Explainable Legal Case Matching via Graph Optimal Transport

Zhongxiang Sun, Weijie Yu, Zihua Si, Jun Xu, *Member, IEEE*, Zhenhua Dong, Xu Chen, Hongteng Xu, *Member, IEEE*, and Ji-Rong Wen

***Abstract*—Providing human-understandable explanations for the matching predictions is still challenging for current legal case matching methods. One difficulty is that legal cases are semi-structured text documents with complicated case-case and case-law article correlations. To tackle the issue, we propose a novel graph optimal transport (GOT)-based legal case matching model that is able to provide not only the matching predictions but also plausible and faithful explanations for the prediction. The model, called GEIOT-Match, first constructs a heterogeneous graph to explicitly represent the semi-structured nature of legal cases and their associations with the law articles. Therefore, matching two legal cases amounts to identifying the rationales from the paired legal case sub-graphs in the heterogeneous graph and then aligning between them. An inverse optimal transport (IOT) model on graphs is learned to extract rationales from paired legal cases. The extracted rationales and the heterogeneous graph demonstrate the key legal characteristics of legal cases, which can be further used to conduct matching and generate explanations for the matching. Experimental results showed that GEIOT-Match outperformed state-of-the-art baselines in terms of matching prediction, rationale extraction, and natural language explanation generation.**

***Index Terms*—Legal retrieval, explainable matching.**

I. Introduction

LEGAL case matching which identifies the relationship between paired legal cases, plays a central role in intelligent legal systems. This task has a high demand on the explainability of matching results because of its critical impacts on downstream applications — the matched legal cases may provide supportive evidence for the judgments of target cases and thus influence the fairness and justice of legal decisions.

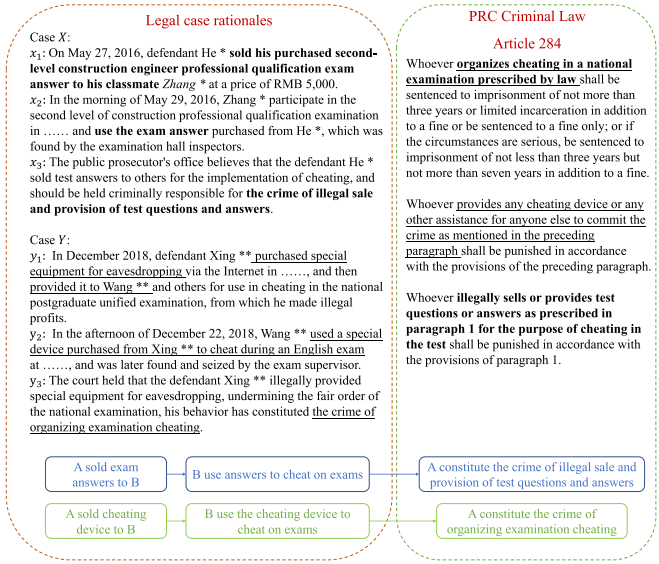


Fig. 1. Example for a pair of legal cases (left) and the corresponding law article (right). The bold and underlined denote the rationales of case X and case Y, respectively. The bottom boxes indicate the sequential order of the criminal acts in X and Y.

Many research efforts have been made to achieve promising matching results, including the early attempts of rule-based strategies [2], [3], [4] and the recent learning-based methods like the Precedent Citation Network (PCNet) [5], and the BERT-based methods [6], [7]. Existing methods suffer from the following challenges in providing plausible and faithful explanations associated with the matching results. First, it is challenging for existing deep matching models to understand the legal cases. This is not only caused by the long-form with complicated contents nature of legal cases, where only the rationales [8], [9], [10] represent the key legal characteristics and support the matching results, but also due to the semi-structured nature of legal cases. In general, rationales in legal cases are organized in the sequential order implicating the transition of criminal acts [11], [12], as shown in the bottom boxes of Fig. 1. However, existing methods tend to overlook the striking different roles between the rationales and other sentences [6], [7], [13]. They cannot take full advantage of the structural information in the rationales. Second, in legal case matching, an ideal explanation

Manuscript received 20 December 2022; revised 2 September 2023; accepted 30 September 2023. Date of publication 13 October 2023; date of current version 19 April 2024. This work was supported in part by the Renmin University of China, National Natural Science Foundation of China under Grants 62376275 and 62106271, in part by the Fundamental Research Funds for the Central Universities, in part by the Research Funds of Renmin University of China under Grant 23XNKJ13, and in part by Intelligent Social Governance Interdisciplinary Platform, Major Innovation and Planning Interdisciplinary Platform for the "Double-First Class" Initiative, and Public Computing Cloud, Renmin University of China. Recommended for acceptance by Y. Tong. *(Corresponding author: Jun Xu.)*

Zhongxiang Sun, Zihua Si, Jun Xu, Xu Chen, Hongteng Xu, and Ji-Rong Wen are with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China (e-mail: 2022000149@ruc.edu.cn; zihua_si@ruc.edu.cn; junxu@ruc.edu.cn; successcx@gmail.com; hongtengxu@ruc.edu.cn; jrwen@ruc.edu.cn).

Weijie Yu is with the School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China (e-mail: yuweijie@ruc.edu.cn).

Zhenhua Dong is with the Noah's Ark Lab, Huawei, Shenzhen 518116, China (e-mail: dongzhenhua@huawei.com).

Digital Object Identifier 10.1109/TKDE.2023.3321935

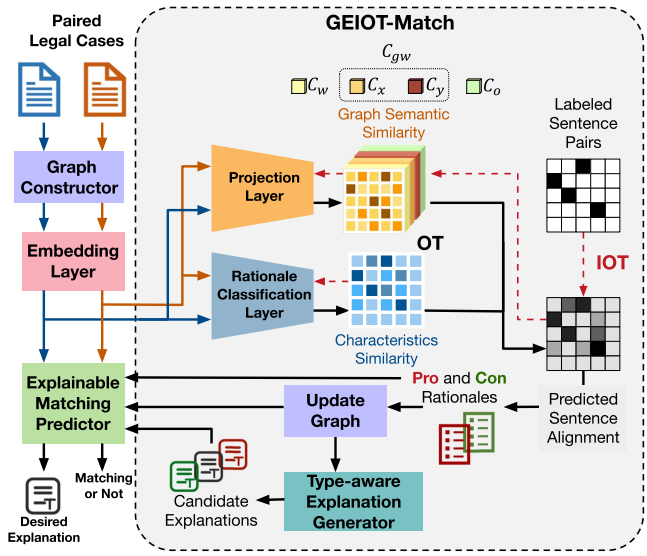


Fig. 2. Architecture of our model GEIOT-Match. C_w, C_{gw}, and C_o represent Wasserstein Distance, Gromov-Wasserstein Distance, and Order-Preserveing Wasserstein Distance, respectively. Note that the red dotted arrows indicate the back-propagation achieved by inverse optimal transport, which are used only in the training phase.

is expected to offer reasons for one side and to refute arguments for the other side [14]. Existing methods often fail to distinguish the pro rationales and con rationales that respectively support the matching and mismatching decisions [9], [15], [16]. Third, for the countries following the Civil Law Systems (e.g., Germany, Japan, and China[1]), only legislative enactments (e.g., law articles) are considered binding for all, and judges' decisions are made heavily based on law articles. Therefore, it is essential to involve law articles in legal case matching. Moreover, law articles are the abstraction of rationales, and thus only the rationales can be the evidence for matching and explanations. As shown in Fig. 1, the bold and underlined sentences on the left denote the rationales of case X and Y. Both correspond to the content in PRC Law Article 284 shown on the right.

To tackle the above challenges, we create a heterogeneous graph and propose a graph optimal transport-based model called GEIOT-Match, as illustrated in Fig. 2 to provide two types of explanations for matching predictions, i.e., rationales and natural language explanations. Concretely, we first construct a heterogeneous graph that regards each legal case as a sub-graph, as shown in Fig. 3(a). In the graph, the legal case nodes connect to the corresponding sentence nodes, which are connected sequentially. Moreover, each sub-graph (legal case) also connects to the corresponding legal article nodes. As a result, the heterogeneous graph explicitly models the structure of the legal cases and provides a basis for the case-case and case-article alignments.

Given the heterogeneous graph, GEIOT-Match formulates the extraction and alignment of pro and con rationales as a graph optimal transport (GOT) problem. In GEIOT-Match, the identified rationales and their alignments are derived from the transport plan of the GOT solution. The GOT is guided by a learnable affinity matrix that reflects semantics, legal characteristic, and the structures of legal cases. The semantics and structure are measured by the Wasserstein distance (for node alignment) and Gromov-Wasserstein distance (for edge alignment) between two cases in the paired sub-graph. The legal characteristic is measured by predicting whether two cross-case sentence nodes have identical rationale types. Thus, the affinity matrix is learned by an inverse GOT process, which corresponds to solving a bi-level optimization problem. In this way, GEIOT-Match learns to extract the pro and con rationales directly.

Once the rationales are extracted, GEIOT-Match updates the heterogeneous graph by removing the non-rationale nodes and adding edges connecting the aligned rationale nodes, as shown in Fig. 3(b). Following the practices in [17], [18], the associated sub-graphs (including the extracted rationales nodes and law article nodes) are fed to a pre-trained language model to generate label-specific natural language explanations. The explanations stand for the pro and con reasons of matching. For weighing the pro and con reasons and involving the law articles, the final matching results are made based on the extracted rationales, the associated sub-graphs, and the generated explanations.

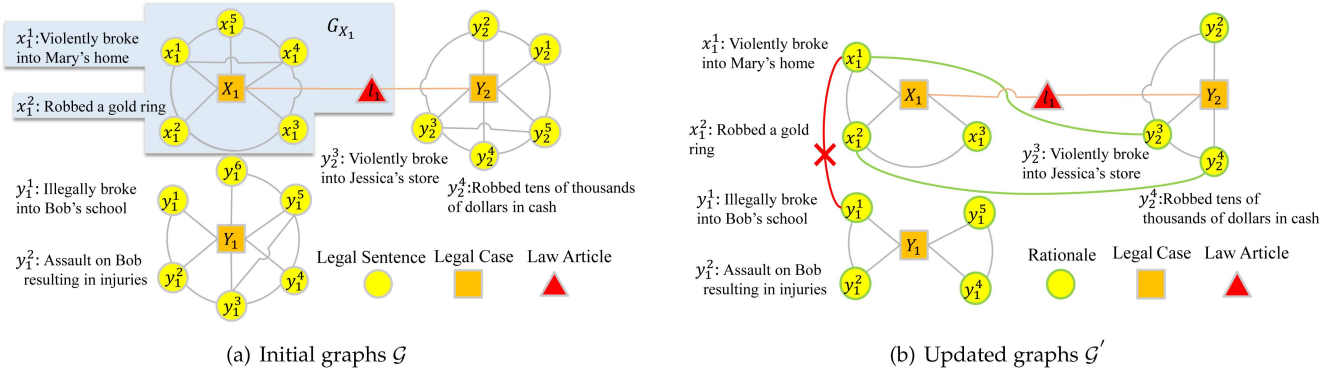
The contributions of the paper can be summarized as follows: (1) We construct a heterogeneous graph to explicitly model the semi-structured nature of legal cases and the case-case and case-article correlations. This graph provides a basis for explainable legal case matching, especially under Civil Law Systems; (2) We propose a novel graph optimal transport-based approach to learn the legal case matching model based on the heterogeneous graph. The model also provides rationales and natural language explanations for matching predictions; (3) Experimental results indicate GEIOT-Match not only achieved state-of-the-art matching accuracy but also produced plausible and faithful explanations for the matching predictions.

