applied sciences

MDPI

Article

A Construction and Representation Learning Method for a Traffic Accident Knowledge Graph Based on the Enhanced TransD Model

Xiaojia Liu [1,2], Haopeng Wu [1], Dexin Yu [1,*], Yunjie Chen [1] and Hao Wu [1]

[1] College of Navigation, Jimei University, Xiamen 361021, China; happylxj1314@163.com (X.L.); 202312861043@jmu.edu.cn (H.W.); 202211823018@jmu.edu.cn (Y.C.); 202314823002@jmu.edu.cn (H.W.)
[2] Marine Traffic Safety Institute, Jimei University, Xiamen 361021, China
* Correspondence: yudx@jmu.edu.cn

Abstract: With rapid urbanization and surging traffic volumes, traffic accident data have become high-dimensional, multi-source, heterogeneous, and spatiotemporally dynamic, posing challenges for traditional statistical methods and machine learning models to simultaneously account for data heterogeneity and nonlinear interactions. Knowledge graphs, by constructing structured semantic networks that integrate accident events, participants, environmental factors, and other multidimensional elements, inherently support multi-source information fusion and reasoning. In this study, following a top-down ontology design principle, we construct a California Traffic Accident Knowledge Graph (TAKG) encompassing over one hundred elements, and propose an enhanced TransD embedding model. Our model introduces entity–attribute projection vectors into the dynamic mapping mechanism to explicitly encode domain attributes, and designs a dual-limit scoring loss function to independently regulate the positive and negative sample boundaries. Experimental results demonstrate that our method significantly outperforms traditional translation-based models on the self-built TAKG as well as on the FB15K-237 and WN18RR benchmark datasets. This research provides a solid data foundation and algorithmic support for downstream traffic accident risk prediction and intelligent traffic safety management.

Keywords: traffic safety; traffic accident; data mining; traffic accident knowledge graph; risk prediction

Academic Editor: Arkadiusz Gola

Received: 27 April 2025
Revised: 22 May 2025
Accepted: 26 May 2025
Published: 27 May 2025

Citation: Liu, X.; Wu, H.; Yu, D.; Chen, Y.; Wu, H. A Construction and Representation Learning Method for a Traffic Accident Knowledge Graph Based on the Enhanced TransD Model. *Appl. Sci.* **2025**, *15*, 6031. https://doi.org/10.3390/app15116031

1. Introduction

In recent years, the rapid advance of urbanization and the widespread adoption of diverse transport modes have led to traffic accident data characterized by high dimensionality, multi-source heterogeneity, and complex structures [1], posing unprecedented challenges for accident analysis and prevention. These data—sourced from traffic surveillance systems, sensor networks, and social media—often suffer from uneven quality, missing values, redundancy, and noise [2], greatly complicating cleaning and preprocessing. Furthermore, significant discrepancies in temporal synchronization and spatial alignment across data sources make effective information fusion a critical technical hurdle. The accident process itself exhibits pronounced spatiotemporal dynamics and uncertainty, involving complex nonlinear interactions among driver behavior, road environment, and meteorological conditions, which further impedes the development of high-precision predictive models and effective countermeasures. To achieve substantive progress in accident prediction, prevention strategy formulation, and risk assessment, an urgent need exists for a framework that

can both integrate multi-source data and capture fine-grained semantic representations, thereby more accurately supporting risk prediction and safety decision-making.

In earlier research on traffic accident data processing, traditional statistical analysis methods predominated. For example, Yu [3] classified causes of death according to ICD-10 and employed Excel 2016 and SPSS 19.0 for data analysis, revealing that motor vehicle accidents accounted for approximately 20.90% of fatalities in the Haidian District. Mao et al. [4] examined shortcomings in road traffic accident data mining—specifically incomplete data, unreasonable criteria for identifying high-incidence locations, and a narrow dimensional scope for causal analysis—and proposed an optimized set of data-collection items together with a GIS-based, multidimensional method for accident cause analysis. Xu et al. [5], using a large dataset of accidents in Guiyang, developed a novel practical numerical model grounded in rough-set theory to assess the significance of objective factors and thus mined the key determinants of road traffic accidents. Although these conventional approaches facilitate straightforward processing of traffic accident data, they typically rely on linear relationship assumptions and are therefore ill-suited to capture the complex nonlinear correlations and multivariate interactions inherent in accident data. Given the high dimensionality, multi-source heterogeneity, and spatiotemporal dynamics of such datasets, traditional models often fail to fully reflect the underlying mechanisms of accident occurrence. Moreover, they tend to be sensitive to noise, missing values, and outliers, and their robustness and predictive accuracy are compromised when fusing data from disparate systems, limiting their utility for effective accident prevention and risk assessment.

