

# MedAlign: Enhancing Combinatorial Medication Recommendation with Multi-modality Alignment

Hang Lv  
Fuzhou University  
Fuzhou, China  
lvhang@gmail.com

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

Zijie Wu  
Clinical Oncology School of Fujian  
Medical University  
Fuzhou, China  
wuzijie@fjzlhospital.com

Yanchao Tan\*  
Fuzhou University  
Fuzhou, China  
yctan@fzu.edu.cn

Guofang Ma  
Zhejiang Gongshang University  
Zhejiang, China  
maguofang@zjgsu.edu.cn

Zhigang Lin  
The First Affiliated Hospital of Fujian  
Medical University  
Fuzhou, China  
1135@mju.edu.cn

Xiping Chen  
Hangzhou Bywin Technology Co., Ltd. The Chinese University of Hong Kong  
Zhejiang, China  
chenxp@bywin.cn

Hong Cheng  
Hong Kong, China  
hcheng@se.cuhk.edu.hk

Carl Yang  
Emory University  
Atlanta, USA  
j.carlyang@emory.edu

## Abstract

Combinatorial Medication Recommendation (CMR) based on multi-modal Electronic Health Records (EHRs) is a promising yet challenging frontier in AI-driven healthcare. Existing approaches usually rely on feature extraction from individual modalities without explicitly aligning information across different data sources. As a result, they may ignore complementary information from other modalities, leading to suboptimal representations for CMR. To this end, we propose MedAlign, a novel combinatorial Medication recommendation framework with multi-modality Alignment. Specifically, we first design a distribution-aware multimodal medication alignment module. This aligns distinct modality distributions of medications within a unified latent space, generating consistent medication representations. Furthermore, we introduce a longitudinal multi-view patient aggregation module, which aggregates the historical visits of patients with multi-view information to form informative patient representations. Finally, we propose a combinatorial medication recommendation module, enabling an accurate and safe medication recommendation combination for each patient. Extensive experiments on two real-world multimodal EHR datasets demonstrate the effectiveness of our MedAlign.

## CCS Concepts

• Applied computing → Health informatics.

\*Yanchao Tan is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permission from permissions@acm.org.  
MM '25, Dublin, Ireland.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-2035-2/2025/10  
<https://doi.org/10.1145/3746027.3755265>

## Keywords

Combinatorial Medication Recommendation, Multi-modality Alignment, Electronic Health Records

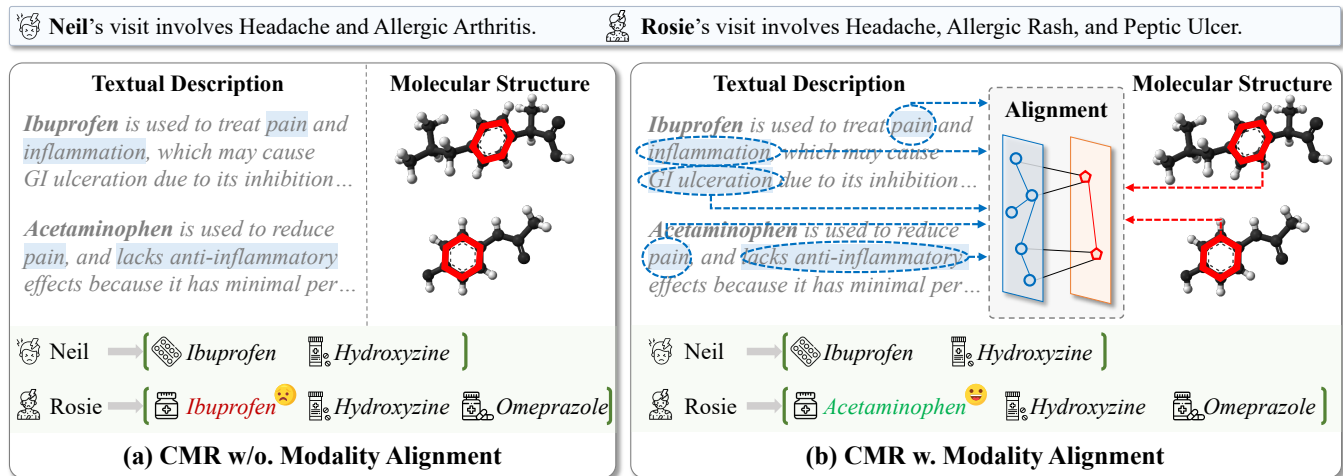
### ACM Reference Format:

Hang Lv, Zixuan Guo, Zijie Wu, Yanchao Tan, Guofang Ma, Zhigang Lin, Xiping Chen, Hong Cheng, and Carl Yang. 2025. MedAlign: Enhancing Combinatorial Medication Recommendation with Multi-modality Alignment. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3755265>

## 1 Introduction

Electronic Health Records (EHRs) are widely used in various real-world healthcare applications, such as diagnose prediction [21, 28] and Combinatorial Medication Recommendations (CMR) [20, 26]. These records encompass a wide range of data modalities, including texts (e.g., clinical notes and medication textual descriptions), structures (e.g., molecular structures of medications), images (e.g., X-rays and ultrasound scans), and signals (e.g., sensor records). The integration of these diverse data modalities can improve the accuracy and safety of CMR, by providing more effective representations of patient status and medication characteristics. For instance, [35, 36, 40] incorporate the molecular structures of medications and capture the interactions among substructures, improving the recommendation safety.

Although these CMR methods leverage multimodal data, they typically extract features from each modality independently without explicitly aligning information across modalities. Consequently, they often capture only shared inter-modal correlations while overlooking cross-model complementary information, leading to suboptimal medication combinations. As shown in Figure 1(a), traditional methods tend to learn the common pain-relief effect of *Ibuprofen* and *Acetaminophen* based on their shared textual description *pain* and molecular substructure *aromatic ring*, thereby recommending



**Figure 1: An illustrative example of Combinatorial Medication Recommendations (CMR). (a) Medications are recommended based on individual modalities without considering the alignment between them. (b) Alignment between different modalities (e.g., textual description and molecular structure) improves the consistent and complementary information among medications.**

inaccurate medication *Ibuprofen* to treat Rosie’s headache. This may increase the risk of *gastrointestinal complications* [13].

In contrast, by aligning heterogeneous modalities into a shared and comparable latent space, the model effectively captures both consistent and complementary characteristics of medications from textual and molecular modalities (illustrated in Figure 1(b)). As a result, *Acetaminophen* is preferred for Rosie, diagnosed with *Peptic Ulcer*, because it causes less *gastrointestinal irritation* [1]. Moreover, for Neil, who has *Allergic Arthritis*, *Ibuprofen* is recommended due to its additional *anti-inflammatory properties*. This illustrates that explicitly aligning multimodal information enables the model to uncover valuable correlations and complementary information among modalities, generating precise and safe medication combinations.

To this end, we propose a novel combinatorial **Medication recommendation framework with multi-modality Alignment (MedAlign)**. Specifically, we first design a distribution-aware multimodal medication alignment module. We align three distinct modality distributions of medications within a unified latent space, generating consistent medication representations. Then, we introduce a longitudinal multi-view patient aggregation module. By aggregating the historical visits of patients with multi-view information, we form informative longitudinal patient representations. Finally, we propose a combinatorial medication recommendation module, enabling an accurate and safe medication recommendation combination.

The main contributions of our work are summarized as follows: (1) We propose a novel distribution-aware multimodal medication alignment module for CMR. To the best of our knowledge, this is the first work from a distribution alignment perspective to explicitly capture both correlations and complementary information among modalities, achieving an accurate and safe CMR. (2) We integrate the rich temporal visit sequences with multi-view information from diagnoses, procedures, and fused historical medication, obtaining comprehensive longitudinal patient representations for medication combination prediction. (3) Extensive experiments on two real-world multimodal EHR datasets demonstrate the superiority of MedAlign over state-of-the-art baselines.

