2

[3]https://github.com/lvhangkenn/SCQMed

TABLE II

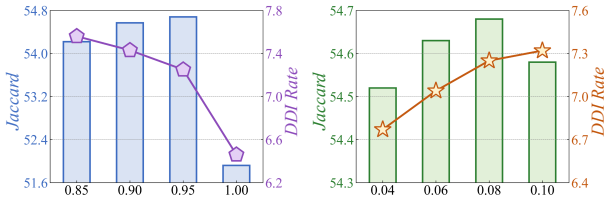
EXPERIMENTAL RESULTS (%) ON TWO EHR DATASETS. THE BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLDFACE**, INDICATING STATISTICALLY SIGNIFICANT IMPROVEMENTS ACCORDING TO THE WILCOXON SIGNED-RANK TEST [19], WHILE THE SECOND-BEST RESULTS ARE UNDERLINED. GROUND-TRUTH # MED. ON THE TEST SETS IS 19.79 FOR MIMIC-III AND 11.98 FOR MIMIC-IV, RESPECTIVELY.

Dataset	**MIMIC-III**					**MIMIC-IV**				
Method	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓	# Med.	Jaccard ↑	F1 ↑	PRAUC ↑	DDI ↓	# Med.
LR	$46.47_{\pm 0.12}$	$62.65_{\pm 0.18}$	$74.72_{\pm 0.13}$	$8.09_{\pm 0.09}$	15.85	$41.01_{\pm 0.13}$	$55.89_{\pm 0.15}$	$67.17_{\pm 0.19}$	$7.76_{\pm 0.07}$	10.03
ECC	$45.56_{\pm 0.16}$	$61.67_{\pm 0.14}$	$71.92_{\pm 0.19}$	$8.18_{\pm 0.12}$	15.48	$39.27_{\pm 0.18}$	$53.74_{\pm 0.13}$	$66.91_{\pm 0.18}$	$7.83_{\pm 0.07}$	7.89
RETAIN	$45.37_{\pm 0.13}$	$61.74_{\pm 0.16}$	$72.19_{\pm 0.12}$	$8.46_{\pm 0.07}$	19.96	$41.64_{\pm 0.12}$	$56.87_{\pm 0.20}$	$67.12_{\pm 0.13}$	$8.09_{\pm 0.09}$	10.53
LEAP	$43.28_{\pm 0.13}$	$59.86_{\pm 0.18}$	$64.65_{\pm 0.13}$	$7.61_{\pm 0.06}$	17.82	$38.98_{\pm 0.17}$	$54.05_{\pm 0.13}$	$54.36_{\pm 0.18}$	$7.28_{\pm 0.07}$	9.78
GAMENet	$47.83_{\pm 0.15}$	$63.79_{\pm 0.17}$	$72.48_{\pm 0.13}$	$8.52_{\pm 0.06}$	24.25	$42.19_{\pm 0.19}$	$57.45_{\pm 0.14}$	$66.29_{\pm 0.18}$	$8.17_{\pm 0.08}$	15.54
MICRON	$48.19_{\pm 0.12}$	$64.03_{\pm 0.18}$	$72.97_{\pm 0.14}$	$7.54_{\pm 0.10}$	18.71	$42.77_{\pm 0.13}$	$57.68_{\pm 0.19}$	$67.33_{\pm 0.13}$	$7.19_{\pm 0.09}$	11.70
SafeDrug	$48.32_{\pm 0.18}$	$64.34_{\pm 0.12}$	$72.61_{\pm 0.19}$	$\mathbf{6.91_{\pm 0.04}}$	18.01	$43.57_{\pm 0.19}$	$58.73_{\pm 0.14}$	$65.43_{\pm 0.19}$	$\mathbf{6.63_{\pm 0.04}}$	10.85
MoleRec	$53.03_{\pm 0.13}$	$68.47_{\pm 0.20}$	$77.75_{\pm 0.14}$	$7.34_{\pm 0.04}$	20.93	$46.74_{\pm 0.13}$	$61.94_{\pm 0.17}$	$70.67_{\pm 0.19}$	$6.94_{\pm 0.02}$	11.85
DEPOT	$53.31_{\pm 0.15}$	$68.68_{\pm 0.13}$	$78.20_{\pm 0.18}$	$7.22_{\pm 0.04}$	20.37	$47.09_{\pm 0.12}$	$62.23_{\pm 0.19}$	$71.11_{\pm 0.11}$	$7.02_{\pm 0.09}$	12.07
UDC	$52.21_{\pm 0.21}$	$67.74_{\pm 0.14}$	$78.24_{\pm 0.11}$	$7.29_{\pm 0.05}$	20.05	$47.81_{\pm 0.16}$	$63.11_{\pm 0.12}$	$71.45_{\pm 0.15}$	$7.12_{\pm 0.07}$	12.15
SCQMed	$\mathbf{54.94_{\pm 0.15}}$	$\mathbf{70.12_{\pm 0.13}}$	$\mathbf{79.29_{\pm 0.16}}$	$7.13_{\pm 0.05}$	20.83	$\mathbf{49.67_{\pm 0.19}}$	$\mathbf{64.64_{\pm 0.12}}$	$\mathbf{73.26_{\pm 0.12}}$	$7.22_{\pm 0.04}$	13.57



(a) MIMIC-III (b) MIMIC-IV

Fig. 2. Ablation studies of SCQMed on MIMIC-III and MIMIC-IV.