II. Related Work

A. Legal Case Matching

Conventional legal case matching methods highly depend on expert knowledge [4], e.g., the decomposition of legal issues [2] and the ontological framework of the problem [3]. In recent years, learning-based legal case matching strategies have shown advantages in exploring the semantics of legal cases, which can be roughly categorized into network-based methods [5], [19], [20], [21] and text-based methods [6], [7]. The network-based methods construct a Precedent Citation Network (PCNet), in which the vertices are legal cases and directed edges indicate the citations of source cases used by target cases. Based on PCNet, [5] used the Jaccard similarity index between the sets of precedent citations to infer the similarity of two legal cases. [20] used whether the sets of precedent citations occurs in the same cluster to measures to what extent the two cases are similar. [21] proposed Hier-SPCNet to capture all domain information inherent in both statutes and precedents. The text-based methods rely on the textual content of the cases and measure the similarity of two legal cases based on their semantics. [6] proposed BERT-PLI to break a case into paragraphs and model the interactions between the paragraphs. It first adopted BERT to encode each

[1]In this paper, we take the legal cases in Chinese as examples.



(a) Initial graphs $\mathcal{G}$ (b) Updated graphs $\mathcal{G}'$

Fig. 3. Examples of legal case graphs: (a) Initial legal case graphs with nodes being Legal Sentence, Legal Case, and Law Article. The legal case node X_1 represents the source case, and the legal case node Y_1 and Y_2 represent the target cases. The blue area G_{X_1} represents the sub-graph of legal case X_1. (b) Legal case graphs are updated by adding edges for aligned cross-case nodes and filtering the irrelevant nodes and edges for matching. The green edges are true cross-graph rationale connections, while the red edge is a misleading edge that the baseline model can easily misidentify.

paragraph in two legal cases, then applied max-pooling to capture their matching signal, and finally, used a recurrent neural network (RNN) with an attention mechanism to predict their matching score. Similarly, [22] proposed to segment two legal cases into paragraphs and aggregate the paragraph-level similarity. Inspired by the success of pre-trained language models in the generic domain, [7] pre-trained a Longformer-based language model with tens of millions of criminal and civil case documents. Although these studies effectively improve the performance, they often have difficulties in explaining their predictions, which limits their practical applications [23].

B. Explainable AI in the Legal Domain

Recently, researchers have made efforts to achieve explainable AI models in various applications of the legal domain [24]. In the task of legal judgment prediction, [25] formalized the court view generation problem as a label-conditioned Seq2Seq task and generated court views based on fact descriptions and charges. [8] proposed a neural based system to jointly extract readable rationales and elevate charge prediction accuracy by a rationale augmentation mechanism. [10] compared various rationale constraints in the form of regularizers and proposed to improve faithfulness and rationale quality in a paragraph-level setup for legal text classification task concerning alleged violations. [26] proposed the Joint Prediction and Generation Model (JPGM) to predict charges and court views. In the task of legal question answering, [27] proposed to first detect elements of fact descriptions by iteratively asking questions about pre-defined charge-specific principles and then used the detected elements for prediction. Different from the above work, we focus on the legal case matching task, extracting rationales and generating natural language-based explanations to support matching results.

C. Graph Neural Networks

Graph neural networks have demonstrated their effectiveness on different graph tasks. Several approaches are proposed to learn representations for nodes and graphs, such as GCN [28], GAT [29], and HGT [30], etc. These methods generally aggregate the characteristics of nearby nodes to define a target node. Graph matching (GM) identifies relationships between two graph-structured objects and has been applied in text-matching tasks. For example, [31] proposed the Concept Interaction Graph to represent an article as a graph of concepts, generated the representation for each vertex, and then used a GCN to obtain the matching score; [32] plugged PageRank into transformer to filter information based on a sentence similarity graph for long-form text matching. Besides, several legal studies have also employed graph neural networks to solve legal challenges. On the basis of the legal knowledge graph, [33] used topic modeling to select the features for training relational graph convolutional networks for citation link prediction and case similarity. To improve computer document similarity estimation, [21] presented a heterogeneous network containing citation links between case papers and statutes, as well as citation and hierarchy relationships among the statutes. This paper examines the application of graph neural networks to explainable legal case matching, which concentrates on developing graph networks from both the semantic and structural perspectives of legal cases.

III. Preliminary: Optimal Transport

Optimal transport (OT) [34], [35] defines a distance between probability distributions, which has been widely used in many machine learning tasks, such as point cloud alignment [36], [37], graph matching [38], [39], data clustering [40], [41], and sequence representations learning [42], [43], [44].

Originally, let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ denotes two discrete distributions, formulated as $\boldsymbol{\mu} = \sum_{m=1}^{M} u_m \delta_{\boldsymbol{x}_m}$ and $\boldsymbol{\nu} = \sum_{n=1}^{N} v_n \delta_{\boldsymbol{y}_n}$, with $\delta_{\boldsymbol{x}}$ as the Dirac function centered on $\boldsymbol{x}$. The weight vectors $\boldsymbol{u} = \{u_m\}_{m=1}^{M} \in \triangle_M$ and $\boldsymbol{v} = \{v_n\}_{n=1}^{N} \in \triangle_N$ respectively belongs M- and N-dimensional simplex, i.e., $\{u_m\}_{m=1}^{M} = \{v_n\}_{n=1}^{N} = 1$, the optimal transport distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is defined as

$$\begin{aligned}\mathbf{A}^* &= \arg\min_{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \mathbb{E}_{m,n\sim\mathbf{A}}[c(x_m, y_n)] \\ &= \arg\min_{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \sum_{m=1}^{M}\sum_{n=1}^{N} a_{mn} \cdot c(x_m, y_n), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\mathbf{A} \in \mathbb{R}_+^{M\times N} | \mathbf{A}\mathbf{1}_N = \boldsymbol{\mu}, \mathbf{A}^\top \mathbf{1}_M = \boldsymbol{\nu}\}$, which represents an arbitrary joint distribution with

marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, and $c(x_m, y_n)$ is the cost function evaluating the distance between x_m and y_n. The OT matrix $\mathbf{A}^*$ provides a soft matching and the element of the optimal transport, *i.e.*, a^*_{mn}, indicates the probability of the coherency of x_m and y_n, which provides the evidence for their matching.

In this work, we apply three variants of OT to conduct multi-level alignment in the heterogeneous graph.

Wasserstein Distance (WD) is a special form of vanilla OT with the cost set as p-norm, i.e., $\mathbf{C}_w := c(x_m, y_n) = ||x_m - y_n||_p$. WD can be fast estimated by introducing an entropic regularizer [45] $\langle \mathbf{A}, \log \mathbf{A} \rangle$ into (1), where $\langle \cdot \rangle$ denotes the Frobenius dot-product. WD defines an optimal transport distance that measures the discrepancy between each pair of samples and thus provides a natural choice for node alignment in the graph.

Order-Preserveing Wasserstein Distance (OPWD) is an extension of WD that concentrates on diagonal entries by transporting the neighboring elements in one sequence into some other neighboring elements in another sequence with a nearby temporal position. To this end, OPWD penalizes the (1) with two additional terms so that the optimal transport plan can be written as

$$\mathbf{A}^* = \arg\min_{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{A}, \mathbf{C} \rangle - \lambda_1 I(\mathbf{A}) + \lambda_2 KL(\mathbf{A} \| \mathbf{P}), \tag{2}$$

where λ_1 and λ_2 are two hyperparameters, KL denotes the Kullback-Leibler divergence, $I(\mathbf{A}) = \sum_{m,n} \frac{a_{mn}}{(\frac{m}{M} - \frac{n}{N})^2 + 1}$ is the inverse difference moment of the transport plan $\mathbf{A}$, and $\mathbf{P}$ is a prior Gaussian distribution whose values decrease gradually from the diagonal to both sides, $\mathbf{P}(m,n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{l^2(m,n)}{2\sigma^2}}$ and $l(m,n) = \frac{|\frac{m}{M} - \frac{n}{N}|}{\sqrt{\frac{1}{M^2} + \frac{1}{N^2}}}$. Please note that (2) equals to

$$\mathbf{A}^* = \arg\min_{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{A}, \tilde{\mathbf{C}} \rangle - \lambda_2 \langle \mathbf{A}, \log \mathbf{A} \rangle. \tag{3}$$

where $\tilde{\mathbf{C}} = \mathbf{C} + \mathbf{C}_o, \mathbf{C}_o = -\lambda_1 \mathbf{E} + \lambda_2 (\frac{\mathbf{F}}{2\sigma^2} + \log(\sigma\sqrt{2\pi}))$, $\mathbf{E} = [1/(m/N - n/M)^2]_{mn}, \mathbf{F} = [l^2(m,n)]_{mn}$. Motivated by the fact that rationales in legal cases are organized in the sequential order, we utilize OPWD to explicitly model the structure of legal cases.

Gromov-Wasserstein Distance (GWD) defines an optimal transport-like distance C_{gw} for metric spaces. In the graph matching scenario, given two graphs $\{\mathbf{C}_x = [c^x_{ij}] \in \mathbb{R}^{M\times M}, \mathbf{C}_y = [c^x_{i'j'}] \in \mathbb{R}^{N\times N}\}$ and their empirical node distributions $\{\boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\nu} \in \mathbb{R}^N\}$, GWD calculates distances between pairs of samples within each graph and measures how these distances compare to those in the other graph, and the corresponding optimal transport plan is

$$\mathbf{A}^* = \underset{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})}{\arg\min} \left(\sum_{i,j=1}^{M} \sum_{i',j'=1}^{N} |c^x_{ij} - c^y_{i'j'}|^2 a_{ii'} a_{jj'} \right)^{\frac{1}{2}}. \tag{4}$$

where c^x_{ij} and $c^y_{i'j'}$ can be regarded as the edges of two graphs, i.e., $\mathbf{C}_{gw} := |c^x_{ij} - c^y_{i'j'}|^2$, and thus GWD aligns edges in different graphs. Inspired by this property, we represent each legal case as a sub-graph of the heterogeneous graph and leverage GWD to not only reflect the topological structure of each pair of graphs (cases) but also yield correspondence across the graphs (case pair).

IV. The Proposed Approach: GEIOT-Match

A. Problem Statement

The explainable legal case matching task provides us a set of labeled data tuples $\mathcal{D} = \{(X, Y, \mathbf{r}^X, \mathbf{r}^Y, \hat{\mathbf{A}}, L^X, L^Y, z, e)\}$. For each tuple $(X, Y, \mathbf{r}^X, \mathbf{r}^Y, \hat{\mathbf{A}}, L^X, L^Y, z, e)$ in the dataset, its elements include 1) a pair of legal cases $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the sets of source and target legal cases; 2) the rationale labels of the paired cases, denoted as $\mathbf{r}^X$ and $\mathbf{r}^Y$, respectively; 3) a binary alignment matrix $\hat{\mathbf{A}}$ indicating the sentence-to-sentence relation between rationales of X and rationales of Y; 4) the law articles of the paired cases, denoted as L^X and L^Y; and 5) the matching label z and the set of sentences (denoted as e) explaining the reasons for z.