## 2 Related Work

### 2.1 Medication Recommendation

Combinatorial Medication Recommendation (CMR) aims to provide accurate and safe prescriptions for patients via personalized treatment [25, 27, 40]. Existing approaches are mainly divided into two categories: instance-based and longitudinal methods. Instance-based methods [8, 22] typically rely on structured features extracted from a single patient visit. For example, LEAP [39] proposed a multi-instance multi-label learning framework to generate medication recommendations based on the patient’s current diagnosis information. However, these methods often neglect valuable historical patient data. In contrast, longitudinal methods [14, 32, 34] integrate temporal information from patients’ hospitalization histories to model long-term disease progression. For instance, GAMENet [24] modeled longitudinal EHR data as a graph structure, incorporating Drug-Drug Interaction (DDI) with medication knowledge bases to mitigate potential conflicts.

To improve the safety and accuracy of CMR, recent studies have explored the integration of multimodal information, including molecular structures and textual descriptions. For instance, SafeDrug [35] enhanced recommendation precision by using molecular structure embeddings, moving beyond reliance on medication history and reducing the risk of DDIs. MoleRec [36] and DEPOT [40] focused on molecular substructure interactions to better reflect pharmacological properties. Furthermore, NLA-MMR [26] applied a cross-modal module to jointly learn from chemical medication structures and textual descriptions. Although the above methods effectively extract features from individual modalities, they often overlook cross-modal correlations and complementary information, thereby limiting their potential to enhance CMR accuracy.

### 2.2 Multimodal Learning in Healthcare

Recently, multimodal Electronic Health Records (EHRs) (e.g., clinical notes, medication molecular structures, and X-rays) have been widely used in various healthcare applications to model accurate

patients’ health states [28, 31, 40]. However, the inherent heterogeneity across modalities hinders the effective multimodal information integration [11]. To bridge the representation gap among heterogeneous modalities, several studies have explored methods to fuse and extract similar features across different data sources [18]. For instance, SMART [3] proposed a deep self-weighted multimodal relevance-weighting approach, which leverages clustering-based contrastive learning and eliminates intra- and inter-modal irrelevancy. MV-Mol [19] leveraged Q-Former, a multi-modal fusion architecture, to extract molecular representations by jointly comprehending molecular structures and view prompts. In addition, FlexCare [33] introduced a multimodal information extraction module to learn intra- and inter-modality features, along with a task-guided hierarchical fusion module for adaptive and task-specific representation learning. DrFuse [37] fused EHR data and medical images via disease-aware attention to handle missing and inconsistent modalities. While these methods employ various fusion strategies across modalities, explicit modality alignment has not yet been explored in the combinatorial medication recommendation task.

### 3 Methodology

#### 3.1 Problem Formulation

Our MedAlign aims to provide an accurate and safe CMR for each patient at the  $t$ -th visit based on multimodal EHRs (shown in Figure 2). It consists of three core modules: the Distribution-aware Multimodal Medication Alignment module (DMMA), the Longitudinal Multi-view Patient Aggregation module (LMPA), and the Combinatorial Medication Recommendation module (CMR).

- In DMMA, we first generate three multimodal medication representations (i.e.,  $X^T$ ,  $X^I$ , and  $X^S$ ) based on their textual descriptions, IDs, and molecular structures, respectively. Subsequently, we adopt optimal transport to align these heterogeneous modality distributions within a unified latent space, thereby ensuring consistent medication embeddings (i.e.,  $\tilde{X}^T$ ,  $\tilde{X}^I$ , and  $\tilde{X}^S$ ).
- In LMPA, we obtain the  $t$ -th visit representations (i.e.,  $V_d^t$ ,  $V_p^t$ , and  $V_m^{t-1}$ ) of each patient with multi-view information  $v^t = [d^t, p^t, m^{t-1}]$ , where  $d^t \in \{0, 1\}^{|D|}$  and  $p^t \in \{0, 1\}^{|P|}$  are multi-hot vectors of the patient’s diagnoses and procedures at the  $t$ -th visit.  $m^{t-1} \in \{0, 1\}^{|M|}$  is a multi-hot vector of the patient’s historical medications at the  $t - 1$ -th visit.  $D$ ,  $P$ , and  $M$  denote the sets of diagnoses, procedures, and medications, respectively. Then, we aggregate these historical visit sequences into the comprehensive longitudinal patient embedding  $P^t$ .
- In CMR, we predict a medication recommendation combination  $\hat{M}^t$  for each patient at the  $t$ -th visit. Notably, we recommend a medication combination while controlling a low Drug-Drug Interaction (DDI) rate via a symmetric binary DDI adjacency matrix  $A \in \mathbb{R}^{|M| \times |M|}$ , where  $A_{ij} = 1$  represents an interaction relationship between medications  $i$  and  $j$ .

#### 3.2 Distribution-aware Multimodal Medication Alignment

To effectively integrate information from different modalities and ensure consistent medication characteristics, in this subsection, we first generate multimodal medication representations based on

their textual description, ID, and molecular structure, respectively. Next, we align these medication embeddings from a distribution alignment perspective via a multi-modality transport mechanism, thereby capturing the correlations and complementary information.

**3.2.1 Multimodal Medication Representations.** To capture effective medication characteristics and enrich medication embeddings, we utilize each medication’s textual description, ID, and molecular structure based on multimodal EHRs, obtaining the corresponding medication representations.

The Pre-trained Language Model (PLM), such as BioBERT [15], is trained on large biomedical corpora with a broad spectrum of clinical domain knowledge. Harnessing its semantic understanding capabilities and domain-specific knowledge, we encode the medication description set  $\{c_i\}_{i=1}^{|M|}$  from DrugBank [29] to generate concise medication representations in textual modality:

$$X^T = \text{MLP}_{\text{Text}} \left( \text{PLM} \left( \{c_i\}_{i=1}^{|M|} \right) \right), \quad (1)$$

where  $\text{MLP}_{\text{Text}} : \mathbb{R}^{|M| \times \text{dim}_{\text{PLM}}} \rightarrow \mathbb{R}^{|M| \times \text{dim}}$  is a multi-layer perceptron and each row of  $X^T \in \mathbb{R}^{|M| \times \text{dim}}$  denotes the textual embedding of a medication. We then initialize learnable ID embeddings of medications  $X^I \in \mathbb{R}^{|M| \times \text{dim}}$ .

Furthermore, we integrate molecular structures to learn the medication’s molecular representations. In particular, we first extract atomic and bond arrangements from SMILES strings using RD-Kit [12], constructing a molecular graph. We adopt a Graph Neural Network (GNN) to model the interactions among all atoms within the molecular structure graph  $\mathcal{G} = (\mathcal{A}, \mathcal{B})$ . Each atom  $a_i \in \mathcal{A}$ , linked by the chemical bond  $b_i \in \mathcal{B}$ , has an initial embedding  $s_{a_i,0} \in \mathbb{R}^{1 \times \text{dim}}$ . The GNN adopts message-passing and aggregation mechanisms to capture complex high-order relationships between atoms, thereby updating the embedding for atom  $a_i$ :

$$\begin{aligned} s_{a_i,l} &= \text{GNN}(s_{a_i,l-1}), \\ &= \text{UPD} \left( s_{a_i,l-1}, \text{AGG} \left( \left\{ \text{MSG} \left( s_{a_i,l-1}, s_{a_j,l-1} \right) \right\}_{a_j \in \mathcal{N}(a_i)} \right) \right), \end{aligned} \quad (2)$$

where  $s_{a_i,l}$  denotes the embedding of atom  $a_i$  in the  $l$ -th layer of the GNN and  $\mathcal{N}(a_i)$  denotes the neighbor atom set of  $a_i$ . The MSG function receives each neighbor atom’s message, the AGG function is used for aggregating the neighbor atom embeddings, and the UPD function updates the embedding of  $a_i$  based on the aggregated neighbor atom embeddings. Inspired by its effectiveness in acquiring long-range collaborative interactions between atoms [40], we employ a graph transformer as our GNN in practical implementation.