To address the limitations of traditional statistical methods, researchers have adopted machine learning and deep learning techniques—including Bayesian frameworks, support vector machines (SVMs), and Long Short-Term Memory (LSTM) networks. Ye et al. [6] applied the Apriori algorithm, Bayesian theory, and fuzzy clustering for large-scale data mining to impute missing accident attributes, assess accident severity, and classify risk levels, validating their approach on Shenzhen traffic accident records from 2014 to 2016. Fan et al. [2] developed a hybrid SVM model that integrates multi-source influencing factors and a deep-neural-network-based dynamic adaptive recognition algorithm to identify urban accident hotspots. Ma et al. [7] employed machine learning and a stacked sparse autoencoder (SSAE) to predict the severity of traffic accident injuries. Hadjidimitriou et al. [8] designed a machine learning system for classifying the severity of powered two-wheeler accidents, achieving approximately 90% accuracy and recall by leveraging features available at the moment of collision. Although these advanced methods can efficiently process traffic accident datasets, they still confront challenges such as limited data volume, imbalanced class distributions, and interference from noisy inputs, which impede the capture of complex spatiotemporal patterns and multifactor interactions. High-dimensional and heterogeneous data fusion often suffers from the “curse of dimensionality” and feature redundancy, leading to degraded predictive performance. Moreover, these approaches frequently lack transparency, hindering their practical deployment for causal analysis and countermeasure planning; their sensitivity to hyperparameters and substantial computational demands further complicate real-world implementation.

Knowledge graphs are, at their core, semantic networks that interconnect real-world entities and the relationships among them, wherein nodes represent entities and edges denote various semantic relations [9]. In scholarly literature, numerous definitions of knowledge graphs have been proposed, tailored to specific application scenarios and technical contexts [10]. Broadly speaking, a knowledge graph is a large-scale factual database with a graph structure, comprising vast numbers of triples formed by entities (nodes) and their relations (edges), and semantically modeling real-world phenomena. Traditional statistical and machine learning approaches exhibit limitations in uncovering

the multi-level causal relationships underlying traffic accidents, whereas knowledge graph technology, by integrating diverse information on accidents, environment, vehicles, and driver behavior, holds promise for mining deep semantic associations within the data. Indeed, traffic accident data not only encapsulate rich spatiotemporal information but also involve multiple influencing factors that typically interact in complex nonlinear ways. Moreover, the heterogeneity of data sources, along with inconsistencies in format and quality, further complicates data fusion and preprocessing. The temporal dynamics and uneven geographic distribution of accident data present severe challenges for pattern recognition and causal inference. Knowledge graph technologies have already proven indispensable in Internet-scale applications such as search engines and intelligent question answering, and they are beginning to see preliminary adoption in domains including finance [11], power systems [12], and healthcare [13].

Although conventional statistical analyses and machine learning techniques have achieved notable results in processing traffic accident data, their ability to capture the nonlinear interactions inherent in high-dimensional heterogeneous data and the spatiotemporal dynamics of accidents remains insufficient [14], making them ill-suited for fine-grained risk assessment and causal analysis. At the same time, existing knowledge graph embedding models often exhibit limited representational capacity or parameter redundancy when handling complex one-to-many, many-to-one, and many-to-many mapping relationships [15], which hampers their broader adoption in the traffic safety domain.