In practice, we represent each legal case from the sentence and document levels. For the sentence-level, $X = \{x_m\}_{m=1}^M$ and $Y = \{y_n\}_{n=1}^N$, where x_m (y_n) denotes the embedding of the mth (nth) sentence in X (Y);[2] for the document-level, $\mathbf{e}(X) \in \mathbb{R}^d$ and $\mathbf{e}(Y) \in \mathbb{R}^d$ are the corresponding embedding of X and Y. The law articles embedding L^X and L^Y are represented as $\{l_{x_q}\}_{q=1}^Q$ and $\{l_{y_p}\}_{p=1}^P$, respectively, where l_{x_q} (l_{y_p}) denotes the embedding of the qth (pth) law article in L^X (L^Y). Typically, each embedding can be calculated by using the output of at the [CLS] token of a BERT model pre-trained on a Chinese legal case corpus.[3] The rationale labels are associated with the sentence embeddings, i.e., $\mathbf{r}^X = \{r_{x_m}\}_{m=1}^M$ and $\mathbf{r}^Y = \{r_{y_n}\}_{n=1}^N$, where the rationale label of a sentence s is designed as following [46]:

$$r_s = \begin{cases} 0 & s \text{ is not a rationale,} \\ 1 & s \text{ is a key circumstance,} \\ 2 & s \text{ is a constitutive element of crime,} \\ 3 & s \text{ is a focus of disputes.} \end{cases} \tag{5}$$

The remaining elements, i.e., $\hat{\mathbf{A}} = [\hat{a}_{mn}] \in \{0,1\}^{M\times N}$, $z \in \{0,1,2\}$, and e, are annotated manually, where

$$\hat{a}_{mn} = \begin{cases} 0 & r_{x_m} \neq r_{y_n}, \\ 1 & r_{x_m} = r_{y_n} \,\&\, x_m \cong y_n, \end{cases} \tag{6}$$

$$z = \begin{cases} 0 & \text{Mismatched } (X, Y), \\ 1 & \text{Partially matched } (X, Y), \\ 2 & \text{Matched } (X, Y), \end{cases} \tag{7}$$

where $x_m \cong y_n$ means the sentences x_m in X can serve as a reference for matching with the sentence y_n in Y. $\hat{a}_{mn} = 1$ means aligned rationales while $\hat{a}_{mn} = 0$ means misaligned rationales. They provide pro and con evidence for matching prediction, respectively.

[2]For brevity, x_m and y_n also denote sentences in equations.

[3]Corpus available at https://github.com/thunlp/OpenCLaP. Note that GEIOT-Match is applicable to the corpus of other languages with the corresponding embeddings.

TABLE I
DESCRIPTIONS OF EDGES IN THE HETEROGENEOUS GRAPH $\mathcal{G}$; FUNCTION "DIS" MEASURES THE SEMANTIC DISTANCE OF TWO SENTENCE EMBEDDINGS, I.E., COSINE SIMILARITY; τ_1 IS A THRESHOLD HYPER-PARAMETER

Edge	Description
$e_{x_i-x_j}$	sentence x_i is next to sentence x_j in legal case X or dis$(x_i, x_j) > \tau_1$
e_{X-x_i}	legal case X contain sentence x_i
e_{X-l_i}	legal case X contains law article l_i
$e_{l_i-l_j}$	law article l_i and law article l_j are in the same level of Law

B. Graph Construction

Legal cases are inherently semi-structured, wherein the rationales are sequentially organized. To better represent rationales and distinguish the rationales from noises, we construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which contains all the legal cases in $\mathcal{D}$. As shown in Fig. 3, $\mathcal{V}$ denotes the set of nodes, including three types of nodes: the case nodes X, the article nodes l, and the sentence nodes x. $\mathcal{E}$ denotes the set of undirected edges, including four types of edges listed in Table I. As shown in Fig. 3(a), each case node X in $\mathcal{G}$, all its 1-hop neighboring sentence nodes and law article nodes form a sub-graph denoted as G_X.

Based on the heterogeneous graph $\mathcal{G}$, we propose to learn the following three modules for explainable legal case matching: 1) f_1 extracts aligned and misaligned rationales from the paired legal case sub-graphs (Section IV-C); 2) f_2 generates candidate explanations based on the heterogeneous graph $\mathcal{G}$ (Section IV-D); and 3) f_3 predicts the final matching label based on the extracted rationales, generated explanations, and graph features (Section IV-E).

C. GEIOT-Based Rationale Extraction

As the key of the above three modules, f_1 learns to extract aligned and misaligned rationales for the paired legal case sub-graphs G_X and G_Y. As the relationship of case nodes and article nodes between the sub-graphs is obvious and does not provide effective explanations for the final matching, we primarily concentrate on extracting and aligning rationales for the sentence nodes, whose labels are expressed as node-to-node alignment matrices $\hat{\mathbf{A}}$. However, the alignment matrices are manually labeled, often very sparse, containing many false negative elements. To learn our model robustly for graph matching, we develop a novel learning paradigm from the viewpoint of Graph Optimal Transport (GOT). As illustrated in Fig. 2, three types of OT distance are adopted for our method; namely, Wasserstein Distance (WD) [39], Gromov Wasserstein Distance (GWD) [37], Order-Preserved Wasserstein Distance (OPWD) [47].

Although WD can capture node similarity between graphs, it cannot be directly applied to graph alignment. Since only the similarity between x_m and y_n is considered, WD does not take context information in graphs into account. In Fig. 3, for example, the sentence pair (x_1^1, y_1^1) has similar semantic meanings as the pair (x_1^1, y_2^3), but the context meanings of the two pairs are completely different, i.e., (x_1^1, x_1^2) describes the crime of robbery, which is similar to (y_2^3, y_2^4). However, (y_1^1, y_1^2) describes the crime of injury and should not be matched. To address the above problem, we propose to use GWD to model the contextual information by aligning the edges in different graphs. GWD can be used to calculate distances between pairs of nodes within each case, as well as measure how these distances compare to those in the counterpart case. Moreover, rationales in legal cases are organized in sequential order, i.e., fact descriptions in legal cases are often written in the temporal order of events [12]. Thus, we use the OPWD to model the order information of legal cases explicitly.

When learning the rationale extraction module f_1 in the above OT framework, the critical learning tasks become 1) unifying these three distances in learning the mutually-beneficial affinity matrix $\mathbf{C}$ to better compute the optimal transport $\mathbf{A}^*$; and 2) fitting the optimal transport $\mathbf{A}^*$ robustly to the manually-labeled (noisy) alignment matrix $\hat{\mathbf{A}}$. In this work, we solve these two tasks jointly by solving the following inverse optimal transport (IOT) problem.

According to the analysis above, we need to learn both the affinity matrix and the optimal transport based on the sentence-node embeddings ($\{x_m\}_{m=1}^M$ and $\{y_n\}_{n=1}^N$) and their annotated alignment matrix $\hat{\mathbf{A}}$, which leads to a so-called graph-based explainable inverse optimal transport (GEIOT) problem [48], [49]

$$\begin{aligned} &\mathbf{C}^* = \arg\min_{\mathbf{C}\in\mathbb{R}^{M\times N}} \mathrm{KL}(\hat{\mathbf{A}} \| \mathbf{A}^*(\mathbf{C})), \\ &s.t.\ \mathbf{A}^*(\mathbf{C}) = \arg\min_{\mathbf{A}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{A}, \mathbf{C}\rangle + \gamma\langle \mathbf{A}, \log \mathbf{A}\rangle. \end{aligned} \quad (8)$$

This problem is a typical bi-level optimization problem, in which the affinity matrix $\mathbf{C}$ is the upper-level variable while the optimal transport $\mathbf{A}$ is the lower-level variable. The upper-level problem minimizes the KL divergence between $\hat{\mathbf{A}}$ and $\mathbf{A}^*$, i.e., $\mathrm{KL}(\hat{\mathbf{A}} \| \mathbf{A}^*) = \sum_{m,n} \hat{a}_{mn} \log \frac{\hat{a}_{mn}}{a^*_{mn}}$, which corresponds to the cross-entropy loss. The optimal transport $\mathbf{A}^*$ is a function of affinity matrix $\mathbf{A}^*(\mathbf{C})$, whose optimization corresponds to the lower-level problem given $\mathbf{C}$.

Solving the GEIOT problem in (8) provides us with a robust method to learn the rationale extraction module. Specifically, on the one hand, the upper-level problem fits the optimal transport to the limited and noisy alignment matrix under the constraint provided by the lower-level optimal transport problem, which suppresses the risk of over-fitting greatly. On the other hand, the lower-level problem provides us with an optimal transport matrix to indicate the aligned rationales, which is determined by the optimized affinity matrix and reveals sentence-level similarity between the paired legal cases. As a result, the optimal transport $\mathbf{A}^*$ derived from the optimal affinity matrix $\mathbf{C}^*$ represents the global alignment between rationales of a legal case pair. Accordingly, we can extract pro (aligned) and con (misaligned) rationales by setting a threshold τ_2, i.e., x_m and y_n are selected as pro rationales if $a^*_{mn} \geq \tau_2$, otherwise, are selected as con rationales.

Note that because the lower-level problem is strictly convex. This IOT problem can be solved efficiently by alternating optimization. Given the current affinity matrix $\mathbf{C}$, we can optimize $\mathbf{A}$ via the Sinkhorn scaling algorithm and then optimize $\mathbf{C}$ via stochastic gradient descent based on fixed $\mathbf{A}$.

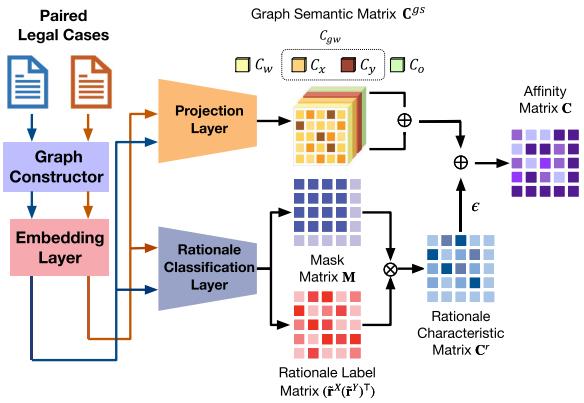


Fig. 4. Illustration of the affinity matrix construction.

We parameterize the affinity matrix $\mathbf{C}$ by a neural network, which takes paired sentence-node embeddings as its input and outputs their discrepancies according to their legal characteristics and semantics jointly. As illustrated in Fig. 4, we model $\mathbf{C}$ as the combination of a rationale characteristic matrix $\mathbf{C}^r$ and a graph semantic matrix $\mathbf{C}^{gs}$ which unify the three distances aforementioned

$$\mathbf{C} = \epsilon \mathbf{C}^r + \mathbf{C}^{gs}, \tag{9}$$

where ϵ is a negative hyper-parameter to encourage the alignment of those sentence pairs with the same rationale label by significantly reducing the transport between them. The $\mathbf{C}^{gs}$ and $\mathbf{C}^r$ are constructed by the following steps.