(a) The impact of λ (b) The impact of ε

Fig. 3. Hyperparameter studies (%) of SCQMed with λ and ε on MIMIC-III.

Regarding medication safety, compared with the safety-oriented baseline SafeDrug, SCQMed achieves accuracy gains of up to 13.70% in Jaccard on MIMIC-III while incurring only a marginal average increase of 0.0282 in DDI rate. Notably, the DDI rates of SCQMed on both datasets remain lower than those in the ground-truth prescriptions (0.0713 vs. 0.0815 on MIMIC-III and 0.0722 vs. 0.0793 on MIMIC-IV). These results demonstrate a well-controlled trade-off between effectiveness and safety, indicating that SCQMed effectively balances recommendation accuracy and clinical risk.

C. Ablation Studies (RQ2)

Fig. 2 compares SCQMed with two ablation variants: SCQMed w/o. SE removes the structure-enhanced quantization module (i.e., SE-RQVAE) and performs semantic-only residual quantization; SCQMed w/o. RQ-VAE replaces the discrete quantized representations with traditional independently learned continuous embedding tables. Across both datasets, SCQMed achieves the highest Jaccard accuracy while requiring lower memory overhead and faster training. Compared with SCQMed w/o. SE, although introducing additional medical concept memory to incorporate structural views, SCQMed significantly reduces code collision across all medical concepts by an average of 31.47% on both EHR datasets. Here, following [6], [11], the code collision rate is defined as the proportion of medical concepts mapped to non-unique discrete codes. This indicates that integrating structural information enables more distinguishable medical concept representations for accurate medication recommendation. Moreover, SCQMed reduces storage overhead by up to 5.19× on MIMIC-IV and converges faster during training than the w/o. RQ-VAE variant. These efficiency gains become more evident as the number of medical concepts increases, showing the scalability of compact quantized representations in large-scale clinical datasets.

D. Hyperparameter Studies (RQ3)

We analyze the impact of two key hyperparameters in the recommendation loss on MIMIC-III, i.e., λ and ε, which control the trade-off between effectiveness and safety.

Impact of λ. The hyperparameter λ balances the BCE loss and the multi-label margin loss. As shown in Fig. 3(a), an excessively large λ diminishes the contribution of the margin loss, resulting in overly conservative recommendations, whereas a too small value leads to unstable training. Following [2], [4], [5], $\lambda = 0.95$ yields the best and most stable performance.

Impact of ε. The hyperparameter ε controls the nonlinearity of the dynamic weighting between prediction loss and DDI constraint. A smaller ε yields a steeper penalty curve emphasizing medication safety, while a larger value produces a smoother curve with less focus on DDI minimization. Fig. 3(b) shows $\varepsilon = 0.08$ achieves the best accuracy-safety balance.

E. Case Studies (RQ4)

To highlight the advantages of SCQMed in mitigating code collisions and capturing fine-grained clinical distinctions, we provide a case study in Fig. 4. We compare SCQMed with SCQMed w/o. SE, which removes SE-RQVAE and adopts residual quantization based solely on semantic similarity.

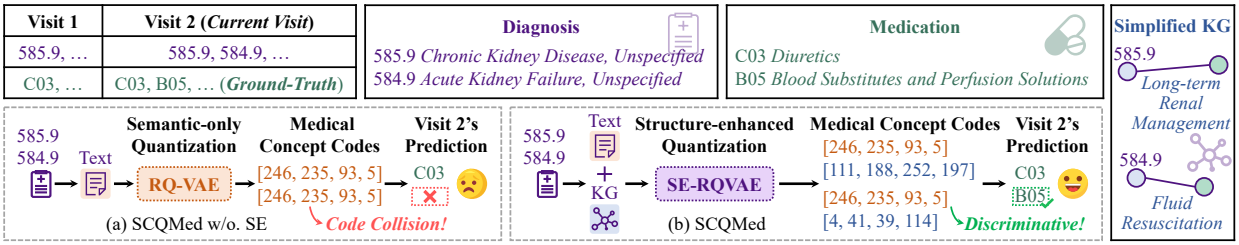


Fig. 4. A case study illustrating how structure-enhanced multimodal medical concept quantization (i.e., the SE-RQVAE in SCQMed) mitigates code collisions inherent in the semantic-only RQ-VAE, thereby improving medication recommendation.

