SEMANTIC ID LEARNING FOR SAFE MEDICATION RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medication recommendation from electronic health records requires jointly modeling longitudinal patient histories, unordered intra-visit clinical concepts, and drug-drug interaction (DDI) safety constraints. Existing methods either represent clinical concepts as flat vocabulary indices without fully leveraging pre-trained biomedical semantics, or enforce DDI constraints through post-hoc filtering, thereby decoupling safety from model optimization. We propose SID-MEDREC, a two-stage framework that addresses these limitations through three designs. First, a residual quantized variational autoencoder (RQ-VAE) learns hierarchical discrete semantic IDs from BioBERT embeddings, providing a unified representation for diagnoses, procedures, and medications. Second, a set-aware temporal architecture combines Set Transformers for intra-visit set aggregation with GRU encoders for inter-visit sequence modeling, capturing the structural characteristics of EHR data. Third, an adaptive DDI-aware loss dynamically balances recommendation accuracy and DDI avoidance based on a predefined target DDI rate. Experiments on MIMIC-III and MIMIC-IV demonstrate that SID-MEDREC achieves consistent improvements over strong baselines in recommendation performance while maintaining competitive DDI rates.

1 INTRODUCTION

Medication recommendation from electronic health records (EHR) is a critical clinical decision support task: given a patient's longitudinal medical history—a sequence of visits each comprising heterogeneous clinical concepts (diagnoses, procedures, medications)—the system predicts an appropriate set of medications for the current visit. The task is inherently challenging along three axes. **First**, patient histories are longitudinal and irregularly spaced, requiring models that capture temporal dependencies across visits. **Second**, each individual visit contains a set of clinical concepts (e.g., multiple co-occurring diagnoses) whose interactions matter; the model must reason over unordered intra-visit co-occurrence patterns. **Third**, medication safety imposes a hard drug-drug interaction (DDI) constraint: predicted medication combinations must avoid known adverse interactions, which are prevalent in polypharmacy regimens for patients with multiple chronic conditions.

Prior work has addressed fragments of this problem. Instance-based methods Zhang et al. (2017); Shang et al. (2019) model patient representations via recurrent neural networks or memory-augmented networks but treat clinical concepts as atomic indices from a flat vocabulary, discarding the rich ontological structure encoded in standardized terminologies (ICD-9, ATC). Graph-based methods Choi et al. (2017); Zhu et al. (2024) leverage hierarchical concept graphs or heterogeneous medical entity graphs but remain in continuous embedding spaces where concept identity is diffuse. Molecular graph-based and retrieval-augmented approaches Yang et al. (2021; 2023) have emerged recently to explicitly model drug substructures, but they still rely on trainable embedding lookups that cannot exploit pre-trained biomedical semantic knowledge. Critically, most methods handle the DDI constraint as a post-hoc filtering step—predicting medications first, then removing conflicting pairs—which decouples prediction quality from safety and often degrades both.

In this paper, we propose SID-MEDREC, a two-stage framework for medication recommendation. In Stage 1, we train a Residual Quantized Variational Autoencoder (RQ-VAE) to compress continuous BioBERT embeddings of medical concepts into compact, multi-layer discrete *semantic IDs*—hierarchical integer codes where each successive layer captures progressively finer-grained seman-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tic distinctions. This quantization step confers two advantages: it converts noisy, high-dimensional continuous embeddings into stable, structured representations with controlled granularity, and it enables a uniform discrete interface for all clinical modalities (diagnoses, procedures, medications). In Stage 2, a patient-level recommendation model takes sequences of these semantic IDs and predicts medication sets. The model employs (a) a *Self-Attention Block* (Set Transformer) to encode intra-visit concept sets without imposing spurious ordering; (b) stacked *GRU encoders* to model inter-visit temporal progression; and (c) a *DDI-adaptive loss* that dynamically balances prediction fidelity against DDI avoidance, guided by a target DDI rate. We evaluate SID-MEDREC on the MIMIC-III and MIMIC-IV datasets. Experimental results demonstrate that our method achieves competitive recommendation accuracy in terms of Jaccard, F1, and PRAUC while maintaining a DDI rate comparable to state-of-the-art methods. Furthermore, the adaptive loss provides a flexible mechanism for controlling the trade-off between recommendation accuracy and DDI risk.

Our contributions are threefold:

1. We introduce RQ-VAE-based semantic ID learning for medical concept representation, producing hierarchical discrete codes that capture ontological structure while serving as compact input tokens for downstream temporal modeling.
2. We propose a set-aware temporal architecture that separates intra-visit set aggregation (via Set Transformer) from inter-visit sequence modeling (via GRU), matching the natural structure of EHR data.
3. We design an adaptive DDI-aware training objective that tightly integrates safety constraints into gradient-based optimization, eliminating the need for post-hoc DDI filtering while achieving a configurable trade-off with recommendation accuracy.