Next, we aggregate the atom embeddings  $\{s_{a_i,L}\}_{i=1}^{|\mathcal{A}|}$  in the molecular graph  $\mathcal{G}$  into a global structural representation via a mean pooling readout function:

$$S_{\mathcal{G}} = \text{Pooling} \left( \{s_{a_i,L}\}_{i=1}^{|\mathcal{A}|} \right), \quad (3)$$

where  $L$  is the total number of GNN layers. We use the same GNN with shared parameters for all  $|M|$  medication molecules and store their corresponding molecular representations as  $X^S \in \mathbb{R}^{|M| \times \text{dim}}$ .

**3.2.2 Distribution-aware Modality Alignment.** As highlighted in the Section 1, existing approaches [26, 27] often rely on feature extraction from individual modalities without explicitly aligning

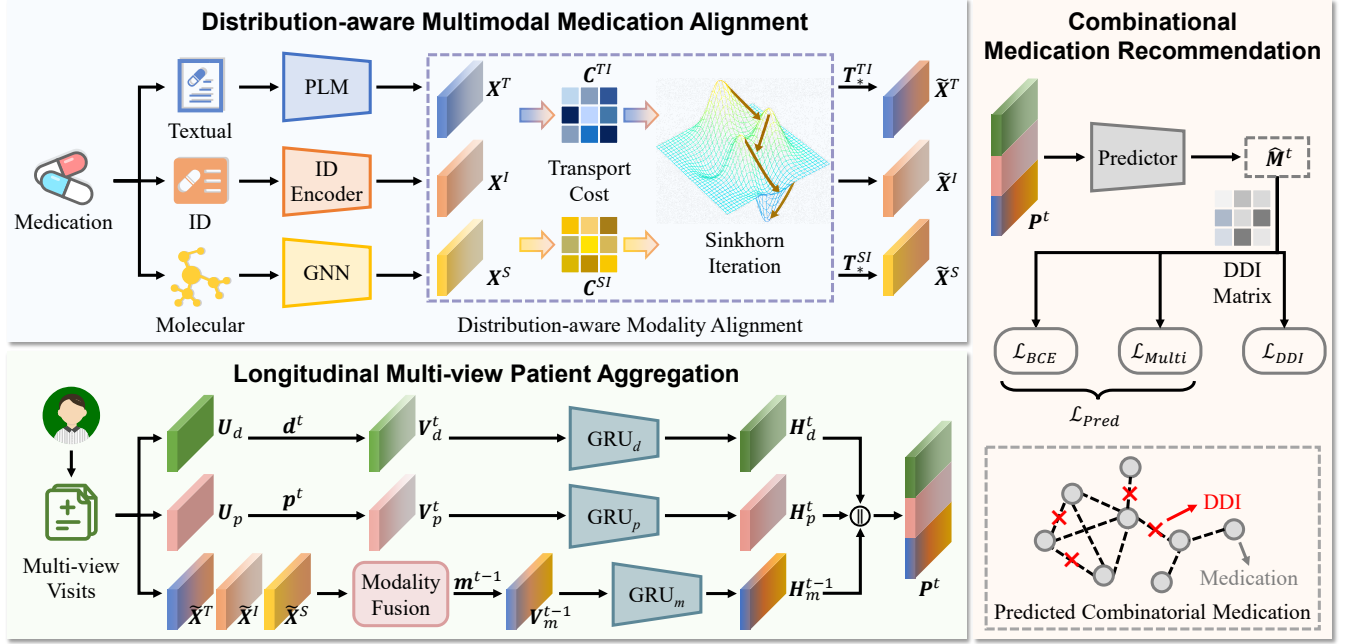


Figure 2: The overall framework of our MedAlign, where red  $\times$  indicates the deletion of the Drug-Drug Interaction (DDI) pair.

information across different data sources. Consequently, they tend to ignore complementary modality-specific information and destroy the intrinsic distributions of modalities [2, 7, 16], leading to suboptimal representations for CMR.

To address these limitations, we draw inspiration from the recent advances in Optimal Transport (OT), which have demonstrated great success in handling heterogeneous data and measuring distributional discrepancies [17, 23, 38]. Building on this, we design a multi-modality transport mechanism for medications from a distribution alignment perspective, where OT is employed to minimize the cost of transferring one modality  $m$  to another modality  $m'$ . Notably, we will form the comprehensive patient embedding with other multi-view information (e.g., diagnoses and procedures) in the latter Section 3.3, where these clinical views share the same ID modality. Therefore, we propose to align textual and structural representations of medications (i.e.,  $X^T$  and  $X^S$ ) into a unified and comparable latent space shared with medication ID embeddings  $X^I$ , ensuring consistent medication representations.

Specifically, we first calculate the transport cost via cosine distance as  $C_{ij}^{mm'} = 1 - \cos(X_i^m - X_j^{m'})$ .  $X_i^m$  and  $X_j^{m'}$  are the row vectors of modality representations  $X^m$  and  $X^{m'}$ . The objective of the above alignment can be formulated:

$$T_*^{mm'} = \arg \min_{T^{mm'} \in \mathbb{R}_+^{|M| \times |M'|}} \sum_{i=1}^{|M|} \sum_{j=1}^{|M'|} C_{ij}^{mm'} T_{ij}^{mm'}, \quad (4)$$

$$\text{s.t. } T^{mm'} \mathbf{1}_{|M'|} = \frac{1}{|M|} \mathbf{1}_{|M|}, (T^{mm'})^\top \mathbf{1}_{|M|} = \frac{1}{|M'|} \mathbf{1}_{|M'|},$$

where  $T_*^{mm'}$  is the optimal transport plan, and  $T_{ij}^{mm'}$  is the amount of information from  $X_i^m$  to  $X_j^{m'}$ .  $\mathbf{1}_{|M'|}$  and  $\mathbf{1}_{|M|}$  are all-ones vectors of dimension  $|M'|$  and  $|M|$ , respectively. We adopt the Sinkhorn [6] algorithm to accelerate the calculation of the optimal

transport problem. Then, we update modality  $m$  representations  $\tilde{X}^m = (T_*^{mm'})^\top X^m$ , aligning distributions between modalities  $m$  and  $m'$ . Consequently,  $\tilde{X}^T = (T_*^{TI})^\top X^T$  and  $\tilde{X}^S = (T_*^{SI})^\top X^S$ .

### 3.3 Longitudinal Multi-view Patient Aggregation

In this subsection, to fully model patient health conditions and provide the basis for the personalized CMR, we leverage multi-view information in multimodal EHRs, obtaining various visit representations of each patient. Subsequently, incorporating the rich temporal sequence data, we form the longitudinal patient representations for medication combination prediction.

**3.3.1 Multi-view Visit Representations.** To learn comprehensive patient health conditions, we encode the visit representations of each patient using multi-view information from diagnoses, procedures, and historical medications in multimodal EHRs, respectively.