1) Construction of the Graph Semantic Matrix $\mathbf{C}^{gs}$: Following [39], we can use a transport plan $\mathbf{A}$ shared by both WD and GWD to better consider the semantic and context similarity. Considering the OPWD can be adapted by adding order-preserving cost to the original cost, the transport plan $\mathbf{A}^*$ can be solved by the Sinkhorn scaling algorithm in (1). GEIOT-Match constructs the graph semantic matrix $\mathbf{C}^{gs} \in \mathbb{R}^{M\times N}$ to unify the three distances

$$\mathbf{C}^{gs} = \rho \mathbf{C}_w + (1-\rho)\mathbf{C}_{gw} + \mathbf{C}_o$$

$$\mathbf{C}_w = \mathrm{dis}(\bar{\mathbf{s}}^X, \bar{\mathbf{s}}^Y), \mathbf{C}_x = \mathrm{dis}(\bar{\mathbf{s}}^X, \bar{\mathbf{s}}^X), \mathbf{C}_y = \mathrm{dis}(\bar{\mathbf{s}}^Y, \bar{\mathbf{s}}^Y),$$

where $\bar{\mathbf{s}}^X$ and $\bar{\mathbf{s}}^Y$ represent the contextual sentence embedding of legal case X and Y, respectively. ρ is the hyper-parameter for controlling the importance of different cost functions. $\bar{\mathbf{s}}^X$ and $\bar{\mathbf{s}}^Y$ are obtained from a trainable two-layer MLP (projection layer in Fig. 4) on the frozen sentence-node embeddings $\{x_m\}_{m=1}^M$ and $\{y_n\}_{n=1}^N$.

2) Construction of Rationale Characteristic Matrix $\mathbf{C}^r$: The rationale characteristic matrix $\mathbf{C}^r$ indicates the rationales having the same legal characteristics, and the legal characteristics are categorized according to the rationale labels shown in (5). Taking the sentence nodes of paired legal case sub-graphs G_X and G_Y as the inputs, our GEIOT-Match predicts the rationale labels of their sentence nodes, denoted as $\hat{\mathbf{r}}^X = \{\hat{r}_{x_m}\}_{m=1}^M$ and $\hat{\mathbf{r}}^Y = \{\hat{r}_{y_n}\}_{n=1}^N$, respectively, which is achieved by solving a sentence-level multi-class classification problem. Formally, given sentence nodes of G_X, our GEIOT-Match would identify the legal characteristics of each sentence embedding x_m in G_X by calculating a probabilistic distribution over the four classes shown in (5)

$$\hat{r}_{x_m} = \arg\max_{k\in\{0,\ldots 3\}} P(r=k|x_m), \tag{10}$$

where $\{P(r=k|x_m)\}_{k=0}^3$ represent the distribution of the rationale labels conditioned on the sentence embedding x_m. In this work, we parameterize the distribution as follows:

$$\{P(r=k|x_m)\}_{k=0}^3 = \mathrm{softmax}(\mathbf{W}\mathbf{s}_{x_m}^{(L)} + \mathbf{b}), \tag{11}$$

where the softmax converts a 4-dimensional vector to a distribution over four classes, matrix $\mathbf{W}$ and vector $\mathbf{b}$ are trainable parameters, and $\mathbf{s}_{x_m}^{(L)}$ is the output of a stacked of L-layer gated convolutional neural network [50] where the lth layer is

$$\mathbf{s}_{x_m}^{(l)} = \mathbf{s}_{x_m}^{(l-1)} + \mathrm{conv}_1(\mathbf{s}_{x_m}^{(l-1)}) \otimes \sigma(\mathrm{conv}_2(\mathbf{s}_{x_m}^{(l-1)})),$$

for $l = 1, \ldots, L$, and $\otimes$ denotes element-wise multiplication, conv_1 and conv_2 denote two dilate convolutional neural Network [51] with the same convolution kernel size. Note that the use of a stacked gated convolutional neural network enables the model to capture further distances without increasing model parameters, which effectively addresses the issue caused by a large number of sentences in a legal case. $\sigma(\cdot)$ denotes a sigmoid gating function controlling which inputs $\mathrm{conv}_1(\mathbf{s}_{x_m}^{(l-1)})$ of the current context are relevant. In the first layer, $\mathbf{s}_{x_m}^{(0)}$ is obtained by adding a trainable one-layer multi-layer perceptron on the frozen sentence embedding x_m.

Similarly, given the sentence nodes of G_Y, the legal characteristics of each sentence embedding y_n in G_Y can also be identified by classifying y_n with the same sentence representations model and neural networks defined above. As a result, the rationale characteristic matrix $\mathbf{C}^r = [c_{mn}^r] \in \{0,1\}^{M\times N}$ can be defined to explicitly indicate whether two sentences have the same predicted legal characteristics

$$\mathbf{C}^r = \mathbf{M} \otimes (\tilde{\mathbf{r}}^X(\tilde{\mathbf{r}}^Y)^\top), \tag{12}$$

where $\mathbf{M} \in \{0,1\}^{M\times N}$ is a mask matrix filtering out the padding sentences, $\tilde{\mathbf{r}}^X \in \{0,1\}^{M\times 4}$ and $\tilde{\mathbf{r}}^Y \in \{0,1\}^{N\times 4}$ are the rationale label matrix, whose rows are one-hot representations of $\hat{\mathbf{r}}^X$ and $\hat{\mathbf{r}}^X$. To incorporate (10) into (12) in a differentiable manner, we apply the Straight-Through Gumbel Trick [52], which replaces the discrete argmax with a continuous and differentiable estimator, to derive $\tilde{\mathbf{r}}^X$ and $\tilde{\mathbf{r}}^Y$. Accordingly, $c_{mn}^r = 1$ means that the mth sentence in X and the nth sentence in Y are identified as rationales (i.e., $\hat{r}_{x_m} \neq 0$ and $\hat{r}_{y_n} \neq 0$) and they belong to the same rationale type ($\hat{r}_{x_m} = \hat{r}_{y_n}$).

D. Generating Candidate Explanations

As aforementioned, the extracted rationales $\hat{r}^X$, $\hat{r}^Y$, and the optimal transport $\mathbf{A}^*$ indicate pro and con rationale pairs, and thus, can help to update the associated heterogeneous graph G and generate explanations (i.e., e's) to support matching results (i.e., z's).

Graph Update: As illustrated in Fig. 3, we update both the edges and nodes of the corresponding heterogeneous graph $\mathcal{G}$ to $\mathcal{G}'$. Based on the extracted rationales $\hat{r}^X$ and $\hat{r}^Y$, we update the

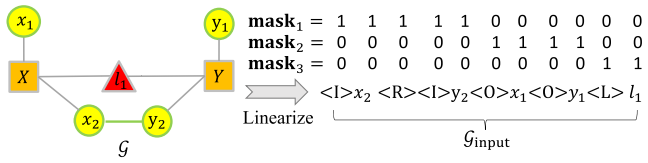


Fig. 5. Illustration of converting paired legal case sub-graphs into text sequences. The sepcical token <I > means the connected ratioales, < R> means the relation, <O > means the unconnected rationales, and < L> means the law articles.

sentence nodes to rationale nodes by removing the non-rationale nodes and their associated edges. Moreover, we add edges connecting the aligned rationale nodes based on $\mathbf{A}^*$, which are represented by green lines in Fig. 3(b).

Explaination Generate: Following the work in [17], our GEIOT-Match exploits the existing pre-trained language model[4] to build three label-specific explanation generators, that is: $f_2 = \{E_z\}, z = 2, 1, 0$ respectively corresponds to matched, partially matched, and mismatched decisions as shown in (6). The three generators are fine-tuned separately. For example, for $z = 0$, the data for fine-tuning E_0 is selected from the training corpus: $\mathcal{D}_0 = \{(G'_X, G'_Y) \subseteq \mathcal{G}', (e, z = 0) \subseteq \mathcal{D}\}$. However, the structural nature of the input legal case graph makes it unsuitable to naively apply sequential encoder-decoder architecture to generate explanations. Following the method in [53], we linearize the graph into sequential input G_{input} shown in Fig. 5.

Moreover, considering the input G_{input} contains three types of sequence (i.e., <I >, < O>, <L>), which play different roles in the explanation generation, we propose a type-aware adapter after each encoder layer to better leverage different types of information. Let $\mathbf{h}$ denote the hidden state after the self-attention. The type-aware adapter function can be defined as follows:

$$\hat{\mathbf{h}} = \mathbf{h}\mathbf{W}_1\mathbf{mask}_1 + \mathbf{h}\mathbf{W}_2\mathbf{mask}_2 + \mathbf{h}\mathbf{W}_3\mathbf{mask}_3, \quad (13)$$

where $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{W}_3$ are trainable parameters, and the masks are used to identify different types of input, i.e., $\mathbf{mask}_1$ indicates the matching information, $\mathbf{mask}_2$ indicates the rationale information contained in a single sub-graph, and $\mathbf{mask}_3$ indicates the law article information.

To fine-tune the parameters in E_0, we optimize a language modeling loss [54] that compares difference between the generated explanation $E_0([G'_X; G'_Y])$ and the human-annotated explanation e. Similarly, E_1 and E_2 are fine-tuned based on corresponding subsets $\mathcal{D}_1$ and $\mathcal{D}_2$.

At explanation generation phase, given a tuple (G'_X, G'_Y), we feed the constructed input text sequence to E_0, E_1, and E_2, generating three candidate explanations $\hat{e}_0, \hat{e}_1$ and $\hat{e}_2$.

E. Matching Prediction

Instead of considering all of the sentences in the paired legal cases, GEIOT-Match learns the f_3 to conduct matching based on the extracted rationales, the generated candidate explanations, and the graph features of legal cases. This strategy not only makes the extracted rationales and generated explanations faithful to the matching predictions, also makes full use of the graph structure information. Formally, given the heterogeneous graph $\mathcal{G}'$, a paired legal case sub-graph (G'_X, G'_Y) sampled from $\mathcal{G}'$ and the candidate explanations, our GEIOT-Match would identify their relation by calculating a probabilistic distribution over the three classes shown in (6)

$$\{P(z = k|(X, Y)\}_{k=0}^{2} = \text{softmax}(\mathbf{W}[\hat{z}_0; \hat{z}_1; \hat{z}_2] + \mathbf{b}), \quad (14)$$

where $[;]$ concatenates vectors, and $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters. As for $\hat{\mathbf{z}}_i$ $(i \in \{0, 1, 2\})$, following the practice in [17], [18], the matching scores are computed based on the rationale nodes, the corresponding candidate explanations, and the graph features of legal case nodes learned from $\mathcal{G}'$

$$\hat{z}_i = \text{MLP}([\mathbf{s}_{r^X}; \mathbf{s}_{r^Y}; \mathbf{s}_{g^X}; \mathbf{s}_{g^Y}; \mathbf{s}_{\hat{e}_i}]), \quad (15)$$

where $\mathbf{s}_{r^X}, \mathbf{s}_{r^Y}, \mathbf{s}_{\hat{e}_i}$ respectively denote the embeddings of rationale nodes of G'_X, G'_Y, and the candidate explanation $\hat{e}_i$ which are obtained by tuning the BERT model[5]; $\mathbf{s}_{g^X}$ and $\mathbf{s}_{g^Y}$ denotes the graph feature of legal case sub-graphs G'_X and G'_Y, which are learned from legal case graph $\mathcal{G}'$ using Heterogeneous Graph Transformer(HGT) [30]; MLP denotes a two-layer perceptron with sigmoid activation functions. In Fig. 6, we illustrate the generating process of $\hat{z}_i$. Accordingly, our GEIOT-Match makes the final matching decision for paired legal cases (X, Y) as

$$\hat{z} = \arg\max_{k \in \{0,1,2\}} P(z = k|(X, Y)), \quad (16)$$

and outputs the explanation corresponding to the highest matching score simultaneously.