2 RELATED WORK

2.1 MEDICATION RECOMMENDATION

Early medication recommendation systems adopted instance-based sequence modeling. LEAP Zhang et al. (2017) formulated the task as multi-instance multi-label learning with an LSTM decoder. GAMENet Shang et al. (2019) introduced a memory bank to store historical medication co-occurrence patterns and a graph-augmented module for DDI encoding. SafeDrug Yang et al. (2021) leveraged molecular substructure graphs of medications and introduced a dual-molecular encoding scheme with a DDI loss term. COGNet Wu et al. (2022) proposed a copy-or-generate mechanism that decides whether to recommend a medication based on its presence in historical visits. More recently, MoleRec Yang et al. (2023) combined molecular graphs with substructure-aware encoding, modelling inter-substructure interactions to inform medication prediction. DrugDoctor Kuang & Xie (2024) addressed the cold-start scenario via visit-level representation learning and iterative training. However, these methods predominantly represent medical concepts via trainable embedding lookup tables, which fail to exploit the rich semantic content of clinical terminologies and cannot generalize to unseen or rare concepts.

2.2 DISCRETE REPRESENTATION LEARNING

Vector Quantized Variational Autoencoders (VQ-VAE) van den Oord et al. (2017) introduced discrete latent variables into generative modeling, demonstrating that discrete codes can capture meaningful structure in images and audio. Residual Quantization (RQ) Zeghidour et al. (2022) extended this by cascading multiple VQ layers on residuals, producing hierarchical codes where earlier layers capture coarse structure and later layers capture fine detail. In NLP and recommender systems, semantic ID-based approaches Tay et al. (2022); Rajput et al. (2023) have used RQ-VAE to assign discrete codes to text passages or items for generative retrieval. These methods demonstrate that discrete representations can serve as compact, compositional indices that outperform random tokenization in both retrieval quality and cold-start generalization. In the biomedical domain, however, RQ-VAE has not been applied to medical concept representation, where the ontology-grounded nature of clinical codes makes hierarchical discretization particularly well-suited.

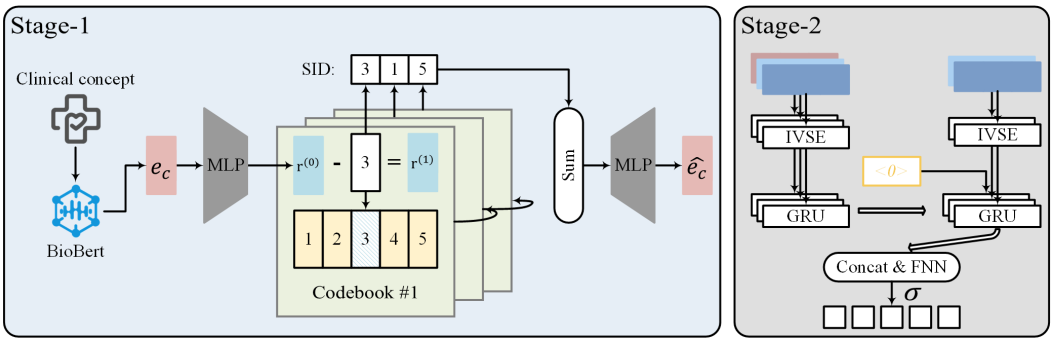


Figure 1: Overview of the two-stage SID-MEDREC framework, consisting of a semantic representation learning stage and a medication recommendation stage.

2.3 SET MODELING FOR CLINICAL DATA

Clinical visits are naturally set-structured: a patient may present with multiple diagnoses and receive multiple procedures and medications, where the ordering of concepts within a visit is generally not meaningful. Standard recurrent architectures, however, may impose an artificial sequential order on these unordered clinical concepts. Set Transformers Lee et al. (2019) address this issue through permutation-invariant attention mechanisms, including MAB, SAB, and ISAB, which enable set representation learning without relying on positional encoding. In healthcare, sequence-of-sets modeling has also been explored for clinical event prediction Zhang et al. (2020), treating patient visits as temporal sequences of unordered event sets and applying permutation-invariant aggregation within each visit. Nevertheless, medication recommendation requires simultaneously capturing the unordered structure of clinical concepts within each visit and the temporal dependencies across visits. This motivates a set-aware temporal architecture that explicitly separates intra-visit set modeling from inter-visit sequential modeling.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

SID-MEDREC operates in two stages (Fig. 1). Stage 1 learns discrete semantic representations for all medical concepts via RQ-VAE. Stage 2 trains a patient-level sequential recommendation model that consumes sequences of semantic IDs and predicts medication sets under an adaptive DDI-aware objective. Let $\mathcal{D}$, $\mathcal{P}$, and $\mathcal{M}$ denote the universal sets of diagnosis codes, procedure codes, and medication codes, respectively. A patient record is a temporally ordered sequence of T visits $\{V_1, V_2, \ldots, V_T\}$, where each visit $V_t = (\mathbf{d}_t, \mathbf{p}_t, \mathbf{m}_t)$ comprises a set of diagnoses $\mathbf{d}_t \subset \mathcal{D}$, a set of procedures $\mathbf{p}_t \subset \mathcal{P}$, and a set of prescribed medications $\mathbf{m}_t \subset \mathcal{M}$. Given a patient's history up to visit t, the task is to predict the medication set $\hat{\mathbf{m}}_t \subset \mathcal{M}$ such that it (1) maximizes agreement with the ground-truth $\mathbf{m}_t$ and (2) minimizes the DDI rate among medications in $\hat{\mathbf{m}}_t$.

