# LAMRec: Label-aware Multi-view Drug Recommendation

Yunsen Tang
Shandong University
Jinan, China
tangyunsen@mail.sdu.edu.cn

Ning Liu*
Shandong University
Jinan, China
liun21cs@sdu.edu.cn

Haitao Yuan
Nanyang Technological University
Singapore
haitao.yuan@ntu.edu.sg

Yonghe Yan
Shandong University
Jinan, China
yanyonghe@mail.sdu.edu.cn

Lei Liu
Shandong Research Institute of
Industrial Technology
Jinan, China
l.liu@sdu.edu.cn

Weixing Tan
Shandong University
Jinan, China
dream_8@sina.com

Lizhen Cui
Shandong University
Jinan, China
clz@sdu.edu.cn

## ABSTRACT

The drug recommendation task aims to predict safe and effective drug prescriptions based on the patients' historical electronic health records (EHRs). However, existing drug recommendation models generally have two limitations. First, they neglect the inherent characteristics of multiple views existing in patients' clinical data (e.g., diagnoses and procedures), leading to fragmented and inconsistent patient representations. Second, they do not fully exploit drug label information. Most models do not explicitly establish a mapping relationship between drug labels and patients' historical visits. To address these two problems, we proposed a label-aware multi-view drug recommendation model named LAMRec. In particular, LAMRec uses a cross-attention module to fuse information from the diagnosis and procedure views, and increases the mutual information of patient multi-view representations through multi-view contrastive loss; the label-wise attention mechanism fully explores drug label information by constructing an adaptive mapping of drug-visit to generate personalized representations that are aware of the drug-related visit information. Experiments on three real world medical datasets demonstrated the superiority of LAMRec, with a relative reduction of 5.25% in DDI compared to the optimal baseline, a relative improvement of 4.20% in Jaccard similarity scores, and a relative improvement of 3.10% in F1 scores. We released the code online at: https://github.com/Tyunsen/LAMRec.

## CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → *Data mining*.

## KEYWORDS

Drug recommendation, label-wise attention, contrastive learning

**ACM Reference Format:**
Yunsen Tang, Ning Liu, Haitao Yuan, Yonghe Yan, Lei Liu, Weixing Tan, and Lizhen Cui. 2024. LAMRec: Label-aware Multi-view Drug Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627673.3679656

## 1 INTRODUCTION

In recent years, deep learning has achieved notable success in supporting clinical decision-making processes [5, 17, 20, 34, 39]. Drug recommendation is a significant task among various applications. Typically, this involves learning representations of medical entities (e.g., diagnoses and procedures) from electronic health records, and using these representations to recommend drug combinations. In addition, these recommendations must align with patients' health conditions and minimize drug-drug interactions (DDI).

In particular, this process is implemented through two primary approaches: 1) *instance-based drug recommendation*, which makes predictions based only on the current admission situation without considering the patient's longitudinal history [43]; 2) *longitudinal-based drug recommendation*, which incorporates a simulation of the patient's longitudinal history to enhance prediction accuracy [4, 26, 36–38]. Recent efforts are increasingly focusing on the latter approach, due to its ability to capture long-term health trends and impact on patient diagnoses.

Drug recommendation models fundamentally rely on the encoding of medical entities into comprehensive patient representations, followed by a decoding process to deduce appropriate drug recommendations. However, the development of precise longitudinal drug recommendation models encounters substantial challenges during both the encoding and decoding stages.

[*] Corresponding author.

*CIKM '24, October 21–25, 2024, Boise, ID, USA.*

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679656

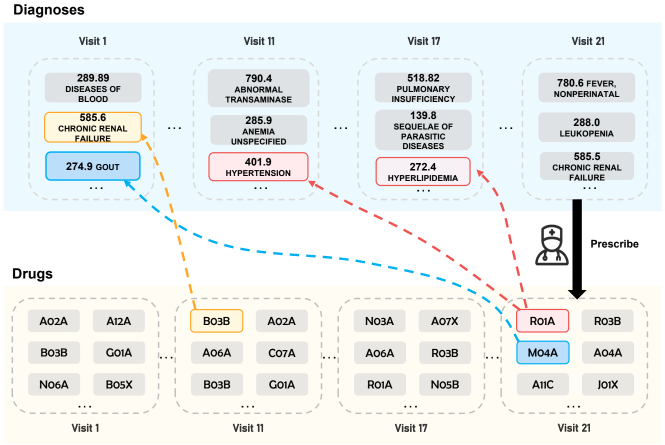

**Figure 1: Example from MIMIC-IV. Diagnosis and physician-prescribed drugs for patient with Subject ID 10516278 at different visit times. According to statistics, the drug "R01A" is typically used to treat "Hyperlipidemia" or "Hypertension", and "M04A" is primarily used to treat "Gout".**

- Most existing drug recommendation models [4, 26, 36, 37, 43] usually treat the patient's diagnosis sequences and procedure sequences as two independent views, and encode them with two recurrent neural networks (RNN). However, this disconnected representation learning approach ignores the importance of dual-view representations. Specifically, the diagnosis and procedure view actually both reflect the health status of the same patient and contain rich complementary information. Encoding the two views separately and directly concatenating the representations cannot fully explore this potential connections between them, resulting in inconsistency of patient representation.
- Existing drug recommendation models have limitations in fully exploiting drug label information. Most models usually focus on the information of the patient's current visit embedding [4, 26, 36–38, 43], ignoring the relevance between drug labels and the patient's historical visits. In the real world, the drug combinations prescribed by doctors are not only based on the patient's recent visit status but may also take into account the diagnoses that appeared in the patient's previous visits. As shown in Figure 1, the doctor prescribed a combination containing drugs "R01A" and "M04A" at the 21-th visit. The use of drug "R01A" is likely to have comprehensively considered the patient's "Hyperlipidemia" at the 17-th visit or "Hypertension" at the 11-th visit, while drug "M04A" may be related to the "Gout" at the 1-st visit. This reflects the long-span and discretely distributed implicit associations between drug labels and historical visits, which are crucial for existing models to capture explicitly.

To address the above issues, we propose LAMRec, a label-aware multi-view drug recommendation model. LAMRec utilizes dual-view representation learning to generate consistent patient representations and fully explores label information through a label-wise attention mechanism to accurately characterize drug-visit associations. Our LAMRec has the following contributions:

- We propose a novel dual-view representation process that fully integrates the potential information from different view sequences through two dual cross-attention networks, and employs a multi-view contrastive learning loss to increase the mutual information between the dual-view representations, thereby obtaining a consistent patient representation.
- We designed a label-wise attention module to construct the relationship between patient historical visits and medications and to generate label-aware vectors that comprehensively consider patient historical information to explore high-value drug combinations. From this perspective, our model fully utilizes patient historical visit representations and fully explores drug label information, resulting in a stronger decoding process.
- We conducted extensive experiments on three real-world EHRs datasets (MIMIC-II, MIMIC-III, and MIMIC-IV). Compared to the state-of-the-art models to our best knowledge, LAMRec achieves a 5.25% relative reduction in DDI rate, a 4.20% relative improvement in Jaccard similarity, and 3.10% in F1 score.