We first initial two learnable embedding tables,  $U_d \in \mathbb{R}^{|\mathcal{D}| \times \dim}$  and  $U_p \in \mathbb{R}^{|\mathcal{P}| \times \dim}$ . Given the multi-hot diagnosis and procedure vectors at the  $t$ -th visit  $d^t, p^t$ , we pick out the corresponding diagnosis and procedure embeddings and sum them up via vector-matrix multiplication, generating the visit representations from two views:

$$V_d^t = d^t U_d, \quad (5)$$

$$V_p^t = p^t U_p. \quad (6)$$

To generate the visit representations from historical medication view, we fuse the aligned medication features (cf., Section 3.2.2) from three modalities,  $\tilde{X}^T, X^I$  (i.e.,  $\tilde{X}^I$ ), and  $\tilde{X}^S$  by a cross-attention mechanism. We denote the aggregated representations as  $M^m = W_v^m \tilde{X}^m$ , where  $m \in \mathcal{O} = \{T, I, S\}$ . The learnable weights  $W_v^m \in$

$\mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}$  are computed by:

$$\begin{aligned} Q &= H_q W_q, K = H_k W_k, \\ \{\mathbf{W}_v^m\}_{m \in \mathcal{O}} &= \text{softmax}\left(QK^\top / \sqrt{\dim}\right), \end{aligned} \quad (7)$$

where  $H_q \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{M}| \times \dim}$  is the matrix stacked by the embeddings in  $\{\tilde{X}^T, \tilde{X}^I, \tilde{X}^S\}$ ,  $H_k = \frac{1}{|\mathcal{O}|} \sum_{m \in \mathcal{O}} \tilde{X}^m \in \mathbb{R}^{|\mathcal{M}| \times \dim}$ ,  $W_q, W_k \in \mathbb{R}^{\dim \times \dim}$  are the learnable weights. Subsequently, we obtain the fused medication representations  $\mathbf{M} = \frac{1}{|\mathcal{O}|} \sum_{m \in \mathcal{O}} \mathbf{M}^m \in \mathbb{R}^{|\mathcal{M}| \times \dim}$ , which can effectively enhance and capture inter-modality correlations and complementary information. Similar to the visit representations from diagnosis and procedure views, the visit embedding from the historical medication view is calculated as follows:

$$\mathbf{V}_m^{t-1} = \mathbf{m}^{t-1} \mathbf{M}, \quad (8)$$

where  $\mathbf{m}^{t-1}$  is a multi-hot vector of the patient's historical medications at the  $t-1$ -th visit. The multi-view visit representation at  $t$ -th visit is defined as:

$$\mathbf{V}^t = \left\{ \mathbf{V}_d^t, \mathbf{V}_p^t, \mathbf{V}_m^{t-1} \right\}. \quad (9)$$

**3.3.2 Longitudinal Patient Representations.** To further leverage the rich temporal sequence data, we utilize three separate Gated Recurrent Units (GRUs) to capture the patient's historical diagnosis, procedure, and medication information:

$$\mathbf{H}_d^t = \text{GRU}_d \left( \{\mathbf{V}_d^i\}_{i=1}^t \right), \quad (10)$$

$$\mathbf{H}_p^t = \text{GRU}_p \left( \{\mathbf{V}_p^i\}_{i=1}^t \right), \quad (11)$$

$$\mathbf{H}_m^{t-1} = \text{GRU}_m \left( \{\mathbf{V}_m^i\}_{i=1}^{t-1} \right). \quad (12)$$

Next, these three obtained embeddings are concatenated to form the final patient embedding  $\mathbf{P}^t = [\mathbf{H}_d^t \| \mathbf{H}_p^t \| \mathbf{H}_m^{t-1}] \in \mathbb{R}^{1 \times 3\dim}$ .

### 3.4 Combinational Medication Recommendation

Given the final patient representation  $\mathbf{P}^t$ , we predict a medication recommendation combination for the patient at the  $t$ -th visit:

$$\hat{\mathbf{M}}^t = \text{MLP}_{\text{Rec}}(\mathbf{P}^t), \quad (13)$$

where  $\text{MLP}_{\text{Rec}} : \mathbb{R}^{1 \times 3\dim} \rightarrow \mathbb{R}^{1 \times |\mathcal{M}|}$  is a multi-layer perceptron. We train our MedAlign using two different loss functions, balancing between the accuracy and safety of CMR. The formulations of two loss functions are summarized as follows:

**3.4.1 Multi-Label Prediction Loss.** We treat the CMR task as a multi-label classification task and employ two commonly adopted loss functions to enhance the robustness of the result [24, 35, 36], i.e., binary cross-entropy loss and multi-label margin loss, which can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{BCE}} &= - \sum_{i=1}^{|\mathcal{M}|} (\mathbf{O}_i^t \log \hat{\mathbf{M}}^t + (1 - \mathbf{O}_i^t) \log (1 - \hat{\mathbf{M}}^t)), \\ \mathcal{L}_{\text{Multi}} &= \sum_{\{i | \mathbf{O}_i^t = 1\}} \sum_{\{j | \mathbf{O}_j^t = 0\}} \frac{\max\{1 - (\hat{\mathbf{M}}_i^t - \hat{\mathbf{M}}_j^t), 0\}}{|\mathcal{M}|}, \\ \mathcal{L}_{\text{Pred}} &= \lambda \mathcal{L}_{\text{BCE}} + (1 - \lambda) \mathcal{L}_{\text{Multi}}, \end{aligned} \quad (14)$$

where  $\{\mathbf{O}_i^t\}_{i=1}^{|\mathcal{M}|}$  is ground-truth medications for patient's  $t$ -th visit. Hyperparameter  $\lambda$  is experimentally set to 0.95 by default. Notably,

**Table 1: Statistics of the datasets used in our experiments.**

Dataset	MIMIC-III	MIMIC-IV
# of patients	6,350	61,264
# of visits	15,031	163,877
# of diagnoses	1,903	2,000
# of procedures	1,409	11,056
# of medications	131	131
Avg. # of visits	2.3671	2.6749
Avg. # of diagnoses per visit	10.2266	8.2343
Avg. # of procedures per visit	3.8244	2.3579
Avg. # of medications per visit	11.4361	6.5055

the multi-label margin loss  $\mathcal{L}_{\text{Multi}}$  ensures that true labels have at least 1 margin larger than others, leading to more stable predictions.

**3.4.2 DDI Loss.** To achieve a lower DDI rate in the predicted medication combinations, we minimize the following loss:

$$\mathcal{L}_{\text{DDI}} = \sum_{i=1}^{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} \left( \hat{\mathbf{M}}_i^t \left( \hat{\mathbf{M}}_j^t \right)^\top \right) \cdot \mathbf{A}_{ij}, \quad (15)$$

where  $\hat{\mathbf{M}}_i^t \left( \hat{\mathbf{M}}_j^t \right)^\top$  denotes the pair-wise DDI probability.

During the training process, the accuracy and DDI rate often increase together. This is because DDI is common in real-world EHR data, and correct or incorrect combinatorial medication predictions may raise the DDI rate. Balancing the model's accuracy and safety is thus crucial for effective CMR. Inspired by [36, 40], we introduce a dynamic weighting strategy to form the final objective function of our MedAlign as follows:

$$\mathcal{L}_{\text{Final}} = \begin{cases} \eta \mathcal{L}_{\text{Pred}} + (1 - \eta) \mathcal{L}_{\text{DDI}}, & \text{DDI rate} \geq \epsilon, \\ \mathcal{L}_{\text{Pred}}, & \text{DDI rate} < \epsilon, \end{cases} \quad (16)$$

where  $\eta = \min\{\tanh(\tau \frac{\epsilon}{\text{DDI rate} - \epsilon}), 1\}$  and  $\epsilon$  is the preset safe DDI threshold. Hyperparameters  $\tau$  and  $\epsilon$  are experimentally set to 0.08 and 0.06, respectively.

### 3.5 Complexity Analysis

We primarily analyze the time and space complexity associated with modality alignment and fusion for multimodal medication representations (cf., Section 3.2.2 and Section 3.3.1).