Table 1: Statistics of processed MIMIC-III and MIMIC-IV.

Item	MIMIC-III	MIMIC-IV
# of patients	6,350	9,036
# of visits	15,032	20,616
# of diagnoses	1,958	1,892
# of procedures	1,430	4,939
# of medications	131	131
avg. # of visit	2.37	2.28
max. # of visit	29	28

3.2 STAGE 1: SEMANTIC ID LEARNING VIA RQ-VAE

Medical Concept Embedding: We first construct a unified medical ontology $\mathcal{C} = \mathcal{D} \cup \mathcal{P} \cup \mathcal{M}$ covering all clinical concepts. For each concept $c \in \mathcal{C}$, we obtain a text description from its corresponding ontology: ICD descriptions for diagnoses and procedures, and ATC descriptions for medi-

cations. These text strings are encoded via a pre-trained biomedical language model, BioBERT Lee et al. (2020), to produce a 768-dimensional continuous embedding:

$$\mathbf{e}_c = \text{BioBERT}(\text{desc}(c)) \in \mathbb{R}^{768}. \tag{1}$$

RQ-VAE Architecture: The RQ-VAE consists of three components: an encoder, a residual vector quantizer, and a decoder. The encoder is a multi-layer perceptron (MLP) that progressively compresses the 768-d BioBERT embedding into a compact 32-d latent vector:

$$\mathbf{z}_c = f_{\text{enc}}(\mathbf{e}_c) \in \mathbb{R}^{32}, \tag{2}$$

where f_{enc} denotes the encoder network that projects the input features into a compact latent representation, with nonlinear transformations and regularization operations employed to enhance representation learning.

The residual vector quantizer (RQ) Zeghidour et al. (2022) consists of L cascaded vector quantization stages, where each stage employs an independent codebook to progressively encode the residual information. Given the latent $\mathbf{z}_c$, the RQ iteratively quantizes the residual across layers:

$$\begin{aligned} \mathbf{r}^{(0)} &= \mathbf{z}_c, \\ \mathbf{z}_q^{(\ell)}, \text{id}^{(\ell)} &= \text{VQ}^{(\ell)}(\mathbf{r}^{(\ell-1)}), \quad \ell = 1, 2, \ldots, L, \\ \mathbf{r}^{(\ell)} &= \mathbf{r}^{(\ell-1)} - \mathbf{z}_q^{(\ell)}, \end{aligned} \tag{3}$$

where each $\text{VQ}^{(\ell)}$ contains a codebook $\mathbf{E}^{(\ell)} \in \mathbb{R}^{K_\ell \times 32}$ with K_ℓ learnable code vectors. The quantization operation maps the residual to its nearest codebook entry:

$$\text{id}^{(\ell)} = \underset{k \in \{1,\ldots,K_\ell\}}{\arg\min} \left\| \mathbf{r}^{(\ell-1)} - \mathbf{E}_k^{(\ell)} \right\|. \tag{4}$$

The decoder is a mirrored MLP that reconstructs the original 768-d embedding from the quantized latent:

$$\hat{\mathbf{e}}_c = f_{\text{dec}} \left(\sum_{\ell=1}^{L} \mathbf{z}_q^{(\ell)} \right) \in \mathbb{R}^{768}, \tag{5}$$

with f_{dec} is a trainable reconstruction network that projects the quantized latent representation into the original embedding space.

Training Objective: The RQ-VAE is trained by jointly optimizing the reconstruction and quantization objectives:

$$\mathcal{L}_{\text{RQ-VAE}} = \underbrace{\|\hat{\mathbf{e}}_c - \mathbf{e}_c\|}_{\text{reconstruction loss}} + \lambda \frac{1}{L} \sum_{\ell=1}^{L} \Big(\underbrace{\|\mathbf{z}_q^{(\ell)} - \text{sg}[\mathbf{r}^{(\ell-1)}]\|}_{\text{codebook loss}} + \beta \underbrace{\|\text{sg}[\mathbf{z}_q^{(\ell)}] - \mathbf{r}^{(\ell-1)}\|}_{\text{commitment loss}} \Big), \tag{6}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, λ controls the contribution of the quantization objective, and β balances the commitment loss term. The codebook loss updates the codebook embeddings toward the encoder outputs, while the commitment loss encourages the encoder representations to remain close to their assigned codebook vectors and stabilizes the quantization process.