## 2 RELATED WORK

### 2.1 Drug Recommendation

Existing drug recommendation methods can be divided into instance-based and longitudinal-based drug recommendation. Instance-based methods focus on the patient's current health status. For example, LEAP [43] extracts feature information from the current visit record and adopts a multi-instance and multi-label drug recommendation setting. However, most studies [3, 34] suggest adopting a longitudinal approach to utilize the time-dependency in the clinical history. Among them, RETAIN [4] employs a two-layer RNN with attention to model longitudinal information, making the model interpretable. GAMENet [26] further models the sequence based on GRU, adopting a memory neural network and storing historical information as reference information for the next step prediction. SafeDrug [37] incorporates drug molecular structure information on this basis, using a graph-based molecular encoder with MPNN and learnable fingerprints. In addition, MICRON [36] does not use GRU as encoder but focuses on learning drug change patterns between different visits, avoiding the use of RNN for longitudinal sequence modeling, making the patient longitudinal representation process more efficient. Recently, MoleRec [38] modeled substructure interactions and patient-substructure relevance to identify critical substructures for effective treatment, improving the accuracy and safety of drug recommendation. This study aims to leverage the complementary value of dual-view data and information of drug label to significantly improve drug recommendation performance compared to state-of-the-art models.

### 2.2 Contrastive Learning

Contrastive learning has emerged as one of the most effective unsupervised learning paradigms in the field of representation learning, achieving remarkable progress in recent years [2, 9, 15, 21, 23, 27–29]. The fundamental principle of contrastive learning is to maximize the similarity between positive sample pairs while minimizing the similarity between negative sample pairs. Some studies [29] have shown that the success of contrastive learning can be attributed to the maximization of mutual information. More precisely, the widely used InfoNCE loss is a lower bound of mutual information. Therefore, MoCo [10] and CPC [23] can be seen as examples of indirectly maximizing mutual information by minimizing the InfoNCE loss. MoCo employs a momentum-updated encoder and a queue of negative samples to achieve this goal. On the other hand, CPC uses a contrastive predictive coding approach to learn representations that maximize mutual information. DCP [16], proposed for the incomplete view problem, suggests that directly maximizing the mutual information of dual views helps to recover missing views, increases the consistency of representations, and proposes a class-level dual-view contrastive learning method to maximize the mutual information of dual views to obtain more discriminative feature representations.

## 2.3 Attention Mechanism

The attention mechanism was first proposed and applied to machine translation tasks in the field of natural language processing (NLP) by [1], and its core idea is a method of weighting the input, which allows the model to selectively focus on different parts of the input and extract more critical information. Since then, the attention mechanism has been widely applied to various tasks in NLP, such as sentence classification [40], reading comprehension [25], and has derived many variants, such as Hierarchical Attention [40], Co-Attention [18], etc. [30] proposed the Transformer model based entirely on the attention mechanism to replace RNN, which has become the new mainstream model in NLP and further promoted the development of the attention mechanism. In recent years, researchers have further explored the internal mechanisms and applicable scope of attention mechanism, such as interpretability [11], the correlation with linguistic knowledge [6], the introduction of prior knowledge [19], etc. The application of attention mechanism on graphs [12, 41] and multimodal [31] has also become a new research direction. Overall, the attention mechanism enhances the model's ability to handle long-distance dependencies and process key information while improving the model's interpretability, which is an important milestone in the development of NLP technology in the era of deep learning.

# 3 PRELIMINARIES

This section formulates the problem of learning a drug recommendation model that generates safe drug combinations by considering a patient's EHRs data.

## 3.1 Definitions

*3.1.1 Electronic Health Records.* A patient's EHRs data are typically represented as a longitudinal sequence of medical codes (e.g., diagnoses, procedures, and medications) that capture the patient's complete medical history. To avoid confusion, we use a single patient to describe our method. Formally, a patient's EHRs can be represented as a sequence $\mathcal{X} = [v_1, v_2, \ldots, v_t]$, where $t$ is the total number of visits for this patient. The $t$-th item in the sequence $\mathcal{X}$ records the patient's $t$-th visit, represented by $v_t = [d_t, p_t, m_t]$, where $d_t \in \{0,1\}^{|\mathcal{D}|}, p_t \in \{0,1\}^{|\mathcal{P}|}, m_t \in \{0,1\}^{|\mathcal{M}|}$ are the multi-hot vectors of the patient's diagnoses, procedures, and medications at the $t$-th visit, respectively. $\mathcal{D}, \mathcal{P}, \mathcal{M}$ representing the corresponding sets of diagnoses, procedures, and medications.

*3.1.2 Safe Drug Recommendation.* We not only focus on drug recommendation but also control the rate of DDI in the recommended drug combinations. We use a symmetric binary adjacency matrix $A \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{M}|}$ to represent the DDI relationships. When $A[i,j] = 1$, it indicates an interaction relationship between drug $i$ and drug $j$, while $A[i,j] = 0$ indicates that the combination of the two medications is a safe co-prescription.

## 3.2 Problem Formulation

This work aims to learn a drug recommendation model $f(\cdot)$ that, for each patient at their $t$-th visit, given the patient's diagnoses and procedures sequences $[d_1, d_2, \ldots, d_t]$, $[p_1, p_2, \ldots, p_t]$ up to the current visit, and the target drug combination $m_t$, as well as the DDI matrix $A$, $f(\cdot)$ can generate a drug recommendation combination:

$$\widehat{m}_t = f([d_i]_{i=1}^t, [p_i]_{i=1}^t) \in \mathbb{R}^{|\mathcal{M}|} \tag{1}$$

During the training process, the objective function of the model $f(\cdot)$ consists of at least two parts: (i) supervised learning using the real drug combination $m_t$ for $\widehat{m}_t$; (ii) imposing unsupervised DDI constraints on the drug combination $\widehat{m}_t$ using the DDI adjacency matrix $A$ to control the DDI rate.

# 4 METHOD

As shown in Figure 2, LAMRec consists of three key modules: (1) A longitudinal cross-attention module that learns multi-view interactive representations of patients from EHRs data. (2) A multi-view contrastive learning module that maximizes the mutual information between diagnoses and procedures representations. (3) The label-wise attention module learns the attention weights between drug labels and each visit of the patient, and generates label-aware vectors based on the current patient representation.

## 4.1 Cross-Attention Module

The cross-attention module aims to learn compact and efficient dual-view representations from a patient's EHRs data. Like other multi-view learning scenarios [8, 42], both the diagnoses view and procedure view both reflect the health status of the same patient, they can be considered as complementary dual-view data. The cross-attention module fully explores the potential associations between the two views during the training process, allowing the patient representations generated from different views to achieve consistency in the latent space, thereby improving the quality and robustness of the representations.