**3.5.1 Time Complexity.** The time complexity of the proposed modality alignment (Eq. (4)) is  $O(\delta \cdot |\mathcal{M}|^2)$ , where  $\delta$  the number of Sinkhorn iterations used in optimal transport and  $|\mathcal{M}|$  is the total number of medications. The complexity of the modality fusion with cross-attention mechanism (Eq. (7)) is  $O(|\mathcal{M}|^2 \cdot \dim)$ , where  $\dim$  is the dimension of medication embeddings.

**3.5.2 Space Complexity.** The matrices in both modality alignment and fusion (Eq. (4) and Eq. (7)) cost  $O(|\mathcal{M}|^2)$  space, where  $|\mathcal{M}|$  is the total number of medications.

## 4 Experiments

In this section, we evaluate our proposed MedAlign framework focusing on the following four key research questions:

- **RQ1:** How does MedAlign perform in comparison to state-of-the-art baselines for combinatorial medication recommendation?

- **RQ2:** What are the effects of the modality alignment and modality fusion components?
- **RQ3:** How do the hyperparameters affect the recommendation performance, and how to choose optimal values?
- **RQ4:** How does MedAlign align and fuse multimodal medication distributions to improve the recommendation accuracy?

## 4.1 Experimental Settings

**4.1.1 Datasets and Evaluation Protocols.** We use two real-world EHR datasets to verify the effectiveness of compared methods, i.e., **MIMIC-III** [9] and **MIMIC-IV** [10]. Both datasets are fully anonymized and carefully sanitized before our access. Following [36, 40], we chose patients who made at least two visits for both datasets and the ATC third-level code as the target label. The statistics are summarized in Table 1. For evaluation metrics, we use Jaccard Similarity Score (Jaccard), Average F1 Score (F1), Precision Recall AUC (PRAUC), Drug-Drug Interaction Rate (DDI), and Average Number of Medications (Avg. # of Med) indicating how well the model aligns with real-world prescribing patterns, which are consistent with [24, 36, 40].

**4.1.2 Methods for Comparison.** To comprehensively evaluate our proposed MedAlign, we adopt 12 representative state-of-the-art methods as baselines for the performance comparison:

### (1) Instance-based methods:

- **LR** [4] is a traditional recommendation method that incorporates L1 regularization. We train a separate binary classifier for each label in multi-label classification.
- **ECC** [22] applies boosting-based ensemble learning for multi-label classification.
- **LEAP** [39] formulates MR to sequential decision-making, with a recurrent decoder to model label dependencies and content-based attention for label instance mapping.

### (2) Longitudinal methods:

- **RETAIN** [5] predicts a patient’s future medication combinations based on their historical visit records, utilizing RNN and reverse time attention mechanism.
- **MICRON** [34] utilizes a recurrent residual learning model to capture drug and disease changes between consecutive visits, enabling incremental learning of new patient characteristics.
- **GAMENet** [24] leverages memory neural networks and DDI conflict relations to capture historical medication data for improving medication recommendations.
- **COGNet** [32] retrieves patients’ historical diagnoses/drugs and mines their relationship with the current diagnosis, which is embedded in the Transformer as a plug-in.
- **LAMRec** [27] is a label-aware multi-view medication recommendation model that integrates cross-attention and contrastive learning to enhance patient representation and label utilization.

### (3) Modality-aware methods:

- **SafeDrug** [35] encodes the molecular structure information of medications for enhancing the accuracy and safety of medication recommendations.
- **DrugRec** [25] extracts all the drug representations with the molecule pre-trained transformer model and designs a causal inference-based drug recommendation model.

- **MoleRec** [36] improves drug recommendation by leveraging molecular substructure interactions and patient-substructure relevance to identify efficacy-driving substructures.
- **DEPOT** [40] is a state-of-the-art medication recommendation framework that decomposes drug molecules into semantic motif trees and models interactions among these motifs.

**4.1.3 Implementation Details.** We split training, validation, and test sets by 2/3, 1/6, and 1/6, consistent with [35, 36, 40]. We optimize the compared baselines with standard Adam and tune all hyperparameters on training sets through grid search. In particular, we set  $\lambda = 0.95$ ,  $\tau = 0.08$ , and  $\epsilon = 0.06$  by default. The dimension of embeddings obtained from PLM  $dim_{PLM}$  is 768. For the adopted GNN architecture, we use 2 layers with a hidden dimension of 64. The learning rate is set to  $5e-4$ . We set the embedding dimension  $dim$  as 64 and the batch size as 32 for all compared methods on both datasets. We carefully tune the hyperparameters of baselines as suggested in the original papers to achieve their best performance. All experiments are performed with two NVIDIA GTX 3090 Ti GPUs. The full code for this work is available<sup>1</sup>.

## 4.2 Overall Performance Comparison (RQ1)

We compare the combinatorial medication recommendation results of the proposed MedAlign framework to those of the competitive baseline models. Table 2 shows the Jaccard, F1, PRAUC, DDI, and Avg. # of Med on both MIMIC-III and MIMIC-IV datasets. We have the following observations:

Overall, our MedAlign outperforms all compared baselines across all recommendation evaluation metrics (i.e., Jaccard, F1, and PRAUC) on both datasets. This answers RQ1, showing that our proposed recommendation framework with multi-modality alignment is capable of precise medication combinations. Compared with the second-best performance (e.g., DEPOT), the performance gains of MedAlign achieve up to 2.02% with Jaccard on MIMIC-IV.

Instance-based methods are traditional approaches that focus on binary classification for each medication label, which are often limited by their inability to capture the complex relationships between multiple modalities of patient and medication data. Therefore, MedAlign significantly outperforms them across all metrics, achieving 11.81% on average. Longitudinal methods incorporate sequential patient records to track historical medication and diagnoses over time. However, they still struggle with modeling interactions across diverse modalities. MedAlign outperforms these models by better capturing both temporal and inter-modality correlations, with significant gains in both Jaccard (10.54% on MIMIC-III) and F1 (6.93% on MIMIC-III).

Among Modality-aware methods, DEPOT achieves superior performance over most baselines through modeling the molecular structures of medications, highlighting the significance of medication characteristics modeling. This suggests that multi-modality alignment provides an effective solution for tackling complex clinical scenarios. Compared with DEPOT, MedAlign captures inter-modality correlations and complementary modality-specific information, further enhancing the accuracy of recommendations.

<sup>1</sup><https://github.com/lvhangkenn/MedAlign>

**Table 2: Experimental results on MIMIC-III and MIMIC-IV, datasets, where \* denotes a significant improvement according to the Wilcoxon signed-rank test [30]. The best performances are highlighted in boldface and the second runners are underlined. Ground-truth Avg. # of Med in the test sets of both datasets is 19.7937 and 11.9788, respectively.**