After training, each medical concept c is assigned a L-layer hierarchical semantic ID $\mathbf{s}_c = (\text{id}^{(1)}, \text{id}^{(2)}, \ldots, \text{id}^{(L)})$ by passing its BioBERT embedding through the encoder and RQ with hard assignment.

3.3 STAGE 2: SET-AWARE TEMPORAL MEDICATION RECOMMENDATION

Layered Semantic ID Embedding: Given a concept's semantic ID $\mathbf{s}_c = (id^{(1)}, id^{(2)}, \ldots, id^{(L)})$, we embed each layer independently and aggregate the resulting representations:

$$\begin{aligned} \mathbf{h}_c^{(\ell)} &= \mathbf{W}_{\text{emb}}^{(\ell)}[id^{(\ell)}] \in \mathbb{R}^{d_{\text{emb}}}, \quad \ell = 1, 2, \ldots, L, \\ \mathbf{h}_c &= \begin{cases} \sum_{\ell=1}^{L} \mathbf{h}_c^{(\ell)}, & \text{if Agg. = sum} \\ \mathbf{h}_c^{(1)} \oplus \ldots \oplus \mathbf{h}_c^{(l)} \oplus \ldots \oplus \mathbf{h}_c^{(L)}, & \text{if Agg. = concat} \end{cases} \end{aligned} \tag{7}$$

where $\mathbf{W}_{\text{emb}}^{(\ell)} \in \mathbb{R}^{K\ell \times d_{emb}}$ is a learnable two-dimensional embedding matrix for the ℓ-th quantization layer, and $\oplus$ denotes concatenation. The embedding dimension $d_{\text{emb}} = 64$ controls the representation capacity. Let dim denote the dimension of the resulting Semantic ID representation. Sum aggregation produces a representation of dimension $dim = d_{\text{emb}}$, whereas concatenation preserves the layer-wise representations and yields $dim = L \times d_{\text{emb}}$.

Intra-Visit Set Encoding: For a visit V_t, we collect the semantic ID embeddings of its diagnoses $\{\mathbf{h}_{diag} : diag \in \mathbf{d}_t\}$, procedures $\{\mathbf{h}_{proc} : proc \in \mathbf{p}_t\}$ and medications $\{\mathbf{h}_{med} : med \in \mathbf{m}_t\}$. Each set is independently encoded using a self-attention block inspired by the Set Transformer framework Lee et al. (2019). Specifically, a learnable start-of-sequence token is prepended to each concept embedding sequence:

$$\begin{aligned} \mathbf{H}_t^{\text{diag}} &= \text{SAB}\left(\mathbf{E}_{\text{sos}} \oplus \mathbf{E}_t^{\text{diag}}\right)[0], \\ \mathbf{H}_t^{\text{proc}} &= \text{SAB}\left(\mathbf{E}_{\text{sos}} \oplus \mathbf{E}_t^{\text{proc}}\right)[0], \\ \mathbf{H}_t^{\text{med}} &= \text{SAB}\left(\mathbf{E}_{\text{sos}} \oplus \mathbf{E}_t^{\text{med}}\right)[0], \end{aligned} \tag{8}$$

where $\mathbf{E}_t^{\text{diag}} \in \mathbb{R}^{|\mathbf{d}_t| \times dim}$, $\mathbf{E}_t^{\text{proc}} \in \mathbb{R}^{|\mathbf{p}_t| \times dim}$ and $\mathbf{E}_t^{\text{med}} \in \mathbb{R}^{|\mathbf{m}_t| \times dim}$ are stacked concept embeddings, $\mathbf{E}_{\text{sos}} \in \mathbb{R}^{1 \times dim}$ is a learnable start token prepended before self-attention, and $[0]$ extracts the summary token's output after self-attention.

Inter-Visit Temporal Modeling: The visit representations across time form three parallel sequences:

$$\begin{aligned} \boldsymbol{X}^{\text{diag}} &= [\mathbf{H}_1^{\text{diag}}, \mathbf{H}_2^{\text{diag}}, \ldots, \mathbf{H}_{T-1}^{\text{diag}}, \mathbf{H}_T^{\text{diag}}], \\ \boldsymbol{X}^{\text{porc}} &= [\mathbf{H}_1^{\text{proc}}, \mathbf{H}_2^{\text{proc}}, \ldots, \mathbf{H}_{T-1}^{\text{proc}}, \mathbf{H}_T^{\text{proc}}], \\ \boldsymbol{X}^{\text{med}} &= [\mathbf{H}_1^{\text{med}}, \mathbf{H}_2^{\text{med}}, \ldots, \mathbf{H}_{T-1}^{\text{med}}, \vec{0}]. \end{aligned} \tag{9}$$