*4.1.1 Embedding Layer.* To fully utilize the rich information in the time sequence data of each view, we construct two embedding tables, $\mathcal{E}_d \in \mathbb{R}^{|\mathcal{D}|\times dim}$ and $\mathcal{E}_p \in \mathbb{R}^{|\mathcal{P}|\times dim}$, where $dim$ is the dimension of the representation space. Given the multi-hot vetor $d_t \in \{0,1\}^{|\mathcal{D}|}$ of a patient's $t$-th diagnosis, we map it to the embedding space through vector-matrix multiplication to obtain the embedding of the $t$-th diagnosis:

$$d'_t = d_t \mathcal{E}_d \in \mathbb{R}^{dim} \tag{2}$$

To capture the relationships between diagnoses at different visits, we add positional encoding to the diagnosis embedding of the $t$-th

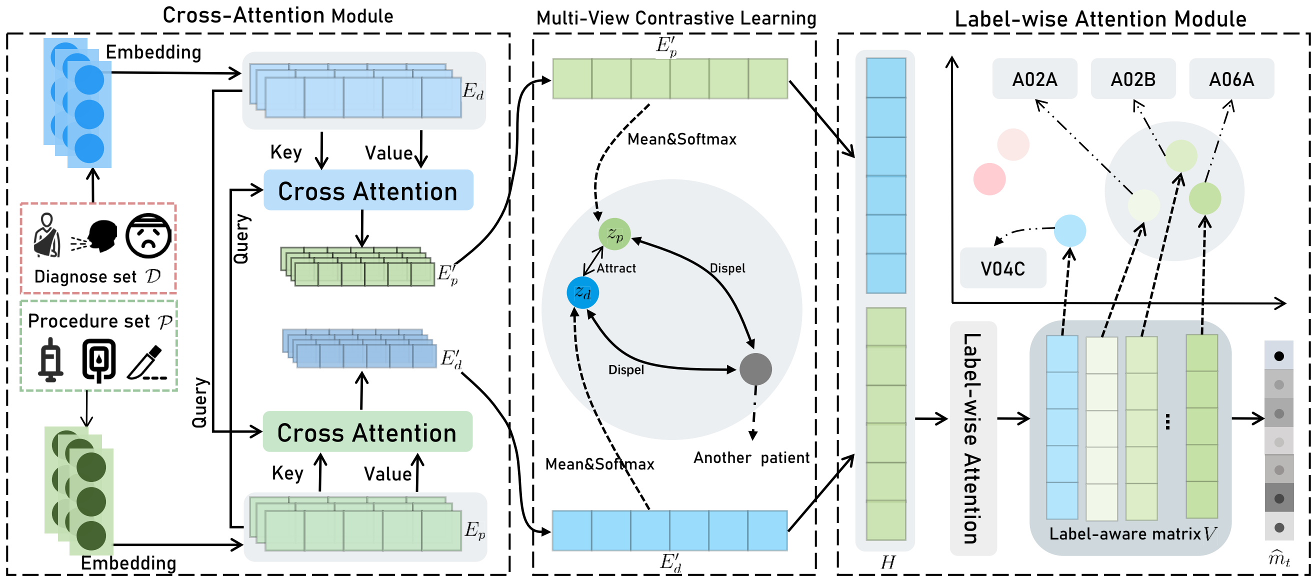

**Figure 2: Overall architecture of the LAMRec model. We first encode the diagnoses and procedure sequences through two cross-attention encoders to obtain two-view representations $E_d'$ and $E_p'$. Then we use a multi-view contrastive learning module to maximize the mutual information between the two-view representations. Afterwards, the two-view representations are concatenated to obtain a consistent patient representation $H$, which is then input into the label-wise attention module to obtain a label-ware matrix $V$. Finally, the prediction layer outputs the final drug recommendation $\widehat{m}_t$.**

visit like [30]:

$$PE_{(t,2i)} = sin(t/10000^{2i/dim}) \tag{3}$$

$$PE_{(t,2i+1)} = cos(t/10000^{2i/dim}) \tag{4}$$

where $i$ denotes the dimension. In this way, we can incorporate the temporal information of visits into the diagnosis embeddings. Overall, we can represent a patient's diagnosis representation as a matrix:

$$E_d = \begin{bmatrix} d_1' + PE_{(1)} \\ d_2' + PE_{(2)} \\ \dots \\ d_t' + PE_{(t)} \end{bmatrix} \in \mathbb{R}^{t\times dim} \tag{5}$$

Using the same method, we can obtain the patient's procedure representation $E_p \in \mathbb{R}^{t\times dim}$ through $\mathcal{E}_p$ and positional encoding.

*4.1.2 Cross-Attention Block.* This block aims to learn the attention weights between the historical visit data of each view to obtain view representations that incorporate information from another view. First, we define the scaled dot-product attention layer as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{6}$$

Here, $Q$, $K$, and $V$ represent the query matrix, key matrix, and value matrix, respectively. Each row corresponds to a visit vector from a certain view. Intuitively, this layer computes the weighted sum of all value vectors. The scaling factor $d$ is used to avoid large inner products, especially in high-dimensional spaces.

In the field of medical imaging, the information contained in different modalities (e.g., medical images and text reports) is often complementary. To fully exploit this cross-modal information, some works [33] have drawn inspiration from the cross-attention mechanism. We also adopt the cross-attention mechanism in LAMRec to ensure that information from different views can be fully integrated, thereby better representing the overall condition of the same patient. Specifically, we use the representation of one view (e.g., diagnoses) to calculate the query matrix Q and the representation of another view (e.g., procedures) to calculate the key matrix K and value matrix V. Formally, the cross-attention layer can be defined as:

$$C = CA\left(E_d, E_p\right) = Attention\left(E_d W^Q, E_p W^K, E_p W^V\right) \tag{7}$$

Here, $W^Q, W^K, W^V \in \mathbb{R}^{dim\times dim}$ are learnable weight matrices. In the computation process of cross-attention layer, for each visit vector in the diagnosis matrix $E_d$, the cross-attention mechanism assigns a weight score based on its similarity with each sub-vector in the procedure matrix $E_p$. These weight scores reflect the relevance between procedures at different time periods and the current diagnosis. In the aggregation stage, each sub-vector in the procedure matrix $E_p$ contributes to the final diagnosis representation according to its associated weight. In other words, the more relevant historical procedure information will have a greater proportion in the new diagnosis representation. Therefore, the output diagnosis matrix actually incorporates key information from the procedure view. It can reflect the patient's diagnosis condition from the view of procedures, thereby better representing the overall condition of the patient.

Although cross-attention layer can aggregate information from different views, it is still a linear model in the end. To endow the model with nonlinearity and consider the interactions between different latent dimensions, we apply a two-layer feedforward neural network to the cross-attention layer:

$$F = FFN(C) = ReLU(CW^1 + b^1)W^2 + b^2 \tag{8}$$

where $W^1, W^2 \in \mathbb{R}^{dim\times dim}$ are learnable parameter matrices, and $b^1, b^2 \in \mathbb{R}^{dim}$ are learnable bias parameters.