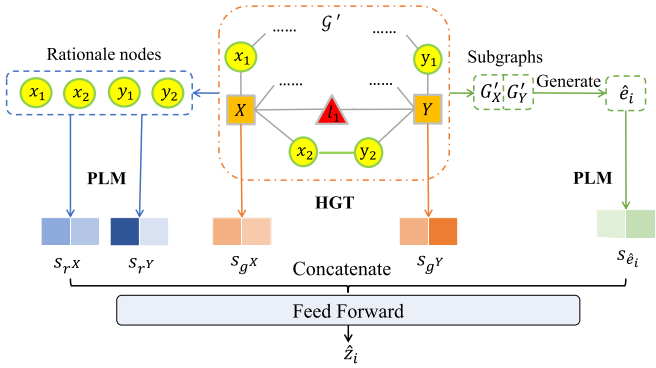


Fig. 6. Illustration of embedding generation process for $\hat{z}_i$.

F. Model Training

GEIOT-Match has parameters to determine during the training, including those in the pro and con rationales extraction (f_1), those in the candidate explanations generation (f_2), and those in the matching (f_3). These models parameters, respectively denoted as $\theta_{f_1}, \theta_{f_2}, \theta_{f_3}$ are trained sequentially, and the output of the f_1 is used as the input of the f_2, and the outputs of the f_1

[4]We adopt Chinese T5-PEGASUS model: https://github.com/ZhuiyiTechnology/t5-pegasus. Note that other pre-trained language models are also applicable.

[5]https://github.com/thunlp/OpenCLaP

and f_2 are used as the input of the f_3. Specifically, in the f_1, the learning objective is defined to measure the loss of the pro and con rationales extraction

$$\mathcal{L}_{f_1} = \sum_{(\mathbf{r}^X, \mathbf{r}^Y, \hat{\mathbf{A}}) \in \mathcal{D}, (G_X, G_Y) \in \mathcal{G}} \mathcal{L}_{\mathcal{R}} + \gamma_1 \mathcal{L}_{\mathcal{A}}, \tag{17}$$

where, for each legal case pair in the dataset, the loss function consists of two parts: the rationale identification loss $\mathcal{L}_{\mathcal{R}}$ and the affinity matrix loss $\mathcal{L}_{\mathcal{A}}$. The $\gamma_1 > 0$ is a hyper-parameter controlling their weights. The rationale identification loss $\mathcal{L}_{\mathcal{R}}$ is defined as the cross-entropy loss between the ground-truth rationale labels of each sentence and the corresponding predictions

$$\mathcal{L}_{\mathcal{R}} = -\sum_{k=0}^{3} \left(\sum_{m=1}^{M} \delta(r_{x_m}, k) \log(P(\hat{r}_{x_m} = k | x_m)) + \sum_{n=1}^{N} \delta(r_{y_n}, k) \log(P(\hat{r}_{y_n} = k | y_n)) \right), \tag{18}$$

where $\delta(r, k) = 1$ if $r = k$ else 0. The loss $\mathcal{L}_{\mathcal{A}}$ is based on the GEIOT problem in (8)

$$\mathcal{L}_{\mathcal{A}} = \mathrm{KL}(\hat{\mathbf{A}} || \mathbf{A}^*(\mathbf{C})) + \gamma_2 \sum_{m=1}^{M} \sum_{n=1}^{N} \delta(\hat{r}_{x_m}, \hat{r}_{y_n}) c_{mn}, \tag{19}$$

where the first term corresponds to the GEIOT problem that optimize the affinity matrix and the associated optimal transport to fit a small number of alignment labels (i.e., $\hat{\mathbf{A}}$). The second term is an unsupervised loss based on the predicted rationale labels, which explicitly regularizes the affinity matrix $\mathbf{C}$ to minimize the discrepancy between identical rationales and maximize that between different rationales. Here, $\delta(\hat{r}_{x_m}, \hat{r}_{y_n}) = 1$ if $\hat{r}_{x_m} = \hat{r}_{y_n} \neq 0$ else 0, γ_2 is a coefficient to balance the supervised loss and the unsupervised loss.

In the f_2, its learning objective $\mathcal{L}_{f_2}$ is identical to that used in the fine-tuning phase of the pre-trained language models [54]

$$\mathcal{L}_{f_2} = -\sum_{e \in \mathcal{D}, (G'_X, G'_Y) \in \mathcal{G}'} \sum_{l=1}^{L} \log(P(s_l | \mathbf{s}_{1:l-1})), \tag{20}$$

where $\mathbf{s}$ stands for a sample sequential input of G'_X and G'_Y, which contains L tokens, s_l denotes for the lth token of $\mathbf{s}$, and $\mathbf{s}_{1:l-1}$ denotes the prefix of s_l.

In the f_3, the loss function consists of three parts

$$\mathcal{L}_{f_3} = \sum_{(z, e, \hat{e}) \in \mathcal{D}, (G'_X, G'_Y) \in \mathcal{G}'} \mathcal{L}_{\mathcal{M}} + \gamma_3 (\mathcal{L}_{\mathcal{E}} + \mathcal{L}_{\mathcal{C}}), \tag{21}$$

where $\gamma_3 > 0$ is a coefficient to balance $\mathcal{L}_{\mathcal{M}}$, $\mathcal{L}_{\mathcal{E}}$ and $\mathcal{L}_{\mathcal{C}}$. $\mathcal{L}_{\mathcal{C}}$ is the cross-entropy loss between the ground-truth matching label z and the matching score of rationales and candidate explanations

$$\mathcal{L}_{\mathcal{M}} = -\sum_{k=0}^{2} \delta(z, k) \log\left(P(\hat{z}_k = k | (X, Y))\right), \tag{22}$$

where $\delta(z, k) = 1$ if $z = k$ else 0.

We also design two auxiliary tasks for learning a better representation for rationales and explanations. To ensure that the human-annotated explanation e accurately reflects the matching relation between rationales, the similarity between $[\mathbf{s}_{r^X}, \mathbf{s}_{r^Y}]$ and $\mathbf{s}_e$ should be larger than that between $[\mathbf{s}_{r^X}, \mathbf{s}_{r^Y}]$ and the generated explanation $\mathbf{s}_{\hat{e}_k}$. Therefore, the first task is designed as

$$\mathcal{L}_{\mathcal{E}} = \sum_{k=0}^{2} \max\Big(0, \cos(\mathrm{MLP}[\mathbf{s}_{r^X}; \mathbf{s}_{r^Y}], \mathbf{s}_{\hat{e}_k}) - \cos(\mathrm{MLP}[\mathbf{s}_{r^X}; \mathbf{s}_{r^Y}], \mathbf{s}_e)\Big), \tag{23}$$

where MLP denotes a one-layer multi-layer perceptron. Moreover, inspired by the success of contrastive learning [55] and the observations in [17] that explanations with the same label tend to have the same form, and the form of explanations may be the noise for matching, the second auxiliary task is designed to avoid the classifier only using the form of explanations to infer the matching prediction. Specifically, the candidate explanations in current data are regarded as positive samples $\mathbf{s}_{\hat{e}_k}$, and the explanations with the same label in the mini-batch are regarded as negative samples $\mathbf{s}_{\hat{e}^l_k -}$. Then, the cosine similarity between rationales and positive/negative explanations are calculated and compared

$$\mathcal{L}_{\mathcal{C}} = \sum_{k=0}^{2} \sum_{l} \max\Big(0, \cos(\mathrm{MLP}[\mathbf{s}_{r^X}; \mathbf{s}_{r^Y}], \mathbf{s}_{\hat{e}^l_k -}) - \cos(\mathrm{MLP}[\mathbf{s}_{r^X}; \mathbf{s}_{r^Y}], \mathbf{s}_{\hat{e}_k})\Big), \tag{24}$$

where l is the number of negative samples.

Remark: The efficiency and practicality of GEIOT-Match are comparable to the original IOT-Match [1]. In contrast to previous generation-based explanation models like NILE [17] and LIREx [18], our model primarily introduces the component of rational extraction. The predominant time complexity of rational extraction lies in the Sinkhorn algorithm [45], which has a complexity of $O(N^2)$, where $N = \max(m, n)$, with m and n being the numbers of sentences in case X and case Y, respectively. This complexity is significantly lower than that of the explanation Generation and Matching Prediction parts, whose time complexity is approximately $O(k^2\, d)$, where k represents the sequence length and d represents the dimension of the representation. The aforementioned time complexity is reasonable for most online systems. Regarding practical deployment, we recommend employing our model in the fine-grained ranking phase of legal case retrieval tasks to minimize time expenditures. Despite the associated increase in time costs, our model remains feasible for deployment, showcasing a harmonious balance between effectiveness, explainability, and efficiency.

V. Experiments

In this section, we conduct experiments to evaluate GEIOT-Match for its performance with baseline models, as well as the quality of its explanations and its efficiency with limited rationale alignment labels. The source code, ELAM and eCAIL datasets, and all experiments have been shared at: https://github.com/Jeryi-Sun/GEIOT-Match.

A. Experimental Settings

1) Datasets: To the best of our knowledge, there exist few datasets that contained explanation labels for our explainable legal case matching model. The experiments were conducted based on two publicly available datasets: ELAM [1] and eCAIL [1].

ELAM is an explainable legal case matching dataset. It contains 1,250 source legal cases, each associated with four target cases. Each legal case pair is manually assigned a matching label which is either match (2), partially match (1), or mismatch (0). The ratio of match : partially match : mismatch is about 0.41 : 0.27 : 0.32. Explainable labels such as rationales, their alignments, and free-form explanations are also provided in the dataset.

eCAIL is an extension of CAIL (Challenge of AI in Law) 2021 dataset.[6] In CAIL data, each legal case is associated with tags about private lending. Following the practices in [1], we constructed 1,875 source cases, each associated with four target cases. Each legal case pair is assigned a matching label according to the number of overlapping tags (match if overlapping > 10, mismatch if < 1, and partially match otherwise). The ratio of match, partially match and mismatch is equal. Explainable labels such as rationales, their alignments, and free-form explanations are also constructed following [1].

More statistics of the two datasets can refer to [1].

2) Baselines and Evaluation Metrics: In the experiments, four types of text matching models are selected as baselines.

The first type includes state-of-the-art legal case matching models without explanations:

1) *Sentence-BERT* [56] uses BERT pre-trained on the legal case corpus[7] to encode two cases and uses a MLP to conduct matching.

2) *Lawformer* [7] leverages a Longformer-based [57] pre-trained language model for Chinese legal long documents understanding.

3) *BERT-PLI* [6] uses BERT to capture paragraph-level semantic relations and then aggregates them with RNN and attention.