Each sequence is processed by a dedicated GRU encoder:

$$\begin{aligned} (\mathbf{O}^{\text{diag}}, \mathbf{h}_T^{\text{diag}}) &= \text{GRU}_1(\boldsymbol{X}^{\text{diag}}), \\ (\mathbf{O}^{\text{proc}}, \mathbf{h}_T^{\text{proc}}) &= \text{GRU}_2(\boldsymbol{X}^{\text{proc}}), \\ (\mathbf{O}^{\text{med}}, \mathbf{h}_T^{\text{med}}) &= \text{GRU}_3(\boldsymbol{X}^{\text{med}}), \end{aligned} \tag{10}$$

where $\mathbf{O}^{(\cdot)} \in \mathbb{R}^{T \times dim}$ are the full output sequences and $\mathbf{h}_T^{(\cdot)} \in \mathbb{R}^{1 \times dim}$ are the final hidden states. The separation of diagnoses, procedures, and medications into independent GRU streams allows each modality to evolve at its own temporal granularity while preventing cross-modal interference during sequence encoding.

Patient Representation and Prediction: The final patient representation concatenates two complementary signals: the *terminal hidden states* (summarizing the full temporal context) and the *last-step outputs* (capturing the most recent clinical state):

$$\mathbf{p} = \left[\mathbf{h}_T^{\text{diag}} \oplus \mathbf{h}_T^{\text{proc}} \oplus \mathbf{h}_T^{\text{med}} \oplus \mathbf{O}_T^{\text{diag}} \oplus \mathbf{O}_T^{\text{proc}} \oplus \mathbf{O}_T^{\text{med}}\right] \in \mathbb{R}^{6dim}, \tag{11}$$

where $\oplus$ denotes concatenation and $\mathbf{O}_T^{(\cdot)}$ is the T-th step output. Medication recommendation scores are computed via a linear projection:

$$\hat{\mathbf{y}} = \mathbf{W}_{\text{out}} \cdot \text{ReLU}(\mathbf{p}) \in \mathbb{R}^{|\mathcal{M}|}, \tag{12}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{|\mathcal{M}| \times 6dim}$. A sigmoid activation produces per-medication probabilities for inference.

Adaptive DDI-Aware Training Objective: The training loss integrates three components with dynamic balancing. Let $\hat{\mathbf{y}}$ be the predicted logits and $\mathbf{y} \in \{0,1\}^{|\mathcal{M}|}$ be the ground-truth multi-hot medication vector.

Binary Cross-Entropy Loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \left[y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i)) \right]. \tag{13}$$

Multi-Label Margin Loss:

$$\mathcal{L}_{\text{Multi}} = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \frac{\max(0, 1 - \sigma(\hat{y}_i) + \sigma(\hat{y}_j))}{|\mathcal{P}| \cdot |\mathcal{N}|}, \tag{14}$$

where $\mathcal{P}$ and $\mathcal{N}$ are the positive and negative medication index sets.

DDI Loss: Let $\mathbf{A} \in \{0,1\}^{|\mathcal{M}| \times |\mathcal{M}|}$ be the DDI adjacency matrix. The DDI loss penalizes co-predicted interacting drug pairs:

$$\mathcal{L}_{\text{DDI}} = \frac{1}{|\mathcal{M}| \cdot |\mathcal{M}|} \cdot \sum_{i=1}^{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} \sigma(\hat{y}_i) \cdot \sigma(\hat{y}_j) \cdot \mathbf{A}_{ij}, \tag{15}$$

Prediction Loss: The BCE and multi-label margin losses are combined with learnable uncertainty weights $w_{\text{BCE}}, w_{\text{Multi}} \in \mathbb{R}$:

$$\mathcal{L}_{\text{Pred}} = e^{-w_{\text{BCE}}} \cdot \mathcal{L}_{\text{BCE}} + e^{-w_{\text{Multi}}} \cdot \mathcal{L}_{\text{Multi}} + (w_{\text{BCE}} + w_{\text{Multi}}). \tag{16}$$

The exponential parameterization and the regularizer $(w_{\text{BCE}} + w_{\text{Multi}})$ follow the multi-task uncertainty weighting framework Kendall et al. (2018), preventing the model from collapsing to a trivial solution where one loss dominates.

Adaptive Trade-Off: The final loss is a convex combination of the prediction loss and the DDI loss, with the mixing coefficient β determined by the current DDI rate:

$$\begin{aligned} \beta &= \sigma\left(-\left(\frac{\text{DDI}_{\text{current}}}{\text{DDI}_{\text{target}}} - 1 \right) \right), \\ \mathcal{L} &= \beta \cdot \mathcal{L}_{\text{Pred}} + (1 - \beta) \cdot \mathcal{L}_{\text{DDI}}, \end{aligned} \tag{17}$$

where $\text{DDI}_{\text{target}}$ denotes the predefined DDI tolerance threshold. The sigmoid gate β dynamically adjusts the trade-off between recommendation accuracy and safety constraints.