After going through the first cross-attention layer, $F$ essentially aggregates information from the embeddings of different views. However, using the output of the previous layer's cross-attention module as the input for the next layer allows for gradually learning more complex and higher-level view interaction patterns. Based on this understanding, we designed a stacked cross-attention structure. The output of the $b$-th cross-attention module is:

$$F^b = FFN(C^b) \tag{9}$$

$$C^b = CA\big(F^{(b-1)}\big) \tag{10}$$

Here, for the first layer, we define $C^1 = C$ and $F^1 = F$.

However, as the network becomes deeper, several problems become more severe. First, the increase in model parameters leads to overfitting, and models with more parameters often require more training time. Second, the training process becomes unstable due to issues like vanishing gradients. Inspired by [30], we perform the following operations to mitigate these problems:

$$g'(x) = LayerNorm\big(Dropout\big(g(x)\big) + x\big) \tag{11}$$

Here, $g(x)$ represents the output of the cross-attention sublayer or the feedforward neural network sublayer. In other words, for the output of each sublayer $g(x)$, we first perform a dropout operation, followed by residual connection and normalization to ensure stable convergence of the model.

To summarize, for a patient's temporal diagnosis representation matrix , we can obtain it using the following equation:

$$E'_d = Enc_d(E_d) = F^b \tag{12}$$

where $E'_d \in \mathbb{R}^{t\times dim}$ is the diagnosis representation matrix that incorporates information from the procedure view. Similarly, we can obtain a patient's procedure representation matrix $E'_p \in \mathbb{R}^{t\times dim}$ on the same way.

## 4.2 Multi-View Contrastive Learning

In the representation space learned by the cross-attention encoder, we consider the diagnoses and procedures information of the same patient as two views. Inspired by existing contrastive learning research [10, 16, 23], we directly maximize the mutual information between representations from different views through multi-view contrastive loss. In mathematics, our loss function is defined as:

$$\mathcal{L}_{mcl} = -\big(I\big(z_d; z_p\big) + \lambda\big(H(z_d) + H\big(z_p\big)\big)\big) \tag{13}$$

$$z_d = softmax\big(MEAN\big(E'_d\big)\big) \tag{14}$$

$$z_p = softmax\big(MEAN\big(E'_p\big)\big) \tag{15}$$

Here, $I$ and $H$ represent mutual information and entropy, respectively. The balancing factor $\lambda$ is to normalize the value of entropy. $z_d$ and $z_p \in \mathbb{R}^{dim}$ represent the patient's global representations from the diagnoses and procedure views, respectively.

The considerations for designing this loss function are as follows. First, a higher entropy value implies richer patient information. According to information theory [7], entropy can measure the average amount of information in an event. Therefore, maximizing entropy helps to retain more key information in individual views. Second, maximizing entropy can also avoid trivially mapping all instances to the same point in the representation space, thereby encouraging the learning of more diverse and discriminative patient representations.

To compute the basic measures of information in the multi-view contrastive loss, we first define the joint probability distribution $P$. Specifically, after mapping the global representation vectors through the softmax function, each element can be seen as the probability of a certain cluster. In other words, $z_d$ and $z_p$ can be considered as the distribution of two discrete clustering distribution variables $z$ and $z'$ over $dim$ classes, where $dim$ is the dimension of the representation. Therefore, we can define the joint probability distribution $P \in \mathbb{R}^{dim\times dim}$ as:

$$P = z_d(z_p)^T \tag{16}$$

Additionally, the marginal probability distributions $P(z = i)$ and $P(z' = j)$ are defined as $P_i$ and $P_j$, respectively. They can be obtained by summing over the $i$-th row and $j$-th column of the joint probability distribution matrix. Accordingly, the multi-view contrastive loss $\mathcal{L}_{mcl}$ can be represented as:

$$\mathcal{L}_{mcl} = -\sum_{i=1}^{dim}\sum_{j=1}^{dim} P_{dd'} In \frac{P_{ij}}{P_i^{\lambda+1} . P_{d^j}^{\lambda+1}} \tag{17}$$

where $P_{ij}$ represents the element in the $i$-th row and $j$-th column of $P$, and $\lambda$ is the balancing factor mentioned in the Eq.(13).

## 4.3 Label-wise Attention Module

*4.3.1 Patient Hidden Representation.* To generate a more compact patient representation $H$, we concatenate the patient's temporal diagnosis representation and procedure representation into a patient representation matrix by using a feedforward neural network $fc_1(.)$:$\mathbb{R}^{t\times 2dim} \rightarrow \mathbb{R}^{t\times dim}$:

$$H = fc_1([E'_d \# E'_p]) \tag{18}$$

where # is the concatenation operation.

*4.3.2 Label-wise Attention Mechanism.* Each patient has a different number of inpatient visits, and each visit corresponds to multiple medication codes. There are certain co-occurrence patterns between different medication codes, and our goal is to explicitly model the relationship between medication codes and the patient's medical history. To achieve this goal, we use the label-wise attention mechanism [22, 32]. Overall, the label-wise attention mechanism takes the patient representation $H$ as input and outputs a label-aware matrix $V$. This mechanism can be described in two steps. First, we compute a label-aware weight matrix:

$$Z = \tanh(HW) \tag{19}$$

$$A = softmax(ZU) \tag{20}$$

Here, $W \in \mathbb{R}^{dim \times s}$ and $U \in \mathbb{R}^{s \times |\mathcal{M}|}$ are linear transforms. $s$ is a hyperparameter to regulate the model. $A \in \mathbb{R}^{t \times |\mathcal{M}|}$ is the label-aware weight matrix, where the element at the $i$-th row and $j$-th column represents the relevance weight between the $j$-th medication code and the $i$-th visit. The softmax function is applied at the column level to ensure that the sum of attention weights for each medication code over all visits is 1. Afterwards, the attention weight matrix $A$ is multiplied with the patient representation matrix $H$ to generate a label-aware matrix which incorporates information from all historical visits:

$$V = A^T H \tag{21}$$

The $i$-th row of $V \in \mathbb{R}^{|\mathcal{M}| \times dim}$ represents the $i$-th medication code representation in the set , and this medication code representation integrates all visit information related to that code, capturing the evolving patterns in the medical history.

## 4.4 Prediction Layer

The prediction layer is the final step that utilizes the label-aware matrix to predict multi-label medications. For the label-aware matrix $V$, we feed it into a feedforward neural network $fc_2 : \mathbb{R}^{|\mathcal{M}| \times dim} \to \mathbb{R}^{|\mathcal{M}|}$ with only $|\mathcal{M}|$ output nodes:

$$\widehat{m}_t = \sigma(fc_2\,(V)) \tag{22}$$

where $\sigma$ is the sigmoid activation function, and the $i$-th element of $\widehat{m}_t \in \mathbb{R}^{|\mathcal{M}|}$ represents the probability of recommending the $i$-th medication at the $t$-th visit.