Dataset	Method	Jaccard $\uparrow$	F1-score $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$	Avg. # of Med
MIMIC-III	LR	0.4647 $\pm$ 0.0021	0.6265 $\pm$ 0.0025	0.7472 $\pm$ 0.0022	0.0809 $\pm$ 0.0011	15.8472 $\pm$ 0.1836
	ECC	0.4556 $\pm$ 0.0027	0.6167 $\pm$ 0.0019	0.7192 $\pm$ 0.0023	0.0818 $\pm$ 0.0009	15.4763 $\pm$ 0.2174
	LEAP	0.4328 $\pm$ 0.0019	0.5986 $\pm$ 0.0022	0.6465 $\pm$ 0.0026	0.0761 $\pm$ 0.0012	17.8219 $\pm$ 0.1725
	RETAIN	0.4537 $\pm$ 0.0023	0.6174 $\pm$ 0.0021	0.7219 $\pm$ 0.0024	0.0846 $\pm$ 0.0005	19.9605 $\pm$ 0.2139
	MICRON	0.4819 $\pm$ 0.0014	0.6403 $\pm$ 0.0019	0.7297 $\pm$ 0.0019	0.0754 $\pm$ 0.0013	18.7138 $\pm$ 0.2017
	GAMENet	0.4783 $\pm$ 0.0024	0.6379 $\pm$ 0.0026	0.7248 $\pm$ 0.0027	0.0852 $\pm$ 0.0003	24.2536 $\pm$ 0.1971
	COGNet	0.4873 $\pm$ 0.0026	0.6472 $\pm$ 0.0028	0.7295 $\pm$ 0.0025	0.0836 $\pm$ 0.0004	22.3725 $\pm$ 0.2014
	LAMRec	0.4915 $\pm$ 0.0024	0.6504 $\pm$ 0.0020	0.7699 $\pm$ 0.0028	0.0803 $\pm$ 0.0002	21.4461 $\pm$ 0.1537
	SafeDrug	0.4832 $\pm$ 0.0021	0.6434 $\pm$ 0.0023	0.7261 $\pm$ 0.0028	<b>0.0691 <math>\pm</math> 0.0007*</b>	18.0294 $\pm$ 0.1923
	DrugRec	0.4956 $\pm$ 0.0018	0.6536 $\pm$ 0.0021	0.7332 $\pm$ 0.0019	0.0739 $\pm$ 0.0008	17.8413 $\pm$ 0.2658
	MoleRec	0.5303 $\pm$ 0.0029	0.6847 $\pm$ 0.0023	0.7775 $\pm$ 0.0024	0.0734 $\pm$ 0.0006	20.9342 $\pm$ 0.1743
	DEPOT	<u>0.5331 <math>\pm</math> 0.0022</u>	<u>0.6868 <math>\pm</math> 0.0025</u>	<u>0.7820 <math>\pm</math> 0.0021</u>	0.0722 $\pm$ 0.0003	19.6914 $\pm$ 0.1432
	MedAlign	<b>0.5433 <math>\pm</math> 0.0019*</b>	<b>0.6955 <math>\pm</math> 0.0023*</b>	<b>0.7869 <math>\pm</math> 0.0024*</b>	<u>0.0715 <math>\pm</math> 0.0002</u>	20.7513 $\pm$ 0.1968
MIMIC-IV	LR	0.4101 $\pm$ 0.0025	0.5589 $\pm$ 0.0024	0.6717 $\pm$ 0.0019	0.0776 $\pm$ 0.0007	8.7963 $\pm$ 0.2195
	ECC	0.3927 $\pm$ 0.0022	0.5374 $\pm$ 0.0027	0.6691 $\pm$ 0.0023	0.0783 $\pm$ 0.0004	7.8924 $\pm$ 0.2013
	LEAP	0.3898 $\pm$ 0.0021	0.5405 $\pm$ 0.0019	0.5436 $\pm$ 0.0021	0.0728 $\pm$ 0.0009	9.7836 $\pm$ 0.1961
	RETAIN	0.4164 $\pm$ 0.0024	0.5687 $\pm$ 0.0018	0.6712 $\pm$ 0.0025	0.0809 $\pm$ 0.0002	10.5347 $\pm$ 0.1879
	MICRON	0.4277 $\pm$ 0.0017	0.5768 $\pm$ 0.0021	0.6733 $\pm$ 0.0029	0.0719 $\pm$ 0.0006	11.6973 $\pm$ 0.1636
	GAMENet	0.4219 $\pm$ 0.0023	0.5745 $\pm$ 0.0022	0.6629 $\pm$ 0.0020	0.0817 $\pm$ 0.0002	15.5426 $\pm$ 0.1784
	COGNet	0.4347 $\pm$ 0.0018	0.5884 $\pm$ 0.0025	0.6517 $\pm$ 0.0024	0.0802 $\pm$ 0.0003	13.2374 $\pm$ 0.1589
	LAMRec	0.4409 $\pm$ 0.0024	0.5912 $\pm$ 0.0025	0.7075 $\pm$ 0.0027	0.0796 $\pm$ 0.0005	12.3147 $\pm$ 0.1765
	SafeDrug	0.4357 $\pm$ 0.0019	0.5873 $\pm$ 0.0026	0.6543 $\pm$ 0.0025	<b>0.0663 <math>\pm</math> 0.0004*</b>	10.8532 $\pm$ 0.2132
	DrugRec	0.4473 $\pm$ 0.0025	0.5938 $\pm$ 0.0023	0.6704 $\pm$ 0.0019	0.0675 $\pm$ 0.0008	10.5933 $\pm$ 0.2058
	MoleRec	0.4674 $\pm$ 0.0022	0.6194 $\pm$ 0.0019	0.7067 $\pm$ 0.0022	0.0694 $\pm$ 0.0007	11.8511 $\pm$ 0.2012
	DEPOT	<u>0.4709 <math>\pm</math> 0.0021</u>	<u>0.6223 <math>\pm</math> 0.0024</u>	<u>0.7111 <math>\pm</math> 0.0026</u>	0.0702 $\pm$ 0.0008	12.0751 $\pm$ 0.2278
	MedAlign	<b>0.4804 <math>\pm</math> 0.0022*</b>	<b>0.6313 <math>\pm</math> 0.0020*</b>	<b>0.7223 <math>\pm</math> 0.0027*</b>	<u>0.0671 <math>\pm</math> 0.0006</u>	12.8769 $\pm$ 0.1264

**Table 3: Ablation studies on MIMIC-III and MIMIC-IV, where Avg. denotes the average number of medications.**

Method	Jaccard $\uparrow$	F1 $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$	Avg.
MIMIC-III					
MedAlign	<b>0.5433</b>	<b>0.6955</b>	<b>0.7869</b>	<b>0.0715</b>	20.7513
- w/o. MA	0.5237	0.6695	0.7480	0.0753	22.2475
- w/o. MF	0.5376	0.6865	0.7749	0.0737	21.1125
MIMIC-IV					
MedAlign	<b>0.4804</b>	<b>0.6313</b>	<b>0.7223</b>	<b>0.0671</b>	12.8769
- w/o. MA	0.4531	0.5921	0.6983	0.0714	14.4574
- w/o. MF	0.4694	0.6204	0.7173	0.0691	13.6971

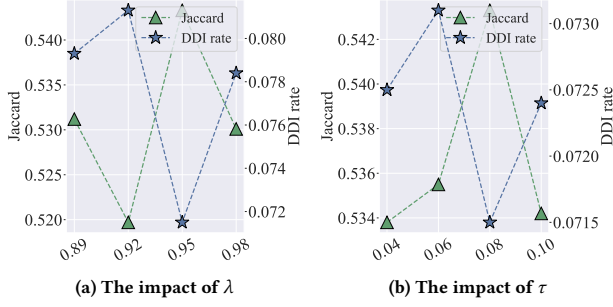
Regarding the safety evaluation metric DDI, SafeDrug leads with its DDI-controllable loss function. However, our method demonstrates significant improvements over SafeDrug in all accuracy metrics (achieved up to 12.44% in Jaccard on MIMIC-III), while keeping the DDI rate only 0.0016 higher than SafeDrug on average. As discussed in Section 3.4.2, DDI is common in real-world EHR data. Pursuing the lowest DDI rate without considering efficacy may lead to suboptimal medication combinations [20]. Therefore, this performance trade-off reflects a practical balance between prediction accuracy and safety, with the slight increase in DDIs offset by substantial gains in recommendation effectiveness.

### 4.3 Ablation Studies (RQ2)

To better understand the contribution of our proposed modality alignment (cf., Eq. (4) in Section 3.2.2) and modality fusion (cf., Eq. (7) in Section 3.3.1) components, we conduct ablation studies as follows: MedAlign w/o. MA is our MedAlign without the modality alignment strategy and directly applies the modality fusion mechanism; MedAlign w/o. MF is our MedAlign without the modality fusion mechanism and directly averages medication embeddings after the modality alignment strategy.