4 Experiments

4.1 Datasets and Preprocessing

Datasets. We evaluate SID-MedRec on the MIMIC-III Johnson et al. (2016) and MIMIC-IV Johnson et al. (2023) datasets, two large-scale and de-identified clinical databases released by the Beth Israel Deaconess Medical Center, together covering nearly two decades of critical care data. MIMIC-III contains ICU records collected between 2001 and 2012, while MIMIC-IV extends the temporal

Table 2: Performance comparison on MIMIC-III and MIMIC-IV. Values are reported as MIMIC-III / MIMIC-IV.

Model	Jaccard	DDI	F1	PRAUC	Avg.#Med
LR	0.4924 / 0.4569	0.0830 / 0.0783	0.6490 / 0.6064	0.7548 / 0.6613	16.05 / 8.57
ECC	0.4856 / 0.4327	0.0817 / 0.0764	0.6438 / 0.6129	0.7590 / 0.6530	16.26 / 8.79
LEAP	0.4526 / 0.4254	0.0762 / 0.0688	0.6147 / 0.5794	0.6555 / 0.6659	18.62 / 11.36
DMNC	0.4864 / 0.4673	0.0842 / 0.0711	0.6529 / 0.6322	0.7580 / 0.7389	20.00 / 12.88
GAMENet	0.4994 / 0.4565	0.0890 / 0.0898	0.6560 / 0.6103	0.7656 / 0.6829	27.77 / 18.59
SafeDrug	0.5154 / 0.4487	0.0655 / 0.0604	0.6722 / 0.6014	0.7627 / 0.6948	19.41 / 13.69
4SDrug	0.4837 / 0.4792	0.0693 / 0.0729	0.6456 / 0.6368	0.7625 / 0.7353	16.22 / 20.68
DPG	0.5028 / 0.4795	0.0739 / 0.0701	0.6632 / 0.6407	0.7648 / 0.7383	21.35 / 20.65
DAI-Net	0.5279 / 0.4956	0.0795 / 0.0765	0.6794 / 0.6518	0.7702 / 0.7424	22.52 / 20.87
SID-MedRec	**0.5414 / 0.5129**	**0.0803 / 0.0718**	**0.6949 / 0.6694**	**0.7841 / 0.7649**	**21.51 / 19.19**

coverage to more recent clinical practice, spanning approximately 2008–2022 and including both ICU and emergency department admissions.

We follow the standard preprocessing procedures described in Chen et al. (2023). The datasets are split into training, validation, and test sets with a ratio of 4:1:1. Statistical information of the preprocessed datasets is summarized in Table 1.

Evaluation Metrics. Following previous work Yang et al. (2021; 2023), we use five metrics: Jaccard and F1 for recommendation accuracy, PRAUC for ranking performance, DDI for medication safety, and Avg.#Med for prediction volume.

Baselines. We compare SID-MEDREC with a diverse set of representative baseline models covering different modeling paradigms. These baselines include traditional machine learning methods such as LR Indra et al. (2016) and ECC Read et al. (2011), as well as deep learning–based approaches including LEAP Zhang et al. (2017), DMNC Le et al. (2018), GAMENet Shang et al. (2019), SafeDrug Yang et al. (2021), 4SDrug Tan et al. (2022), DPG Zheng et al. (2023) and DAI-Net Zou et al. (2024).

4.2 PERFORMANCE COMPARISON

Table 2 summarizes the performance of all methods on MIMIC-III and MIMIC-IV. SID-MEDREC achieves the highest recommendation accuracy among the compared methods, with Jaccard/F1/PRAUC scores of 0.5414/0.6949/0.7841 on MIMIC-III and 0.5129/0.6694/0.7649 on MIMIC-IV, corresponding to improvements of 2.56%/2.28%/1.80% and 3.49%/2.70%/3.03% over DAI-Net, respectively. Meanwhile, it maintains competitive DDI rates of 0.0803 and 0.0718, with predicted medication counts of 21.51 and 19.19. These results indicate that SID-MEDREC achieves accurate medication set recovery while maintaining a reasonable prediction volume and safety level.

The results further reveal a clear accuracy–safety trade-off among existing methods. SafeDrug achieves lower DDI rates but lower recommendation accuracy, whereas GAMENet produces larger medication sets accompanied by higher DDI rates. In contrast, SID-MEDREC maintains a balanced prediction range and competitive DDI rates across both datasets, suggesting that its improvements in recommendation accuracy do not rely on excessive medication prediction or relaxed safety constraints.

Overall, SID-MEDREC demonstrates consistent performance across MIMIC-III and MIMIC-IV, indicating robustness under dataset variation. This performance is supported by three complementary components: semantic ID learning provides structured representations of heterogeneous medical concepts, Set Transformer captures unordered intra-visit dependencies, and the adaptive DDI-aware objective regulates the accuracy–safety trade-off. The following ablation study further examines the contribution of each component.

Table 3: Ablation study results on MIMIC-III and MIMIC-IV. Values are reported as MIMIC-III / MIMIC-IV.