4) *Thematic Similarity* [22] segments two legal cases into paragraphs and computes the paragraph-level similarities. Maximum or average similarities are used for the overall matching prediction.

The second type of baselines includes the following general text matching models:

1) *RetroMAE* [58] is a retrieval-oriented pre-training paradigm for dense retrieval, which includes a unique MAE workflow, asymmetric model structure, and high masking ratios.

2) *SimLM* [59] is a pre-training method for dense passage retrieval. It uses a simple bottleneck architecture to compress passage information into a dense vector.

The third type of baselines includes the following graph-base text matching models designed for text matching:

1) *Match-GCN* is adjusted from GCN [28] to perform our matching task. We use GCN to learn graph features of pair legal cases and feed concatenated graph features to FNN for the final matching prediction.

2) *Match-CIG* [31] proposes the Concept Interaction Graph to represent an article as a graph of concepts, generates the representation for each vertex, and then uses a GCN to obtain the matching score.

3) *Match-Ignition* [6] plugs PageRank into transformer models to filter information based on a sentence similarity graph for long- form text matching.

The fourth type of baselines includes the following matching models designed for short text matching with explanations:

1) *NILE* [17] adopts GPT2 to generate label-specific explanations for paired sentences, which has three variants that leverage different information to output matching scores: *NILE (Ind)* only uses the generated explanation; *NILE (App)* uses the concatenation of input paired sentences and the generated explanation; and *NILE (Agg)* compares all the generated label-specific explanations.

2) *LIREx* [18] uses an attention mechanism to generate rationale-enabled explanations, which also involves selected explanations to conduct the sentence matching.

Note that both ELAM and eCAIL are in Chinese and do not have precedent information, we do not choose the precedent citation network-based methods [21], [22] as the baselines. Additionally, since NILE and LIREx can generate natural language explanations for matching, we compared GEIOT-Match with them in terms of explanation generation using identical pre-trained language models.

To evaluate the performance of rationale extraction, we also compare GEIOT-Match with the following state-of-the-art rationale extraction models designed for paired documents:

1) *MT-H-LSTM* [60] uses two bi-LSTMs to obtain sentence embeddings and predict the aligned sentences from document pairs.

2) *MLMC* [61] formulates the associative sentence extraction for paired documents as a problem of table filling, in which a matrix is constructed to show whether the sentences are related or not.

3) *DecAtt* [62] adopts attention to indicate the alignments between cross-case sentences. To make fair comparisons, the sentence encoder in DecAtt is set to be identical to that of in GEIOT-Match.

Besides, we compare GEIOT-Match to *IOT-Match* [1] which is proposed in the original SIGIR 2022 paper [1].

Different metrics are adopted to evaluate the different modules of GEIOT-Match. As for rationales extraction and matching prediction, Accuracy, Precision, Recall, and F1 are used. As for natural language explanation generation, the ROUGE score is used because the task is formulated as the Seq2Seq text generation.

3) Hyper-Parameter Settings: All of the hyper-parameters in GEIOT-Match are tuned using grid search on the validation set with Adam [63]. In the rationale extraction, the learning rate η_1 is tuned between $\{1e-4, 1e-3\}$; the batch size n_1 is tuned among $\{32, 64, 128\}$; γ_1 is tuned between $[1, 10]$; γ_2 and ρ is tuned between $[0.1, 1.0]$; the threshold τ_1 for graph edge construction is set to 0.1; the alignment threshold τ_2

[6]Fact Prediction Track data: http://cail.cipsc.org.cn/

[7]https://github.com/thunlp/OpenCLaP

TABLE II
EXPERIMENTAL RESULTS ON ELAM AND ECAIL TEST SETS

Model types	Models	ELAM Acc. (%)	ELAM P. (%)	ELAM R. (%)	ELAM F1 (%)	eCAIL Acc. (%)	eCAIL P. (%)	eCAIL R. (%)	eCAIL F1 (%)
Legal case matching	Sentence-BERT [56]	68.83	69.83	66.88	67.20	71.33	70.83	71.21	70.98
	Lawformer [7]	69.91	72.26	68.34	69.18	70.67	70.20	70.55	69.91
	BERT-PLI [6]	71.21	71.22	71.23	70.88	70.66	70.05	70.54	70.18
	Thematic Similarity (avg) [22]	70.99	71.28	68.97	69.12	71.47	70.88	71.34	71.00
	Thematic Similarity (max) [22]	71.86	71.50	70.07	70.26	68.53	67.25	68.38	67.57
General text matching	RetroMAE	70.08	66.86	65.68	65.12	70.00	70.07	69.55	69.78
	Simlm	70.40	69.66	70.37	69.79	71.73	70.55	71.28	70.73
Graph-based text matching	Match-GCN [28]	71.54	69.92	69.77	69.81	67.73	67.85	6760	67.37
	Match-CIG [31]	69.07	70.80	65.67	61.45	70.77	69.93	69.51	68.04
	Match-Ignition [32]	70.88	69.73	68.33	68.29	70.28	69.14	68.33	68.29
Short text matching with explanations	NILE (Agg) [17]	65.87	65.22	64.89	65.05	71.60	71.44	71.02	70.91
	NILE (App) [17]	68.90	68.90	66.87	67.32	72.53	71.97	71.93	71.95
	NILE (Ind) [17]	69.76	68.30	68.82	68.46	73.33	73.43	72.84	73.05
	LIREx [18]	68.18	68.22	67.34	67.66	70.53	69.68	70.40	69.94
Ours	IOT-Match [1]	73.87	73.02	72.41	72.55	82.00	82.10	81.92	81.90
	GEIOT-Match	**76.55**†	**75.74**†	**75.08**†	**75.34**†	**83.20**†	**82.92**†	**83.11**†	**82.85**†

† Indicates the statistically significant difference between the performance of all baseline models and that of GEIOT-match (p-value < 0.05).
The bold values represent the maximum values of their respective columns.

for ELAM and eCAIL are tuned between $[1e-3, 1e-2]$ and $[1e-3, 5e-3]$, respectively; and the entropic regularizer λ_1 and λ_2 is tuned among $[0.1, 1.0]$; the affinity matrix coefficient ϵ is tuned among $\{0, -10, -50, -100, -200\}$. In the natural explanation generation, the hyper-parameters are set according to those reported in [64]: the learning rate η_2 is set as $2e-5$; the batch size n_2 is set as 2; In the matching, the learning rate η_3 is tuned between $\{2e-5, 2e-4\}$; the batch size n_3 is tuned between $\{4, 8\}$, and γ_3 is tuned among $\{1, 10, 20\}$.

B. Matching Accuracy

We first study the matching performance of our proposed GEIOT-Match. Table II presents the matching performances of GEIOT-Match and the baselines in terms of four evaluation metrics on ELAM and eCAIL. All the methods are trained ten times and the averaged results are reported. Based on the results, we summarize our observations as follows: (1) GEIOT-Match consistently and significantly outperforms all of the baselines on two datasets in terms of all metrics, indicating the effectiveness of GEIOT-Match in enhancing the matching accuracy. (2) Compared to short text matching with explanation models which involve all sentences in a paired legal case during the matching, GEIOT-Match enjoys the advantages from the extracted rationales and achieves consistent improvements on two datasets. The result indicates that the rationale extraction module in GEIOT-Match accurately identified the rationales and filtered out the noise sentences from legal cases. (3) Compared to existing legal case matching models, general text matching models and graph-based text matching models that cannot provide matching explanations, GEIOT-Match also achieves consistent improvements on both datasets. The results indicate that the natural language explanations generated by GEIOT-Match are helpful for legal case matching. (4) Compared to IOT-Match, GEIOT-Match achieves consistent improvements on two datasets, which indicates GEIOT-Match can leverage the semi-structured nature of legal cases to achieve better matching predictions.

TABLE III
PLAUSIBILITY OF EXTRACTED RATIONALES ON ELAM AND ECAIL TEST SETS IN TERMS OF EXTRACTION ACCURACY

Models	ELAM Acc. (%)	eCAIL Acc. (%)
MT-H-LSTM [60]	68.91	95.18
MLMC [61]	68.37	95.30
DecAtt [62]	83.09	94.33
OT [35]	83.09	90.97
IOT-Match [1]	86.82	96.26
GEIOT-Match	**87.36**	**96.30**

The bold values represent the maximum values of their respective columns.

C. Quality of Rationales and Explanations

The major superiority of GEIOT-Match compared to existing legal case matching models is that GEIOT-Match is able to extract rationales and generate explanations for the matching prediction. In this subsection, we conduct experiments to assess the quality of the extracted rationales and the generated natural language explanation by GEIOT-Match. Following [65], we adopt plausibility and faithfulness as the metrics. Plausibility measures how well the explanation aligns with human annotations, and faithfulness measures the degree to which the explanation influences the corresponding predictions.

1) Quality of the Extracted Rationales: In terms of *plausibility*, we compare the rationales extracted by GEIOT-Match and baseline models with human annotations on ELAM and eCAIL. As shown in Table III, the rationales extracted by GEIOT-Match are more consistent with human annotations, especially on the ELAM dataset where the rationales are more diverse (three types of rationales). We also compare the original GEIOT-Match with a modified one with GEIOT ablated but forward OT kept, denoted as OT in Table III. Table III shows that the extraction accuracy drops if we remove GEIOT from GEIOT-Match. The results indicate the effectiveness of GEIOT in learning the adaptive cross-graph sentence affinity and predicting the rationale alignment. Furthermore, our GEIOT-Match outperforms the original

TABLE IV
FAITHFULNESS OF THE EXTRACTED RATIONALES AND THE GENERATED EXPLANATION ON ELAM AND ECAIL TEST SETS

Input	ELAM Acc. (%)	ELAM P. (%)	ELAM R. (%)	ELAM F1 (%)	CAIL Acc. (%)	CAIL P. (%)	CAIL R. (%)	CAIL F1 (%)
$a\backslash r$	71.28	69.67	67.72	67.96	69.86	70.07	70.40	69.89
a	68.83	69.83	66.88	67.20	71.33	70.83	71.21	70.98
r	70.68	69.11	68.48	68.45	71.60	71.30	71.80	71.45
e	71.14	70.02	68.19	68.18	70.66	70.47	70.55	70.29
$a\backslash r + e$	73.69	74.04	71.13	71.65	77.33	77.09	77.61	77.24
$a + e$	75.90	**76.12**	74.93	**75.40**	78.26	78.60	78.19	78.21
$r + e$	**76.55**	75.74	**75.08**	75.34	**83.20**	**82.92**	**83.11**	**82.85**

The column "input" denotes GEIOT-match with different inputs.
The bold values represent the maximum values of their respective columns.

model IOT-Match demonstrating that GEIOT can fully use structural and sequential information by integrating WD, GWD, and OPWD into the IOT framework. In terms of *faithfulness*, we conduct experiments to measure the degree to which the extracted rationales influence the final matching. Specifically, we test the matching performance of GEIOT-Match with explanations and GEIOT-Match without explanations respectively under three conditions: using all sentences as the input (respectively denoted as "GEIOT-Match $(a + e)$" and "GEIOT-Match (a)"), using rationale extracted by GEIOT-Match as the input (respectively denoted as "GEIOT-Match $(r + e)$" and "GEIOT-Match (r)"), and using sentences except those extracted by GEIOT-Match as the input (respectively denoted as "GEIOT-Match $(a\backslash r + e)$" and "GEIOT-Match $(a\backslash r)$").[8] From the results reported in Table IV, we find that the rationales extracted by GEIOT-Match play a critical role in legal case matching. Specifically, if the extracted rationales are removed from a model's input (GEIOT-Match($a\backslash r + e$) or GEIOT-Match($a\backslash r$)), the matching accuracy of the model drops dramatically. In addition, in eCAIL where the legal cases are extremely lengthy, if all sentences are used as a model's input (GEIOT-Match($a + e$) or GEIOT-Match(a)), the model's accuracy still drops to some extent because of the noise from other sentences. On ELAM, the performance of using rationales as the only input is competitive with that of using all sentences. This result indicates the rationales extracted by GEIOT-Match already provide sufficient legal semantics for case matching. Based on the above analysis, we conclude that GEIOT-Match is capable of accurately extracting faithful rationales for legal case matching.