When the patient's diagnosis evolves from *Chronic Kidney Disease, Unspecified* (585.9) to *Acute Kidney Failure, Unspecified* (584.9), this acute deterioration typically necessitates timely treatment adjustment (e.g., fluid management) [24]. However, relying only on textual descriptions, SCQMed w/o. SE assigns identical discrete codes to the two diagnoses, resulting in a code collision that prevents distinguishing the acute event from the chronic condition. Consequently, the clinically appropriate medication B05 is missed in the prediction (shown in Fig. 4(a)). In contrast, Fig. 4(b) illustrates that SCQMed incorporates structural information from the medical KG into quantization. As shown in the simplified KG, 585.9 and 584.9 have distinct neighboring clinical concepts, corresponding to *Long-term Renal Management* and *Fluid Resuscitation*, respectively. By encoding this topological distinction into the structural view, SE-RQVAE assigns different structural-view codes to the two diagnoses, enabling SCQMed to correctly capture the acute progression and recommend C03 and B05.

V. Conclusion

In this paper, we propose SCQMed, a framework for medication recommendation with structure-enhanced multimodal medical concept quantization. By incorporating structural information into residual vector quantization, SCQMed learns compact and clinically discriminative concept representations, effectively alleviating code collision caused by semantic-only quantization. Extensive experiments demonstrate that SCQMed consistently improves medication recommendation performance while significantly reducing memory overhead.

References

[1] Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun, "Change matters: Medication change prediction with recurrent residual networks," in *IJCAI*, 2021, pp. 3728–3734.

[2] Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan, "Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning," in *WWW*, 2023, pp. 4075–4085.

[3] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun, "Gamenet: Graph augmented memory networks for recommending medication combination," in *AAAI*, 2019, pp. 1126–1133.

[4] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun, "Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations," in *IJCAI*, 2021, pp. 3735–3741.

[5] Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li, "Enhancing precision drug recommendations via in-depth exploration of motif relationships," *TKDE*, vol. 36, no. 12, pp. 8164–8178, 2024.

[6] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al., "Recommender systems with generative retrieval," *NeurIPS*, pp. 10299–10315, 2023.

[7] Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao, "Empowering large language model for sequential recommendation via multimodal embeddings and semantic ids," in *CIKM*, 2025, pp. 3209–3219.

[8] Lishi Zuo and Man-Wai Mak, "Vector quantization-based counterfactual augmentation for speech-based depression detection under data scarcity," *JBHI*, vol. 29, no. 10, pp. 7559–7567, 2025.

[9] Yixuan Guan, Jianwei Niu, Tao Ren, and Xuefeng Liu, "Enabling communication-efficient and robust federated learning over packet lossy networks via random interleaved vector quantization," in *ICME*, 2025.

[10] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han, "Autoregressive image generation using residual quantization," in *CVPR*, 2022, pp. 11523–11532.

[11] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua, "Learnable item tokenization for generative recommendation," in *CIKM*, 2024, pp. 2400–2409.

[12] Chuang Zhao, Hui Tang, Jiheng Zhang, and Xiaomeng Li, "Unveiling discrete clues: Superior healthcare predictions for rare diseases," in *WWW*, 2025, pp. 1747–1758.

[13] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun, "Leap: learning to prescribe effective and safe treatment combinations for multimorbidity," in *KDD*, 2017, pp. 1315–1324.

[14] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, pp. 333–359, 2011.

[15] Weiwei Cheng and Eyke Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Mach. Learn.*, vol. 76, pp. 211–225, 2009.

[16] Olivier Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[18] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua, "Learning intents behind interactions with knowledge graph for recommendation," in *WWW*, 2021, pp. 878–887.

[19] Robert F Woolson, "Wilcoxon signed-rank test," *Wiley Encycl. Clin. Trials*, pp. 1–3, 2007.

[20] Alistair EW 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, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[21] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard, "The mimic code repository: enabling reproducibility in critical care research," *JAMIA*, vol. 25, no. 1, pp. 32–39, 2018.

[22] Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *NeurIPS*, 2016, pp. 3512–3520.

[23] Yuhao Wang, Junwei Pan, Xinhang Li, Maolin Wang, Yuan Wang, Yue Liu, Dapeng Liu, Jie Jiang, and Xiangyu Zhao, "Empowering large language model for sequential recommendation via multimodal embeddings and semantic ids," in *CIKM*, 2025.

[24] John A Kellum, Paola Romagnani, Gloria Ashuntantang, Claudio Ronco, Alexander Zarbock, and Hans-Joachim Anders, "Acute kidney injury," *Nat. Rev. Dis. Primers*, vol. 7, no. 1, pp. 52, 2021.