Model	Jaccard	DDI	F1	PRAUC	Avg.#Med
w/o SID	0.5395 / 0.5127	0.0768 / 0.0691	0.6921 / 0.6688	0.7818 / 0.7735	20.92 / 18.69
w/o IVSE	0.5306 / 0.5000	0.0799 / 0.0748	0.6850 / 0.6556	0.7789 / 0.7536	22.11 / 19.12
SID Agg. = sum	0.5354 / 0.5052	0.0789 / 0.0700	0.6889 / 0.6605	0.7822 / 0.7590	21.67 / 18.77
w/o DDI Loss	0.5375 / 0.5068	0.0852 / 0.0771	0.6908 / 0.6616	0.7869 / 0.7568	21.65 / 18.54
SID-MedRec	**0.5414 / 0.5129**	**0.0803 / 0.0718**	**0.6949 / 0.6694**	**0.7841 / 0.7649**	**21.51 / 19.19**

K_ℓ	Jaccard	DDI	F1	PRAUC	Avg.#Med
16	0.5330	0.0793	0.6867	0.7799	20.98
32	**0.5414**	**0.0803**	**0.6949**	**0.7841**	**21.51**
64	0.5377	0.0798	0.6910	0.7837	20.58

Table 4: Effect of per-layer codebook size K_ℓ on MIMIC-III. All three RQ-VAE layers use the same codebook size. Bold indicates the best result.

4.3 Ablation Study

Table 3 presents the ablation results of SID-MedRec. We consider four variants: (1) *w/o SID*, which replaces RQ-VAE-based semantic IDs with independently learned embeddings; (2) *w/o IVSE*, which replaces the Self-Attention Block with mean pooling; (3) *SID aggregation: sum*, which replaces layer-wise SID concatenation with summation; and (4) *w/o DDI Loss*, which removes the adaptive DDI objective by setting $\beta = 1$. Among these variants, *w/o IVSE* causes the largest accuracy degradation, with Jaccard decreasing by 0.0108 and 0.0129 on MIMIC-III and MIMIC-IV, respectively. F1 also decreases by 0.0099 and 0.0138. This result highlights the importance of modeling fine-grained interactions among unordered clinical concepts within each visit.

The adaptive DDI objective mainly contributes to safety, while the SID aggregation strategy affects representation quality. Removing the DDI loss increases the DDI rate by 6.1% and 7.4% on MIMIC-III and MIMIC-IV, respectively, while also causing small Jaccard decreases of 0.0039 and 0.0061. This suggests that the adaptive objective effectively regulates unsafe medication combinations during optimization. Meanwhile, replacing SID concatenation with summation reduces Jaccard by 0.0060 and 0.0077, indicating that concatenation better preserves the complementary information encoded at different residual quantization levels.

Removing SID discretization results in only minor Jaccard decreases of 0.0019 and 0.0002 on MIMIC-III and MIMIC-IV, respectively. This indicates that the primary role of SID learning is not immediate supervised accuracy improvement, but the construction of compact and structured representations through shared hierarchical codebooks. Overall, the results reveal complementary contributions from the proposed components: IVSE primarily enhances recommendation accuracy, the adaptive DDI objective improves safety, and hierarchical SID representations provide richer semantic information. The consistent degradation across all variants validates the effectiveness of their combination.

4.4 Hyperparameter Study

Table 4 investigates the effect of the per-layer codebook size K_ℓ. Increasing K_ℓ from 16 to 32 improves Jaccard from 0.5330 to 0.5414 (+1.6%), F1 from 0.6867 to 0.6949 (+1.2%), and PRAUC from 0.7799 to 0.7841 (+0.5%), while maintaining a similar DDI rate (0.0793 vs. 0.0803). However, further increasing K_ℓ to 64 decreases Jaccard to 0.5377 and F1 to 0.6910. This inverted-U trend suggests a granularity trade-off: small codebooks may lack sufficient semantic resolution, whereas overly large codebooks may fragment related concepts. Thus, $K_\ell = 32$ provides an appropriate balance between representation capacity and semantic consolidation.

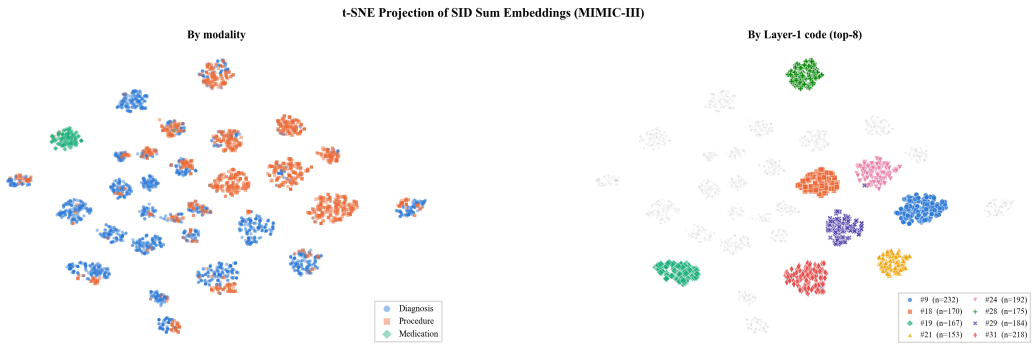


Figure 2: t-SNE visualization of SID representations on MIMIC-III with $K_\ell = 32$. (a) Modality-level visualization. (b) Layer-1 code visualization, achieving NMI = 0.974 and Silhouette = 0.402 for the top-8 frequent codes.