## 4.5 Loss Function

We train our LAMRec model using three distinct loss functions: (1) multi-label prediction loss $\mathcal{L}_{bce}$ for supervised training of medication combinations; (2) DDI control loss $\mathcal{L}_{ddi}$ for constraining the recommended medication combinations; (3) multi-view contrastive loss $\mathcal{L}_{mcl}$ for enhancing the mutual information between dual views.

*4.5.1 Multi-label Prediction Loss.* We treat the prediction of each medication as a subtask of drug recommendation and use the common binary cross-entropy loss function, which can be represented as:

$$\mathcal{L}_{bce} = -\sum_{i=1}^{|\mathcal{M}|} m_{t,i} \log(\widehat{m}_{t,i}) + (1 - m_{t,i}) \log(1 - \widehat{m}_{t,i}) \tag{23}$$

where $\widehat{m}_t \in \mathbb{R}^{|\mathcal{M}|}$ is the predicted medications for the patient's $t$-th visit, and $m_t \in \mathbb{R}^{|\mathcal{M}|}$ is the target medications for the patient's $t$-th visit.

*4.5.2 DDI Control Loss.* The DDI control loss is used to control the drug-drug interaction reactions in the medication combinations. For the predicted medication combinations, we want it to achieve a lower DDI rate. We can achieve our goal by minimizing the following loss:

$$\mathcal{L}_{ddi} = \sum_{i=1}^{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} \left( A \odot \left( \widehat{m}_t (\widehat{m}_t)^T \right) \right) [i, j] \tag{24}$$

where $\widehat{m}_t (\widehat{m}_t)^T \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}$ represents the probability of two medications appearing simultaneously in the recommended combination. $\odot$ is the element-wise product.

*4.5.3 Combined Loss functions.* During the training process, both the accuracy and DDI rate tend to increase simultaneously. This is because in the real labels of EHRs data, medication combinations often have DDI as well. Therefore, whether the predicted medications are incorrect or correct, they may both increase the DDI rate. This leads to a difficulty in balancing the model's accuracy and safety, and how to trade off between the two becomes a key issue. Inspired by [26], we set a safe DDI threshold $\theta$ to balance the model's accuracy and safety. In summary, our final loss function is as follows:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{bce} + \alpha \mathcal{L}_{ddi} + \beta \mathcal{L}_{mcl} & if\ ddi \geq \theta \\ \mathcal{L}_{bce} + \beta \mathcal{L}_{mcl} & if\ ddi < \theta \end{cases} \tag{25}$$

where $ddi$ is the DDI score calculated based on the current $\widehat{m}_t$, and $\theta$ is our preset DDI threshold. $\alpha$ and $\beta$ are balancing factors used to control the DDI loss function and the multi-view contrastive loss, respectively.

# 5 EXPERIMENTS

## 5.1 Datasets

In the experiments, we used the following three datasets: MIMIC-II [24], MIMIC-III [14], and MIMIC-IV [13]. These datasets are widely-used de-identified health record datasets for ICU patients, publicly available at https://mimic.physionet.org/. We employed PyHealth [35] to process these datasets. Table 1 summarizes the statistics of the processed datasets.

**Table 1: Statistics of the Datasets**

| Items | MIMIC-II | MIMIC-III | MIMIC-IV |
|---|---|---|---|
| # of patients | 3632 | 5449 | 32633 |
| # of visits | 9779 | 14141 | 105441 |
| # of diagnosis | 3840 | 4491 | 17267 |
| # of procedure | 1247 | 1412 | 8923 |
| # of medication | 192 | 193 | 199 |
| avg # of visits | 2.6925 | 2.5952 | 3.2311 |
| avg # of diagnosis | 31.5896 | 31.6520 | 47.1405 |
| avg # of procedure | 10.2384 | 9.4601 | 8.0635 |
| avg # of medication | 27.2665 | 29.7427 | 24.6683 |

In the data preprocessing stage, we standardized the diagnosis and procedures codes according to the International Classification of Diseases, Ninth Revision (ICD-9), and normalized the medication data according to the ATC level 3 codes.

## 5.2 Baselines

We compared LAMRec with the following baseline models (MLP and Transformer are designed by ourselves):

- **MLP** is a simple multi-layer perceptron neural network. It directly inputs the patient representation $H$ into a feedforward neural network, without performing sequence modeling on $d'_t$ and $p'_t$. thus not capturing temporal relationships between visits.

**Table 2: Performance on MIMIC-II, MIMIC-III and MIMIC-IV in terms of DDI rate, Jaccard, F1-score, PRAUC, ROCAUC and Avg.# of medications. The best and runner-up results are highlighted in bold and underline respectively.**