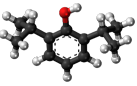
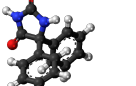
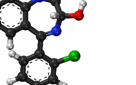
As shown in Table 3, compared with MedAlign w/o. MA, our MedAlign leads to performance gains ranging from 3.44% (achieved in PRAUC on MIMIC-IV) to 6.62% (achieved in F1 on MIMIC-IV). Such results show the effectiveness of aligning multi-modal medication representations via optimal transport from a distribution perspective. Additionally, MedAlign outperforms MedAlign w/o. MF ranging from 0.70% in PRAUC on MIMIC-IV to 2.99% in DDI on MIMIC-III. This indicates that the modality fusion based on the cross-attention mechanism can effectively integrate information between diverse modalities, modeling more consistent medication characteristics. These results also affirm that our proposed components for multimodal medications capture inter-modality correlations and complementary information.





**Figure 3: Performance regarding Jaccard and DDI rate with varying the weight of  $\mathcal{L}_{Pred}$  (i.e.,  $\lambda$  in Eq. (14)) and the weight factor of  $\mathcal{L}_{DDI}$  (i.e.,  $\tau$  in Eq. (15)) on MIMIC-III.**

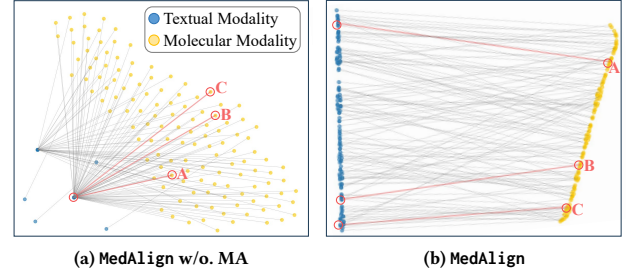
**Table 4: Multimodal information of three example medications on the MIMIC-III dataset. The textual descriptions provide clinical indications, while the molecular structures capture pharmacological properties, highlighting the complementary information among multimodal data.**

Medication	A	B	C
Textual Description	<i>Propofol</i> manages status epilepticus and maintains anesthesia.	<i>Phenytoin</i> presents and controls various types of seizures.	<i>Lorazepam</i> treats panic disorders, severe anxiety, and seizures.
Molecular Structure			

#### 4.4 Hyperparameter Studies (RQ3)

To verify the impact of two main hyperparameters  $\lambda$  and  $\tau$  introduced by our MedAlign for CMR, which collectively optimize model performance, we conduct hyperparameter studies on the MIMIC-III dataset. Particularly,  $\lambda$  in Eq. (14) controls the weight between the hybrid BCE and multi-label margin loss.  $\tau$  in Eq. (15) adjusts the nonlinearity of the tanh function, which regulates the trade-off between DDI and recommendation loss, balancing safety and accuracy. We analyze their impact on model performance and provide insights into optimal parameter selection.

As shown in Figure 3(a), by adjusting the weight  $\lambda$ , MedAlign can fine-tune the separation between correct and incorrect medication combination predictions. A too large  $\lambda$  decreases the impact of  $\mathcal{L}_{Multi}$ , restricting medication recommendations and degrading performance, while an excessively small  $\lambda$  induces instability. Following [35, 36, 40], empirical results suggest that  $\lambda = 0.95$  is optimal, ensuring that both loss components contribute effectively to the model’s convergence. As described in Section 3.4.2, the adopted dynamic weighting strategy is essential for modulating the penalty on unsafe medication combinations based on the current DDI rate. A smaller  $\tau$  results in a steeper penalty curve, sharply prioritizing DDI avoidance, while a larger  $\tau$  smooths the curve and lessens the emphasis on safety. Figure 3(b) demonstrates that  $\tau = 0.08$  achieves the



**Figure 4: Visualizations of textual and molecular medication embedding distributions on MIMIC-III learned by different models. Best viewed in color.**

best balance between maintaining high recommendation accuracy and enforcing medication safety.

#### 4.5 Case Studies (RQ4)

To highlight the advantages of MedAlign in integrating multimodal medications for recommendations, we provide three example medications from MIMIC-III. Their detailed multimodal information is presented in Table 4. Additionally, Figure 4 visualizes the learned embedding distributions in textual and molecular modalities by MedAlign w/o. MA and MedAlign, respectively.

As shown in Figure 4(a), the textual representations of medications A, B, and C significantly overlap, making them indistinguishable when relying solely on textual features. This issue arises from their similar textual indications, highlighting the inability of MedAlign w/o. MA to align information effectively across different modalities, as it depends only on the cross-attention mechanism.

In contrast, Figure 4(b) demonstrates the effectiveness of our MedAlign in capturing both inter-modality correlations and complementary modality-specific information. For example, MedAlign successfully distinguishes the unique roles of medications A *Propofol* (for maintaining *general anesthesia*) and C *Lorazepam* (for addressing *panic disorders*) by leveraging both textual and molecular features. Moreover, MedAlign captures accurate relationships between medications, as reflected by the closer proximity of embeddings for medications B and C across the two modalities. This corresponds to their similar pharmacological actions (e.g., controlling various types of *seizures*). These results showcase MedAlign’s ability to precisely align modalities and represent medications.

## 5 Conclusion

In this paper, we propose MedAlign, a novel framework for combinatorial medication recommendation using multimodal electronic health records. Particularly, we design a distribution-aware multimodal medication alignment strategy with optimal transport. This effectively captures inter-modality correlations and complementary information from different modalities, learning consistent and accurate medication representations. Furthermore, through aggregating the rich temporal visit sequences with multi-view information from diagnoses, procedures, and fused historical medications, we form longitudinal patient representations for medication combination prediction. Comprehensive experiments show that our MedAlign achieves significant improvements over state-of-the-art methods.



## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (No.62302098), Fujian Provincial Natural Science Foundation of China under Grants (2025J01540), Zhejiang Provincial Natural Science Foundation of China under Grants (LQ23F020007), and Zhejiang Provincial Department of Agriculture and Rural Affairs Project under Grants (2024SNJF044). Carl Yang was not supported by any fund from China.