The t-SNE visualization under $K_\ell = 32$ (Fig. 2) further demonstrates the structure of the learned SIDs. Layer-1 assignments closely align with embedding clusters (NMI = 0.974; Silhouette = 0.402), while the top-8 codes exhibit strong modality preferences, achieving an average modality purity of 87.4% (e.g., 99.1% for procedures in Code #9 and 96.2% for diagnoses in Code #29). Medication concepts also form a relatively compact and separated region, with an inter-intra-class distance ratio of 14.9×. These results show that hierarchical semantic structure and modality organization can emerge without explicit modality supervision, while $K_\ell = 32$ represents these patterns using only 96 code vectors, providing a compact discrete vocabulary for medical concepts.

5 Conclusion

This paper presented SID-MedRec, a two-stage framework that integrates hierarchical semantic ID learning, set-aware temporal modeling, and adaptive DDI-aware optimization for medication recommendation. RQ-VAE converts BioBERT embeddings into compact hierarchical semantic IDs, while Set Transformers and GRU capture intra-visit set structures and inter-visit dependencies, respectively. Experiments on MIMIC-III and MIMIC-IV show that SID-MedRec improves Jaccard over DAI-Net by 2.56% and 3.49%, respectively, while maintaining competitive DDI rates. Ablation studies confirm the complementary contributions of set-aware modeling and adaptive DDI optimization, while semantic analysis demonstrates strong codebook organization and modality alignment (NMI = 0.974; modality purity = 87.4%). Limitations include the lack of joint representation–recommendation optimization, globally defined DDI targets, and limited modeling of temporal semantic evolution. Future work will investigate task-adaptive representations, personalized safety calibration, and dynamically evolving clinical semantics. Overall, SID-MedRec provides a unified approach to accurate and safety-conscious medication recommendation.

References

Qianyu Chen, Xin Li, Kunnan Geng, and Mingzhong Wang. Context-aware safe medication recommendations with molecular graph and ddi graph embedding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7053–7060, 2023.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795. Association for Computing Machinery, 2017.

ST Indra, Liza Wikarsa, and Rinaldo Turang. Using logistic regression method to classify tweets into the selected topics. In *2016 international conference on advanced computer science and information systems (icacsis)*, pp. 385–390. IEEE, 2016.

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*, 3(1):160035, 2016.

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

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Yabin Kuang and Minzhu Xie. Drugdoctor: enhancing drug recommendation in cold-start scenario via visit-level representation learning and training. *Briefings in Bioinformatics*, 25(6):bbae464, 11 2024.

Hung Le, Truyen Tran, and Svetha Venkatesh. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1637–1645, 2018.

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*, 36(4):1234–1240, 02 2020.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753. PMLR, 2019.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10299–10315. Curran Associates, Inc., 2023.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85:333–359, 2011.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1126–1133, 2019.

Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S. Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3970–3980. Association for Computing Machinery, 2022.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21831–21843. Curran Associates, Inc., 2022.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. Conditional generation net for medication recommendation. In *Proceedings of the ACM Web Conference 2022*, pp. 935–945. Association for Computing Machinery, 2022.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3735–3741. International Joint Conferences on Artificial Intelligence Organization, 8 2021.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, pp. 4075–4085. Association for Computing Machinery, 2023.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

Tianran Zhang, Muhao Chen, and Alex A. T. Bui. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In *Artificial Intelligence in Medicine*, pp. 348–358. Springer International Publishing, 2020.

Yutao Zhang, Robert Chen, Jie Tang, Walter F. Stewart, and Jimeng Sun. Leap: Learning to prescribe effective and safe treatment combinations for multimorbidity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1315–1324. Association for Computing Machinery, 2017.

Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Xiangyu Zhao, Baoxing Huai, Xian Wu, and Enhong Chen. Interaction-aware drug package recommendation via policy gradient. *ACM Trans. Inf. Syst.*, 41(1), 2023.

Fanglin Zhu, Xu Zhang, Batuo Zhang, Yonghui Xu, and Lizhen Cui. Medicine package recommendation via dual-level interaction aware heterogeneous graph. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2294–2303, 2024.

Xin Zou, Xiao He, Xiao Zheng, Wei Zhang, Jiajia Chen, and Chang Tang. Dai-net: Dual adaptive interaction network for coordinated medication recommendation. *IEEE Journal of Biomedical and Health Informatics*, 28(10):6201–6211, 2024.