| Dataset | Method | DDI ↓ | Jaccard ↑ | PRAUC ↑ | F1-score ↑ | ROCAUC ↑ | Avg.# of medications |
|---|---|---|---|---|---|---|---|
| MIMIC-II | MLP | 0.0727 ± 0.0022 | 0.4354 ± 0.0041 | 0.6990 ± 0.0026 | 0.5961 ± 0.0038 | 0.9180 ± 0.0021 | 32.4238 ± 0.7586 |
| | RETAIN | 0.0824 ± 0.0039 | 0.4375 ± 0.0031 | 0.6907 ± 0.0030 | 0.5976 ± 0.0031 | 0.9145 ± 0.0020 | 28.6522 ± 0.7783 |
| | Transformer | 0.0737 ± 0.0031 | 0.4464 ± 0.0033 | 0.7119 ± 0.0042 | 0.6055 ± 0.0033 | 0.9225 ± 0.0015 | 24.5889 ± 0.6128 |
| | GAMENet | 0.0756 ± 0.0015 | 0.4495 ± 0.0028 | 0.7047 ± 0.0061 | 0.6105 ± 0.0029 | 0.9181 ± 0.0034 | 34.8824 ± 1.8680 |
| | SafeDrug | 0.0738 ± 0.0026 | 0.4519 ± 0.0052 | 0.7074 ± 0.0071 | 0.6121 ± 0.0053 | 0.9196 ± 0.0027 | 30.6591 ± 1.2472 |
| | MICRON | 0.0667 ± 0.0013 | 0.4670 ± 0.0039 | 0.7398 ± 0.0051 | 0.6250 ± 0.0036 | 0.9310 ± 0.0016 | 36.7053 ± 1.1301 |
| | MoleRec | 0.0697 ± 0.0027 | 0.4702 ± 0.0039 | 0.7232 ± 0.0060 | 0.6298 ± 0.0035 | 0.9234 ± 0.0044 | 31.1766 ± 1.1817 |
| | LAMRec | **0.0632 ± 0.0044** | **0.4866 ± 0.0027** | **0.7448 ± 0.0022** | **0.6444 ± 0.0025** | **0.9336 ± 0.0012** | 29.5951 ± 1.0414 |
| MIMIC-III | MLP | 0.0713 ± 0.0039 | 0.4312 ± 0.0030 | 0.6949 ± 0.0031 | 0.5908 ± 0.0029 | 0.9116 ± 0.0010 | 33.3912 ± 2.1828 |
| | RETAIN | 0.0722 ± 0.0016 | 0.4479 ± 0.0068 | 0.7078 ± 0.0061 | 0.6062 ± 0.0061 | 0.9145 ± 0.0020 | 31.1821 ± 1.2127 |
| | Transformer | 0.0652 ± 0.0012 | 0.4436 ± 0.0087 | 0.7039 ± 0.0042 | 0.5995 ± 0.0086 | 0.9163 ± 0.0013 | 28.8262 ± 0.9151 |
| | GAMENet | 0.0700 ± 0.0018 | 0.4519 ± 0.0019 | 0.7142 ± 0.0050 | 0.6114 ± 0.0017 | 0.9164 ± 0.0025 | 35.8954 ± 0.9630 |
| | SafeDrug | 0.0704 ± 0.0020 | 0.4500 ± 0.0060 | 0.7082 ± 0.0087 | 0.6084 ± 0.0057 | 0.9140 ± 0.0035 | 33.0478 ± 1.5002 |
| | MICRON | 0.0627 ± 0.0008 | 0.4635 ± 0.0028 | 0.7399 ± 0.0045 | 0.6202 ± 0.0026 | 0.9272 ± 0.0013 | 38.8207 ± 0.5758 |
| | MoleRec | 0.0658 ± 0.0014 | 0.4650 ± 0.0032 | 0.7186 ± 0.0097 | 0.6232 ± 0.0024 | 0.9190 ± 0.0025 | 32.6386 ± 1.6881 |
| | LAMRec | **0.0616 ± 0.0020** | **0.4796 ± 0.0037** | **0.7448 ± 0.0036** | **0.6360 ± 0.0030** | **0.9301 ± 0.0015** | 30.9211 ± 1.7195 |
| MIMIC-IV | MLP | 0.0668 ± 0.0036 | 0.4106 ± 0.0027 | 0.6755 ± 0.0039 | 0.5639 ± 0.0025 | 0.9170 ± 0.0008 | 26.7705 ± 0.8927 |
| | RETAIN | 0.0684 ± 0.0014 | 0.4250 ± 0.0032 | 0.6931 ± 0.0025 | 0.5783 ± 0.0031 | 0.9253 ± 0.0011 | 23.3900 ± 0.7389 |
| | Transformer | 0.0650 ± 0.0008 | 0.4403 ± 0.0020 | 0.7065 ± 0.0035 | 0.5935 ± 0.0022 | 0.9285 ± 0.0009 | 26.0508 ± 1.0818 |
| | GAMENet | 0.0683 ± 0.0010 | 0.4415 ± 0.0014 | 0.7051 ± 0.0029 | 0.5960 ± 0.0014 | 0.9236 ± 0.0018 | 28.0243 ± 0.3344 |
| | SafeDrug | 0.0681 ± 0.0022 | 0.4333 ± 0.0015 | 0.6921 ± 0.0033 | 0.5872 ± 0.0013 | 0.9205 ± 0.0013 | 26.0622 ± 0.8937 |
| | MICRON | 0.0644 ± 0.0022 | 0.4544 ± 0.0026 | 0.7454 ± 0.0016 | 0.6075 ± 0.0022 | 0.9405 ± 0.0009 | 33.4993 ± 0.9084 |
| | MoleRec | 0.0625 ± 0.0009 | 0.4368 ± 0.0030 | 0.7005 ± 0.0068 | 0.5924 ± 0.0029 | 0.9233 ± 0.0018 | 25.4723 ± 0.4353 |
| | LAMRec | **0.0615 ± 0.0016** | **0.4683 ± 0.0011** | **0.7467 ± 0.0016** | **0.6198 ± 0.0013** | **0.9415 ± 0.0006** | 26.9675 ± 1.9922 |

↓ means the corresponding metric is the lower the better and ↑ means the opposite.

- **Transformer** is based on the self-attention mechanism. Unlike MLP, it performs sequence modeling on $d'_t$ and $p'_t$, then inputs $H$ into a feedforward neural network for the final output.
- **RETAIN [4]** predicts a patient's future medication combinations based on their historical visit records, utilizing RNN and attention mechanism.
- **GAMENet [26]** innovatively combines RNN, memory networks, and graph neural networks, leveraging the knowledge of drug co-occurrence graphs and DDI graphs.
- **SafeDrug [37]** fuses drug molecular structure and drug interaction information, and uses graph neural networks to learn the relationships between medications.
- **MICRON [36]** utilizes a recurrent residual learning model to capture drug changes between consecutive visits and makes recommendations based on these changes and the patient's recent medication combinations.
- **MoleRec [38]** improves drug recommendation by leveraging molecular substructure interactions and patient-substructure relevance to identify efficacy-driving substructures.

The code for the baselines was implemented with the assistance of the PyHealth [35] framework.

In the following experiments, we used the following metrics to evaluate the performance of the models: Jaccard Similarity Score (Jaccard), Precision Recall AUC (PRAUC), Average F1 score (F1-score), Area Under ROC Curve (AUCROC), DDI rate, and average number of medications.

## 5.3 Hyperparameters

All experiments for the models were conducted on an NVIDIA GeForce RTX 3090 (24GB). We randomly partitioned the dataset into training, validation, and test sets at an 8:1:1 ratio. To identify the optimal hyperparameter combination, we employed the annealing algorithm on the validation set. Specifically, we sampled initial hyperparameters from a prior point and progressively adjusted the sampling point over time to converge closer to the observed best point, thereby determining a suitable hyperparameter combination.

For the MIMIC-III dataset, the main parameters of LAMRec were set as follows: the batch size was set to 256, the values of both $\alpha$ and $\beta$ were 0.1, $\lambda$ was 8, number of cross-attention layers was 2, the number of attention heads was 8, the dropout for the cross-attention module was 0.1, and both the embedding dimension and hidden dimension were set to 512(i.e., $dim$= 512). For a fair comparison, all other baselines that employed GRU, we configured the embedding dimension and hidden layer dimension to 512, the number of GRU layers to 2, and dropout to 0.5. Training was performed using the Adam optimizer with a learning rate of 5e-4 until convergence was achieved. In the validation set, we identified the best-performing model within 50 epochs as the final model and reported its metrics in the test set.

## 5.4 Performance Comparison

We conducted 5 randomized experiments using random seeds and presented the mean and standard deviation of each metric in Table 2.

**Table 3: Ablation study on MIMIC-III in terms of DDI rate, Jaccard, F1-score, and PRAUC. The best and runner-up results are highlighted in bold and underline respectively.**

| Method | DDI ↓ | Jaccard ↑ | PRAUC ↑ | F1-score ↑ |
|---|---|---|---|---|
| LAMRec (our full version) | **0.0614 ± 0.0020** | **0.4792 ± 0.0040** | **0.7450 ± 0.0037** | **0.6356 ± 0.0033** |
| $\mathcal{W}/O$ cross-attention | 0.0667 ± 0.0012 | 0.4726 ± 0.0050 | 0.7394 ± 0.0044 | 0.6305 ± 0.0043 |
| $\mathcal{W}/O$ multi-view CL | 0.0624± 0.0037 | 0.4726 ± 0.0047 | 0.7428 ± 0.0045 | 0.6292 ± 0.0044 |
| $\mathcal{W}/O$ label-wise attention | 0.0668 ± 0.0027 | 0.4713 ± 0.0040 | 0.7398 ± 0.0041 | 0.6284 ± 0.0037 |

↓ means the corresponding metric is the lower the better and ↑ means the opposite.