2) Quality of the Generated Explanation: In terms of *plausibility*, we compare the natural language explanation generated by GEIOT-Match to those generated by NILE [17] and LIREx [18]. Since both ELAM and eCAIL have human-annotated explanations for the matching labels, the popular metrics in machine translation such as ROUGE-1, ROUGE-2, and ROUGE-L are used to evaluate the plausibility. As shown in Table V, the natural language explanations generated by GEIOT-Match are more consistent with human annotations than those generated by NILE and LIREx. Furthermore, our model overcomes the strong baseline IOT-Match, demonstrating the viability of combining graph linearization with type-aware adapters for generating plausible explanations.

Moreover, we also conduct human evaluations to test the quality of the generated explanations. Following [18], we randomly sampled 50 examples respectively from ELAM and eCAIL, and ask two annotators to answer the questions that whether the generated explanation and the label explanation convey the same meaning. Each annotator was provided with the context (legal cases, rationales, explanations), and asked to label them as 1 if they agree to the question, or 0 otherwise. As shown in Table VI, GEIOT-Match obtains a high relevance score between the generated explanations and label explanations. The result verifies the effectiveness of GEIOT-Match in generating plausible explanations.

In terms of *faithfulness*, we conduct experiments to measure the degree to which the generated explanation influences the final matching. Specifically, we compare the performance among GEIOT-Match using the rationales only (GEIOT-Match(r)), using the explanation only (GEIOT-Match(e)), and using rationales and explanations (GEIOT-Match($r + e$)). From the results reported in Table IV, we find: (1) on both ELAM and eCAIL, GEIOT-Match $(r + e)$ performs the best, indicating that the natural language explanations generated by GEIOT-Match contributed to the matching prediction; (2) GEIOT-Match (e) performs better than GEIOT-Match (r) on ELAM, verifying the faithfulness of the generated explanation. The result also indicates that the explanations on ELAM are more sufficient for the matching prediction than the extracted rationales. (3) GEIOT-Match (e) performs worse than GEIOT-Match (r) on eCAIL. We analyze the reasons and find that the labeled explanations on eCAIL are the concatenations of the rationale sentences. Such labeled explanations are not coherent enough and may harm the generated explanations.

Furthermore, we present a case study of the explanation generated by GEIOT-Match in Fig. 7. The right side of the figure (where the red region indicates the difference between the ground truth and predicted labels) shows that the extracted rationale from GEIOT exhibits high accuracy. Moreover, the generated rationale aligns closely with the human-annotated explanation. This demonstrates the effectiveness and quality of the GEIOT-Match in generating explanations. More case studies about the partially match and mismatch cases are shown in https://github.com/Jeryi-Sun/GEIOT-Match.

D. Robustness Under Limited Labels

One advantage of GEIOT-Match is its capability of learning to extract the pro and con rationales from human-labeled rationale alignments in a semi-supervised manner, because, in real practice, manually labeling the rationale alignments is expensive and time-consuming.

We conduct experiments to test the rationale extraction accuracy w.r.t. different amounts of labeled alignments. Specifically, we configure GEIOT-Match to extract rationales given different ratios of labeled alignments $\hat{\mathbf{A}}$ in (8) (from 0% to 100% where

[8]This experiment is also the ablation study of GEIOT-Match. Specifically, "GEIOT-Match $(a + e)$" represents the results obtained after removing the Rational Extraction Module, while "GEIOT-Match (r)" denotes the results obtained after removing the Explanation Generation Module.

TABLE V
PLAUSIBILITY OF GENERATED EXPLANATIONS ON ELAM AND ECAIL TEST SETS IN TERMS OF ROUGE SCORES

Models	ELAM ROUGE-1 (%) $z=2$ $z=1$ $z=0$	ELAM ROUGE-2 (%) $z=2$ $z=1$ $z=0$	ELAM ROUGE- L (%) $z=2$ $z=1$ $z=0$	eCAIL ROUGE-1 (%) $z=2$ $z=1$ $z=0$	eCAIL ROUGE-2 (%) $z=2$ $z=1$ $z=0$	eCAIL ROUGE- L (%) $z=2$ $z=1$ $z=0$
NILE [17]	73.40 70.96 69.93	58.47 55.57 56.08	69.87 65.70 66.84	73.40 70.96 69.93	58.47 55.57 56.08	69.87 65.70 66.84
LIREx [18]	74.15 71.61 70.97	59.78 56.36 56.88	70.89 66.22 67.41	80.35 74.36 74.46	72.00 67.35 63.58	76.84 74.28 66.98
IOT-Match [1]	75.55 73.64 75.18	60.97 58.25 61.84	72.79 68.85 72.54	83.54 76.59 91.21	76.48 69.46 87.03	80.37 76.16 86.91
GEIOT-Match	**75.98 74.38 76.45**	**62.91 59.90 63.04**	**73.15 69.62 73.62**	**85.19 77.68 91.57**	**79.13 70.90 87.57**	**82.10 77.34 88.08**

The bold values represent the maximum values of their respective columns.

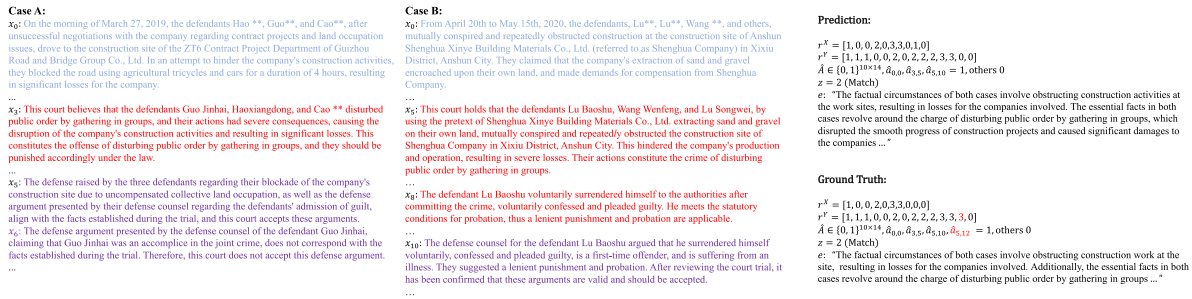


Fig. 7. Case study of the explanation generated by GEIOT-Match. The left and middle sections of the figure show the content of the cases (translated from Chinese). The right section shows the comparison between the ground truth and the GEIOT prediction results.

TABLE VI
HUMAN EVALUATIONS OF THE EXPLANATION QUALITY OVER 50 RANDOMLY SAMPLED DATA FROM ELAM AND ECAIL BY TWO ANNOTATORS WITH THE INTER-RATER AGREEMENT OF 0.95

	NILE [17]	LIREx [18]	IOT-Match [1]	GEIOT-Match
ELAM	35	41	46	**48**
eCAIL	36	38	44	**45**

The bold values represent the maximum values of their respective columns.

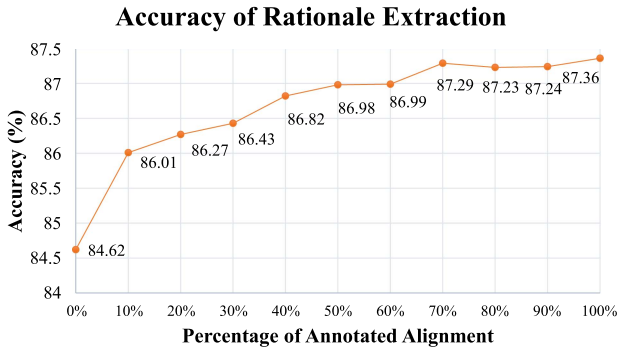


Fig. 8. Rationale extraction accuracy of GEIOT-Match w.r.t. different percentages of labeled alignments.

0% means no labels available, and 100% means fully supervised learning). Fig. 8 illustrates the extraction accuracy w.r.t. the ratio of labeled alignments on ELAM data. We find that GEIOT-Match shows competitive performances when only 10%\,20% of the labeled alignments are involved in learning. The results indicate that with only a small fraction of the alignment labels, GEIOT-Match can still learn the cross-case sentence-level affinity matrix $\mathbf{C}$ with high accuracy, and accurately extract the pro and con rationales.

TABLE VII
AVERAGE ONLINE INFERENCE TIME PER CASE PAIR

Models w/o explanation	**Sentence-BERT**	**Match-GCN**	**RetroMAE**	**BERT-PLI**
Time Cost (s)	0.0502	0.0466	0.0511	0.1034
Models w/ explanation	**NILE**	**LIREx**	**IOT-Match**	**GEIOT-Match**
Time Cost (s)	0.1214	0.1239	0.1173	0.1246

E. Running Time Analysis

In Table VII, we present an analysis of the average online inference time for matching models, both with and without explanations. It is observed that models integrated with explanations inherently necessitate a longer inference time due to the additional process of explanation generation. Among the explainable models, GEIOT-Match demonstrates a higher accuracy and improved explanation quality. Notably, its inference time remains on par with other models in the same category. While there is an observable increase in time, GEIOT-Match maintains a balance of effectiveness, explainability, and efficiency, suggesting its potential suitability for practical deployment.

VI. CONCLUSION

This paper proposes a novel explainable legal case matching model which represents the legal cases as a heterogeneous graph and learns an inverse optimal transport model on the graph to extract rationales. The prediction of the matching results and generation of the explanations can be conducted based on the rationales. The proposed method, called GEIOT-Match, has the advantage of explicitly representing the structures of legal cases and involving the associated law articles for matching and explanation. Comprehensive experimental results showed that GEIOT-Match can consistently outperform the state-of-the-art baselines in terms of matching accuracy. The empirical analysis also verified that the extracted rationales and the generated

explanations are not only consistent with human annotations but also faithful to the final matching predictions.

ACKNOWLEDGMENTS

This is an extension of our previous work (SIGIR 2022) [1]. We make the following extensions: 1) we propose to construct a heterogeneous graph to explicitly model the semi-structured nature of legal cases and the case-case and case-law article correlations; 2) Based on the inverse optimal transport framework, we propose a novel graph optimal transport-based model called GEIOT-Match to learn from the heterogeneous graph; and 3) we conducted extensive experiments and analysis on GEIOT-Match in Section V.