## References

- [1] Phillip Aitken, Ioana Stanescu, Rebecca Playne, Jennifer Zhang, Christopher MA Frampton, and Hartley C Atkinson. 2019. An integrated safety analysis of combined acetaminophen and ibuprofen (Maxigesic®/Combogesic®) in adults. *J Pain Res* (2019), 621–634.
- [2] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in neural information processing systems* 35 (2022), 39090–39102.
- [3] Qiuhui Chen and Yi Hong. 2024. SMART: Self-Weighted Multimodal Fusion for Diagnostics of Neurodegenerative Disorders. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4426–4435.
- [4] Weiwei Cheng and Eyke Hüllermeier. 2009. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning* 76 (2009), 211–225.
- [5] 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).
- [6] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*. 2292–2300.
- [7] Hao Geng, Deqing Wang, Fuzhen Zhuang, Xuehua Ming, Chenguang Du, Ting Jiang, Haolong Guo, and Rui Liu. 2022. Modeling dynamic heterogeneous graph and node importance for future citation prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 572–581.
- [8] Fan Gong, Meng Wang, Haofen Wang, Sen Wang, and Mengyue Liu. 2021. SMR: medical knowledge graph embedding for safe medicine recommendation. *Big Data Research* 23 (2021), 100174.
- [9] 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.
- [10] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. 2018. The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association* 25, 1 (2018), 32–39.
- [11] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2015. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence* 38, 1 (2015), 188–194.
- [12] Greg Landrum et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8, 31.10 (2013), 5281.
- [13] MJS Langman, J Weil, P Wainwright, DH Lawson, MD Rawlins, RFA Logan, M Murphy, MP Vessey, and DG Colin-Jones. 1994. Risks of bleeding peptic ulcer associated with individual non-steroidal anti-inflammatory drugs. *The Lancet* 343, 8905 (1994), 1075–1078.
- [14] Hung Le, Truyen Tran, and Svetha Venkatesh. 2018. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1637–1645.
- [15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
- [16] Hui Li and Xiao-Jun Wu. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion* 103 (2024), 102147.
- [17] Zhenghong Lin, Yanchao Tan, Yunfei Zhan, Weiming Liu, Fan Wang, Chaochao Chen, Shiping Wang, and Carl Yang. 2023. Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6234–6242.
- [18] Yang Liu, Zhonglei Gu, Tobey H Ko, and Jiming Liu. 2018. Identifying key opinion leaders in social media via modality-consistent harmonized discriminant embedding. *IEEE transactions on cybernetics* 50, 2 (2018), 717–728.
- [19] Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, Zikun Nie, Hao Zhou, and Zaiqing Nie. 2024. Learning multi-view molecular representations with structured and unstructured knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2082–2093.
- [20] Shi Mu, Chen Li, Xiang Li, and Shunpan Liang. 2025. Medication recommendation via dual molecular modalities and multi-step enhancement. *Expert Systems with Applications* (2025), 127163.
- [21] Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Li Guo, and Xian Yang. 2024. EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion* 102 (2024), 102069.
- [22] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning* 85 (2011), 333–359.
- [23] Kyeongha Rho, Hyeongkeun Lee, Valentio Iverson, and Joon Son Chung. 2025. LAVCap: LLM-based Audio-Visual Captioning using Optimal Transport. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*.
- [24] 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.
- [25] Hongda Sun, Shufang Xie, Shuqi Li, Yuhao Chen, Ji-Rong Wen, and Rui Yan. 2022. Debaised, longitudinal and coordinated drug recommendation through multi-visit clinic records. *Advances in Neural Information Processing Systems* 35 (2022), 27837–27849.
- [26] Jie Tan, Yu Rong, Kangfei Zhao, Tian Bian, Tingyang Xu, Junzhou Huang, Hong Cheng, and Helen Meng. 2024. Natural Language-Assisted Multi-modal Medication Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2200–2209.
- [27] 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*. 2230–2239.
- [28] Phan Nguyen Minh Thao, Cong-Tinh Dao, Chenwei Wu, Jian-Zhe Wang, Shun Liu, Jun-En Ding, David Restrepo, Feng Liu, Fang-Ming Hung, and Wen-Chih Peng. 2024. Medfuse: Multimodal ehr data fusion with masked lab-test modeling and large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3974–3978.
- [29] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* 46, D1 (2018), D1074–D1082.
- [30] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.
- [31] Jialun Wu, Kai He, Rui Mao, Xuequn Shang, and Erik Cambria. 2025. Harnessing the potential of multimodal EHR data: A comprehensive survey of clinical predictive modeling for intelligent healthcare. *Information Fusion* (2025), 103283.
- [32] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM web conference 2022*. 935–945.
- [33] Muhao Xu, Zhenfeng Zhu, Youru Li, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. 2024. FlexCare: Leveraging Cross-Task Synergy for Flexible Multimodal Healthcare Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3610–3620.
- [34] Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. Change Matters: Medication Change Prediction with Recurrent Residual Networks. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*. International Joint Conferences on Artificial Intelligence, 3728–3734.
- [35] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*. International Joint Conferences on Artificial Intelligence, 3735–3741.
- [36] 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.
- [37] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. 2024. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 16416–16424.
- [38] Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Qing Liu, and Shiyao Yan. 2023. TOT: topology-aware optimal transport for multimodal hate detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4884–4892.
- [39] 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.
- [40] Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li. 2024. Enhancing Precision Drug Recommendations via In-Depth Exploration of Motif Relationships. *IEEE Transactions on Knowledge and Data Engineering* (2024).

## Appendix

### A Efficiency Analysis

We evaluate both training and inference efficiency of different methods on the MIMIC-III dataset. As shown in Table 5, although MedAlign incurs a higher computational cost, MedAlign achieves the highest Jaccard score (0.5433) and the lowest DDI rate (0.0715) among all compared methods, indicating superior recommendation accuracy and safety. This reasonable trade-off demonstrates that MedAlign not only advances predictive quality but also maintains computational efficiency, making it well-suited for real-world clinical deployment.

### B Generalization Analysis

To assess the robustness and generalizability of MedAlign, we analyze its performance under varying PLMs and GNNs, different alignment mechanisms, and different modality absence, focusing on the MIMIC-III dataset.

**Various PLMs and GNNs.** We utilize different PLMs and GNNs as textual and molecular encoders, respectively. Table 6 shows that MedAlign consistently maintains superior performance, demonstrating its robustness and flexibility.

**Different Alignment Mechanisms.** To assess the adaptability of MedAlign using different alignment strategies, we replace our adopted Optimal Transport and Cross-Attention (OT+CA) mechanisms with MvDA-VC [11]. Table 7 shows that regardless of the specific alignment component used, MedAlign consistently benefits from modality alignment and exhibits strong adaptability.

**Different Modality Absence Settings.** To examine the performance of the MedAlign under different modality absence settings, we randomly replace 5% of medication representations in the molecular or textual modalities with random embeddings, respectively. Table 8 shows that MedAlign consistently maintains superior performance despite the partial absence of modalities. These results highlight that our multi-modality alignment and fusion mechanisms effectively capture consistent and complementary information across modalities, thereby improving robustness under incomplete EHR scenarios.

**Table 5: Comparison of training and inference time (s/epoch) for various methods on MIMIC-III.**

Method	Jaccard $\uparrow$	DDI $\downarrow$	Training Time	Inference Time
MoleRec	0.5303	0.0734	2,076	340
DEPOT	0.5331	0.0722	2,137	384
MedAlign	<b>0.5433</b>	<b>0.0715</b>	2,844	477

**Table 6: Ablation results of our proposed MedAlign with various PLMs and GNNs on MIMIC-III.**

Method	Jaccard $\uparrow$	F1 $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$
	MedAlign			
BioBERT	<b>0.5433</b>	<b>0.6955</b>	<b>0.7869</b>	<b>0.0715</b>
ClinicalBERT	0.5362	0.6897	0.7826	0.0723
PubMedBERT	0.5359	0.6893	0.7825	0.0721
SciBERT	0.5354	0.6889	0.7817	0.0715
Graph Transformer	<b>0.5433</b>	<b>0.6955</b>	<b>0.7869</b>	<b>0.0715</b>
GIN	0.5388	0.6919	0.7824	0.0718

**Table 7: Ablation results of our proposed MedAlign with different alignment methods on MIMIC-III. Here “OT + CA” denotes the integration of Optimal Transport and Cross-Attention mechanisms, while “MvDA-VC” refers to an extended version of MvDA [11] incorporating view consistency.**

Method	Jaccard $\uparrow$	F1 $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$
	MedAlign			
OT + CA	<b>0.5433</b>	<b>0.6955</b>	<b>0.7869</b>	<b>0.0715</b>
MvDA-VC	0.5339	0.6875	0.7808	0.0720

**Table 8: Ablation results of our proposed MedAlign under different modality absence settings.**

Method	Jaccard $\uparrow$	F1 $\uparrow$	PRAUC $\uparrow$	DDI $\downarrow$
	MIMIC-III			
MedAlign	<b>0.5433</b>	<b>0.6955</b>	<b>0.7869</b>	<b>0.0715</b>
- w/o. Textual	0.5383	0.6915	0.7826	0.0719
- w/o. Molecular	0.5381	0.6911	0.7829	0.0727