The experimental results demonstrate that LAMRec outperforms all baseline models on all datasets, particularly exhibiting significant improvements in Jaccard and F1-score. This proves that LAMRec can more accurately recommend medication combinations that closely resemble the true medical orders. Moreover, the medication combinations recommended by LAMRec are safer, with a notably lower DDI rate compared to other models.

In contrast, MLP, as a simple multilayer perceptron model, performs the worst across all metrics due to its inability to effectively utilize patients' longitudinal sequence information. Although RETAIN and Transformer encode patients' historical visits through attention mechanisms, their effectiveness remains limited. GAMENet and SafeDrug integrate auxiliary information such as medication knowledge graphs on top of modeling temporal data, resulting in improved performance compared to previous models. MICRON considers the continuity of medications between different patient visits and better captures the changes between visits through a recurrent residual network, achieving superior results. Furthermore, MoleRec achieves superior performance compared to SafeDrug by better integrating information through modeling interactions among molecular substructures and the relevance between patients and substructures.

However, all of the above models overlook the complementary information and consistency contained within different views in EHRs data. They model each view separately, failing to uncover patients' consistent representations across different views. Unlike GAMENet, SafeDrug and MoleRec, which require external information such as medication knowledge graphs, LAMRec relies solely on patients' own EHRs data. Through a more powerful dual-view representation module, LAMRec generates more comprehensive and robust patient features. Furthermore, while the above models focus only on patient feature representation, they neglect the varying importance of different medication labels for the drug recommendation task. By incorporating label-wise attention mechanism, LAMRec enables the model to adaptively focus on the relevance of different medication labels to the historical patient visit, leading to more personalized and precise combination recommendations. With these two major improvements, LAMRec ultimately surpasses existing baseline models by a substantial margin on all datasets.

## 5.5 Ablation Study

To comprehensively validate the effectiveness of each module in LAMRec, we designed the following ablation experiments on the MIMIC-III dataset. Specifically, we considered three variant models:

- $\mathcal{W}/O$ **cross-attention.** This variant removes the cross-attention module of LAMRec and instead uses GRU to encode patients' historical visit sequences, similar to other baseline models.
- $\mathcal{W}/O$ **multi-view CL.** This variant removes the multi-view contrastive learning loss $\mathcal{L}_{mcl}$ from LAMRec. $E'_d$ and $E'_p$ are directly fed into the downstream label-wise attention module after being encoded by the cross-attention mechanism, without imposing any consistency constraints between views.
- $\mathcal{W}/O$ **label-wise attention.** After obtaining the comprehensive patient representation, this variant directly feeds it into a feed-forward neural network for the final medication combination prediction, bypassing the label-wise attention module.

The ablation study results in Table 3 show that the complete LAMRec model significantly outperforms these three variants across all metrics, fully demonstrating the indispensability of each module in the drug recommendation task.

The performance of $\mathcal{W}/O$ cross-attention is notably inferior to LAMRec, confirming that the cross-attention mechanism can effectively integrate complementary information from different views, substantially improving the quality of patient representations. This is because the cross-attention layers allow the diagnosis and procedure representations to mutually learn from and influence each other during the encoding process, enabling the exchange of implicit common features between the two views. In contrast, other sequence modeling methods disconnect the associations between different views, limiting the richness of the representations.

Moreover, the performance of $\mathcal{W}/O$ multi-view CL also decreases significantly, indicating the crucial role of the multi-view contrastive learning loss in enhancing the consistency of patient representations. By minimizing the mutual information divergence between representations from different views, contrastive learning promotes the alignment of diagnosis and procedure representations in the latent space. Without such consistency constraints, simply concatenating the raw information from different views makes it difficult to obtain consistent and robust patient representations.

Lastly, the performance of $\mathcal{W}/O$ label-wise attention is also inferior to the complete model, highlighting the value of the label-wise attention mechanism in personalized medication recommendations. By introducing label-wise attention layer, LAMRec can adaptively adjust the relevance weights of each medication to the current patient visit. Unlike directly matching patient representations to all medication labels, the label-wise attention mechanism can selectively focus on medications that are more important for the current decision, thereby providing more precise and personalized combination recommendations.

**Table 4: Examples of recommended drug sets for given diagnosis groups on MIMIC-IV. Here "wrong" refers to drugs that are not in the ground truth set but are predicted, while "interacting" indicates the combinations in the predicted drug set that have DDI interactions.**

| Method | Recommend medications |
|---|---|
| Ground Truth<br>Num: 23 | **23 correct**:A02A, A02B, A03F, A06A, A11D, A12B, A12C, B02B, B03B, B05A, B05X, C03C, C03D, C07A, H01C, J01D, J01M, N01A, N05A, N05C, R03A, V04C, V06D<br>**10 interacting**:(A02B, C07A), (A02B, J01M), (A02B, N05B), (B01A, N01A), (B01A, N02A), (C07A, J01M), (J01M, J01X), (J01M, N01A), (N01A, N02A), (N02A, N05B) |
| GAMENet<br>Num: 22<br>Recall=13/23<br>Precision=13/22 | **13 correct**:A02B, A06A, A12B, A12C, B05A, B05X, C03C, C07A, J01M, N01A, N05C, V04C, V06D<br>**9 wrong**:A10A, A12A, B01A, C01C, J01X, M03A, N02A, N02B, N05B<br>**13 interacting**:(A02B, C07A), (A02B, J01M), (A02B, N02B), (A02B, N05B), (B01A, N01A), (B01A, N02A), (C07A, J01M), (J01M, J01X), (J01M, N01A), (J01M, N02B), (M03A, N01A), (N01A, N02A), (N02A, N05B) |
| MICRON<br>Num: 17<br>Recall=14/23<br>Precision=14/17 | **14 correct**:A02B, A06A, A12B, B05X, C03C, C03D, C07A, H01C, J01M, N01A, N05C, R03A, V04C, V06D<br>**3 wrong**:N02A, N02B, N05B<br>**9 interacting**:(A02B, C07A), (A02B, J01M), (A02B, N02B), (A02B, N05B), (C07A, J01M), (J01M, N01A), (J01M, N02B), (N01A, N02A), (N02A, N05B) |
| MoleRec<br>Num: 19<br>Recall=14/23<br>Precision=14/19 | **14 correct**:A02B, A06A, A12B, B02B, B05X, C03C, C03D, C07A, H01C, J01M, N01A, N05C, V04C, V06D<br>**5 wrong**:A12A, J01X, N02A, N02B, N05B<br>**11 interacting**:(A02B, C07A), (A02B, J01M), (A02B, N02B), (A02B, N05B), (B02B, N02B), (C07A, J01M), (J01M, J01X), (J01M, N01A), (J01M, N02B), (N01A, N02A), (N02A, N05B) |
| LAMRec<br>Num: 17<br>Recall=16/23<br>Precision=16/17 | **16 correct**:A02B, A06A, A12B, B02B, B05A, B05X, C03C, C03D, C07A, H01C, J01M, N01A, N05C, R03A, V04C, V06D<br>**1 wrong**:N05B<br>**5 interacting**:(A02B, C07A), (A02B, J01M), (A02B, N05B), (C07A, J01M), (J01M, N01A) |

In summary, each innovative module of LAMRec plays a crucial role in improving the overall performance, and they are all indispensable. The seamless integration of the three modules ultimately enables LAMRec to achieve optimal performance across all evaluation metrics, fully demonstrating the rationality and effectiveness of our model design.