REFERENCES

[1] W. Yu et al., "Explainable legal case matching via inverse optimal transport-based rationale extraction," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 657–668.
[2] Y. Zeng, R. Wang, J. Zeleznikow, and E. A. Kemp, "Knowledge representation for the intelligent legal case retrieval," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Melbourne, Australia, 2005, pp. 339–345.
[3] M. Saravanan, B. Ravindran, and S. Raman, "Improving legal information retrieval using an ontological framework," *Artif. Intell. Law*, vol. 17, no. 2, pp. 101–124, 2009.
[4] T. Bench-Capon et al., "AI and law," *Artif. Intell. Law*, vol. 20, no. 3, pp. 215–319, 2012.
[5] S. Kumar, P. K. Reddy, V. B. Reddy, and A. Singh, "Similarity analysis of legal judgments," in *Proc. 4th Bangalore Annu. Compute Conf.*, 2011, Art. no. 17.
[6] Y. Shao et al., "BERT-PLI: Modeling paragraph-level interactions for legal case retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 3501–3507.
[7] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021.
[8] X. Jiang, H. Ye, Z. Luo, W. Chao, and W. Ma, "Interpretable rationale augmented charge prediction system," in *Proc. 27th Int. Conf. Comput. Linguistics: Syst. Demonstrations*, Santa Fe, New Mexico: Association for Computational Linguistics, 2018, pp. 146–151.
[9] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer, "An information bottleneck approach for controlling conciseness in rationale extraction," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 1938–1952.
[10] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, and P. Malakasiotis, "Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2021, pp. 226–241.
[11] M. Black, "Why cannot an effect precede its cause?," *Analysis*, vol. 16, no. 3, pp. 49–58, 1956.
[12] X. Liu, D. Yin, Y. Feng, Y. Wu, and D. Zhao, "Everything has a cause: Leveraging causal inference in legal text analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2021, pp. 1928–1941.
[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: "Preparing the muppets for court"," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 2898–2904.
[14] K. Atkinson, T. J. M. Bench-Capon, and D. Bollegala, "Explanation in AI and law: Past, present and future," *Artif. Intell.*, vol. 289, 2020, Art. no. 103387.
[15] L. Sha, O. Camburu, and T. Lukasiewicz, "Learning from the best: Rationalizing predictions by adversarial information calibration," in *Proc. AAAI Conf. Artif. Intell.*, AAAI Press, 2021, pp. 13771–13779.
[16] K. Luu, X. Wu, R. Koncel-Kedziorski, K. Lo, I. Cachola, and N. A. Smith, "Explaining relationships between scientific documents," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, Association for Computational Linguistics, 2021, pp. 2130–2144.
[17] S. Kumar and P. Talukdar, "NILE: Natural language inference with faithful natural language explanations," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2020, pp. 8730–8742.
[18] X. Zhao and V. G. V. Vydiswaran, "LIREx: Augmenting language inference with relevant explanations," in *Proc. AAAI Conf. Artif. Intell.*, AAAI Press, 2021, pp. 14532–14539.
[19] A. L. Monroy, H. Calvo, A. F. Gelbukh, and G. G. Pacheco, "Link analysis for representing and retrieving legal information," in *Proc. 14th Int. Conf. Comput. Linguistics Intell. Text Process.*, Springer, 2013, pp. 380–393.
[20] A. Minocha, N. Singh, and A. Srivastava, "Finding relevant indian judgments using dispersion of citation network," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1085–1088.
[21] P. Bhattacharya, K. Ghosh, A. Pal, and S. Ghosh, "Hier-SPCNet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, New York, NY, USA, 2020, pp. 1657–1660.
[22] P. Bhattacharya et al., "Methods for computing legal document similarity: A comparative study," 2020, *arXiv:2004.12307v1*.
[23] A. Bibal, M. Lognoul, A. De Streel, and B. Frénay, "Legal requirements on explainability in machine learning," *Artif. Intell. Law*, vol. 29, no. 2, pp. 149–169, 2021.
[24] F. Doshi-Velez et al., "Accountability of AI under the law: The role of explanation," 2017, *arXiv:1711.01134*.
[25] H. Ye, X. Jiang, Z. Luo, and W. Chao, "Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, New Orleans, Louisiana, USA: Association for Computational Linguistics, 2018, pp. 1854–1864.
[26] L. Liu, W. Zhang, J. Liu, W. Shi, and Y. Huang, "Interpretable charge prediction for legal cases based on interdependent legal information," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.
[27] H. Zhong, Y. Wang, C. Tu, T. Zhang, Z. Liu, and M. Sun, "Iteratively questioning and answering for interpretable legal judgment prediction," in *Proc. AAAI Conf. Artif. Intell.*, AAAI Press, 2020, pp. 1250–1257.
[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, Apr. 24-26, 2017.
[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, Apr. 30-May 3, 2018.
[30] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proc. Web Conf.*, 2020, pp. 2704–2710.
[31] B. Liu et al., "Matching article pairs with graphical decomposition and convolutions," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2019, pp. 6284–6294.
[32] L. Pang, Y. Lan, and X. Cheng, "Match-ignition: Plugging PageRank into transformer for long-form text matching," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 1396–1405.
[33] J. S. Dhani, R. Bhatt, B. Ganesan, P. Sirohi, and V. Bhatnagar, "Similar cases recommendation using legal knowledge graphs," 2021, *arXiv:2107.04771*.
[34] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer, 2009.
[35] G. Peyré et al., "Computational optimal transport: With applications to data science," *Found. Trends Mach. Learn.*, vol. 11, no. 5/6, pp. 355–607, 2019.
[36] D. Alvarez-Melis, S. Jegelka, and T. S. Jaakkola, "Towards optimal transport with global invariances," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1870–1879.
[37] D. Alvarez-Melis and T. Jaakkola, "Gromov-wasserstein alignment of word embedding spaces," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1881–1890.
[38] H. Xu, D. Luo, H. Zha, and L. Carin, "Gromov-wasserstein learning for graph matching and node embedding," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6932–6941.
[39] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu, "Graph optimal transport for cross-domain alignment," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1542–1553.
[40] H. Xu, W. Wang, W. Liu, and L. Carin, "Distilled wasserstein learning for word embedding and topic modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1723–1732.
[41] S. Chakraborty, D. Paul, and S. Das, "Hierarchical clustering with optimal transport," *Statist. Probability Lett.*, vol. 163, 2020, Art. no. 108781.
[42] W. Yu et al., "Wasserstein distance regularized sequence representation for text matching in asymmetrical domains," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Association for Computational Linguistics, 2020, pp. 2985–2994.

[43] W. Yu, C. Xu, J. Xu, L. Pang, and J.-R. Wen, "Distribution distance regularized sequence representation for text matching in asymmetrical domains," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 721–733, 2022.

[44] W. Yu, L. Pang, J. Xu, B. Su, Z. Dong, and J.-R. Wen, "Optimal partial transport based sentence selection for long-form document matching," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 2363–2373.

[45] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2013, pp. 2292–2300.

[46] Y. Ma et al., "LeCaRD: A legal case retrieval dataset for chinese law system," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 2342–2348.

[47] B. Su and G. Hua, "Order-preserving optimal transport for distances between sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2961–2974, Dec. 2019.

[48] A. Dupuy, A. Galichon, and Y. Sun, "Estimating matching affinity matrix under low-rank constraints," *Econometric Model.: Theor. Issues Microeconometrics eJournal*, 2016.

[49] R. Li, X. Ye, H. Zhou, and H. Zha, "Learning to match via inverse optimal transport," *J. Mach. Learn. Res.*, vol. 20, pp. 80:1–80:37, 2019.

[50] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.

[51] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 4th Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Juan, Puerto Rico, May 2-4, 2016.

[52] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.

[53] P. Ke et al., "JointGT: Graph-text joint representation learning for text generation from knowledge graphs," in *Proc. Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 2526–2538.

[54] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[55] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2021, pp. 1065–1072.

[56] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empir. Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992.

[57] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[58] S. Xiao, Z. Liu, Y. Shao, and Z. Cao, "RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Association for Computational Linguistics, 2022, pp. 538–548.

[59] L. Wang et al., "SimLM: Pre-training with representation bottleneck for dense passage retrieval," in *Proc. 61st Annual Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada, Jul. 9-14, 2023, pp. 2244–2258. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.125

[60] L. Cheng, L. Bing, Q. Yu, W. Lu, and L. Si, "APE: Argument pair extraction from peer review and rebuttal via multi-task learning," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 7000–7011.

[61] L. Cheng, T. Wu, L. Bing, and L. Si, "Argument pair extraction via attention-guided multi-layer multi-cross encoding," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2021, pp. 6341–6353.

[62] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 2249–2255.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7-9, 2015.

[64] J. Su, "T5 Pegasus - Zhuiyiai," Tech. Rep., 2021. [Online]. Available: https://github.com/ZhuiyiTechnology/t5-pegasus

[65] J. DeYoung et al., "ERASER: A benchmark to evaluate rationalized NLP models," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4443–4458.

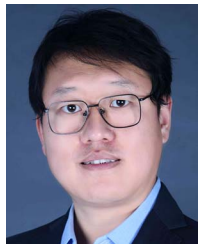
Zhongxiang Sun is currently working toward the PhD degree in Artificial Intelligence with the Gaoling School of Artificial intelligence, Renmin University of China. His current research interests mainly include legal information retrieval and recommender systems.

Weijie Yu received the PhD degree from the School of Information, Renmin University of China, in 2023. He is currently an assistant professor with the School of Information Technology and Management, University of International Business. Her research interests include text matching and legal retrieval.

Zihua Si is currently working toward the ME degree in Artificial Intelligence with the Gaoling School of Artificial intelligence, Renmin University of China. His current research interests mainly include sequential recommender systems.

Jun Xu (Member, IEEE) is a professor with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests focus on learning to rank and semantic matching in web search. He served or is serving as SPC for SIGIR, WWW, and AAAI, editorial board member for *Journal of the Association for Information Science and Technology*, and associate editor for *ACM Transactions on Intelligent Systems and Technology*. He has won the Test of Time Award Honorable Mention in SIGIR (2019), Best Paper Award in AIRS (2010) and Best Paper Runner-up in CIKM (2017).

Zhenhua Dong received the PhD degree in computer science from Nankai University, China, in 2012. He is currently a principal researcher of Huawei Noah's Ark Lab. His research interests include recommender system, counterfactual learning, causal information retrieval, and their applications. He has published papers in refereed conferences and journals such as AAAI, ICDE, SIGIR, WWW, *IEEE Transactions on Knowledge and Data Engineering*, etc.



Xu Chen received the PhD degree from Tsinghua University, China. Before joining Renmin University of China, he was a postdoc researcher with University College London, U.K. In the period from March to September 2017, he was studying with the Georgia Institute of Technology, as a visiting scholar. His research mainly focuses on the recommender system, reinforcement learning, and causal inference.

Hongteng Xu (Member, IEEE) received the PhD degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, in 2017. He is currently an associate professor (Tenure-Track) with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests include machine learning and its applications, especially optimal transport theory, sequential data modeling and analysis, deep learning techniques, and their applications in computer vision and data mining.

Ji-Rong Wen is a professor of the Renmin University of China (RUC). He is also the dean of the School of Information and executive dean of the Gaoling School of Artificial Intelligence with RUC. His main research interests include information retrieval, data mining, and machine learning. He was a senior researcher and group manager of the Web Search and Mining Group with Microsoft Research Asia (MSRA).