## 5.6 Case Study

To intuitively show the advantages of LAMRec compared to other models, we selected representative case from the test set. By comparing the recommendation results and key metrics of each model, We can make the following observations from Table 4.

LAMRec achieves the best performance in both recall and precision for medication combination recommendations. In contrast, GAMENet over-recommends medications to improve recall, resulting in a significant decrease in precision. MICRON and MoleRec attempt to improve precision by recommending higher-quality medication combinations, but due to their limited modeling capability, both recall and precision still lag significantly behind LAMRec.

Moreover, the medication combinations recommended by LAMRec also have clear advantages in terms of safety. LAMRec generates the least number of interacting medication pairs and has the lowest DDI rate, even outperforming the ground truth prescriptions written by doctors. It is noteworthy that we found all baseline models (including the ground truth) failed to avoid certain typical pairing errors. For example, (N01A, N02A) and (N02A, N05B), two pairs of contraindicated medications, frequently co-occurred in their recommendation results. Only LAMRec can consistently avoid these obvious contraindicated pairings. This indicates that by integrating multi-view information and label information, LAMRec can model patient characteristics more comprehensively and meticulously, thereby identifying those implicit pairing risks. The combination of this strong representational capability and innovative mechanisms is the fundamental reason for LAMRec's superiority in medication combination safety.

In conclusion, the case analysis clearly demonstrates LAMRec's comprehensive advantages in the drug recommendation task. It achieves a balance across multiple objectives, including recommendation effectiveness, medication quantity, and medication combination safety, significantly outperforming existing baseline models. These outstanding performances ultimately stem from LAMRec's unique cross-attention mechanism, multi-view contrastive learning, and label-wise attention, among other innovative designs. LAMRec's excellent performance on real-world cases confirms that the model we proposed is effective and outstanding.

# 6 CONCLUSION

In this paper, we propose a label-aware multi-view drug recommendation model (LAMRec) to address the issues of inconsistent patient representations and insufficient label information exploitation in existing medication recommendation models. We extract latent information from dual views using a cross-attention mechanism and design a multi-view contrastive learning loss to maximize the mutual information between the dual views. Furthermore, we fully exploit medication label information by using a label-wise attention mechanism, and establish connections between medications and patient visits. We compare our model with baselines and achieve significant performance improvements on the inpatient MIMIC-II, MIMIC-III, and MIMIC-IV datasets. Finally, both ablation experiments and case studies demonstrate the validity and superiority of our model.

# ACKNOWLEDGMENTS

This work was supported by Shandong Provincial Natural Science Foundation (No.ZR2022QF114) and NSFC(No.92367202). And Lei Liu is supported by Taishan Scholar Program.

# REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[3] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*. PMLR, 301–318.

[4] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).

[5] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2017. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association* 24, 2 (2017), 361–370.

[6] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341* (2019).

[7] TM Cover and JA Thomas. 2006. Elements of Information Theory. Hoboken, NJ, USA: John Wiely & Sons.

[8] Jiajun Cui, Zeyuan Chen, Aimin Zhou, Jianyong Wang, and Wei Zhang. 2023. Fine-grained interaction modeling with multi-relational transformer for knowledge tracing. *ACM Transactions on Information Systems* 41, 4 (2023), 1–26.

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[10] 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*. 9729–9738.

[11] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[12] Nan Jiang, Haitao Yuan, Jianing Si, Minxiao Chen, and Shangguang Wang. 2024. Towards Effective Next POI Prediction: Spatial and Semantic Augmentation with Remote Sensing Data. *arXiv preprint arXiv:2404.04271* (2024).

[13] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)* (2020), 49–55.

[14] 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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[15] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 8547–8555.

[16] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4447–4461.

[17] Ning Liu, Wei Zhang, Xiuxing Li, Haitao Yuan, and Jianyong Wang. 2020. Coupled graph convolutional neural networks for text-oriented clinical diagnosis inference. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I 25*. Springer, 369–385.

[18] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems* 29 (2016).

[19] Chaitanya Malaviya, Pedro Ferreira, and André FT Martins. 2018. Sparse and constrained attention for neural machine translation. *arXiv preprint arXiv:1805.08241* (2018).

[20] Chengsheng Mao, Liang Yao, and Yuan Luo. 2019. Medgcn: Graph convolutional networks for multiple medical tasks. *arXiv preprint arXiv:1904.00326* (2019).

[21] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6707–6717.

[22] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695* (2018).

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[24] Mohammed Saeed, Christine Lieu, Greg Raber, and Roger G Mark. 2002. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in cardiology*. IEEE, 641–644.

[25] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

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

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.

[28] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1255–1265.

[29] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576* (2020).

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[31] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat* 1050, 20 (2017), 10–48550.

[32] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. *arXiv preprint arXiv:2007.06351* (2020).

[33] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 33536–33549.

[34] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 25, 10 (2018), 1419–1428.

[35] Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P Danek, and Jimeng Sun. 2023. Pyhealth: A deep learning toolkit for healthcare applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5788–5789.

[36] Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. Change matters: Medication change prediction with recurrent residual networks. *arXiv preprint arXiv:2105.01876* (2021).

[37] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711* (2021).

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

[39] Xiao Yang, Ning Liu, Jianbo Qiao, Haitao Yuan, Teng Ma, Yonghui Xu, and Lizhen Cui. 2022. Clinical Phenotyping Prediction via Auxiliary Task Selection and Adaptive Shared-Space Correction. In *CAAI International Conference on Artificial Intelligence*. Springer, 438–449.

[40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[41] Haitao Yuan, Gao Cong, and Guoliang Li. 2024. Nuhuo: An Effective Estimation Model for Traffic Speed Histogram Imputation on A Road Network. *Proceedings of the VLDB Endowment* 17, 7 (2024), 1605–1617.

[42] Wei Zhang, Xin Lai, and Jianyong Wang. 2020. Social link inference via multiview matching network from spatiotemporal trajectories. *IEEE transactions on neural networks and learning systems* 34, 4 (2020), 1720–1731.

[43] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. 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*. 1315–1324.