On the basis of the foregoing theoretical gaps, this study addresses the following research questions: (1) How can one construct a traffic accident knowledge graph encompassing multi-level, multi-source heterogeneous elements—integrating accident events, environmental factors, participants, and victims—so as to fully represent the multidimensional nature of traffic incidents? (2) What limitations do prevailing translation-based embedding methods (TransE, TransH, TransR, and TransD) exhibit when modeling one-to-many, many-to-one, and many-to-many relations within a traffic accident knowledge graph? (3) To what extent does the proposed enhanced TransD model generalize on both a self-constructed California Traffic Accident Knowledge Graph (TAKG) and the benchmark datasets FB15K-237 and WN18RR?

To answer these questions, the main contributions of this paper are as follows: (1) A California Traffic Accident Knowledge Graph (TAKG) comprising more than 110 entity and relation types is constructed following a top-down ontology design principle, thereby filling the gap in fine-grained traffic accident graph resources. (2) An enhanced TransD embedding model is proposed, which introduces entity attribute projection vectors into the dynamic mapping mechanism and employs a dual-limit loss to efficiently model complex multi-mapping relations. (3) Extensive benchmark comparisons and ablation studies on FB15K-237, WN18RR, and the self-constructed TAKG systematically demonstrate robust improvements across multiple key metrics. Given TransD's advantageous balance between expressive capacity and computational efficiency in handling complex multi-mapping relations, it was selected as the baseline for enhancement. The TAKG was built from the publicly available California traffic accident dataset on Kaggle, which was manually cleaned and field-standardized.

The remainder of this paper is organized as follows: In Section 2, we review related works; Section 3 presents the methodology for constructing the traffic accident knowledge graph; Section 4 details the theoretical foundations and algorithmic implementation of the enhanced TransD model; Section 5 describes the experimental setup, presents results, and provides analysis; Section 6 discusses the implications of our findings; finally, Section 7 concludes the paper and outlines future research directions.

2. Related Works

In the field of traffic accident data analysis, the evolution of distinct methodological streams reflects researchers' deepening understanding of data characteristics and modeling requirements: traditional statistical analyses focus on descriptive statistics and parameter estimation for accident frequency and influencing factors; machine learning approaches introduce nonlinear models to enhance analytical accuracy; and knowledge graph techniques further integrate multi-source heterogeneous information while emphasizing semantic relationship representation and inference.

(1) Traditional Statistical Analysis Methods

In the early stage of traffic accident research, we relied primarily on parametric models and clustering analyses to provide macroscopic characterization and causal inference for large-scale accident datasets. For example, Qiu et al. [16] proposed an improved DBSCAN clustering algorithm that intelligently identifies accident hot spots by carefully selecting ε (epsilon) and minPts; Li et al. [17] built a decision tree model using 2006–2013 Wenli Expressway accident data—incorporating time, roadway geometry, and driver characteristics as predictors—to quantify each factor's impact; Depaire et al. [1] partitioned a heterogeneous traffic accident dataset into seven clusters corresponding to distinct accident types and then analyzed injury severities within each cluster. Traditional statistical analysis methods offer strong interpretability in identifying accident hot spots and quantifying risk factors, but they struggle to capture the complex correlations in high-dimensional, heterogeneous data.

(2) Machine Learning Methods

As both the volume and dimensionality of traffic data surged, unsupervised and supervised learning algorithms were adopted to uncover richer nonlinear patterns and improve analytical accuracy. Lv et al. [18] used support vector machines (SVMs) with real-time traffic data to distinguish patterns that lead to accidents from those that do not; Hu et al. [19] introduced a ConvLSTM U-shaped network with densely connected convolutional layers (BCDU-Net) to mine hidden regularities and extract deep spatiotemporal features from accident sequences; Shokry et al. [20] applied k-means clustering to five years of aggregated and disaggregated Egyptian accident data to optimize data utilization; Wang et al. [21] developed an ensemble time-series forecasting model by mapping current accident data to historical traffic–weather datasets covering Greater London (2000–2019); Wang et al. [22] employed spatiotemporal data association to identify 2126 accident-related violations and then used the FP-growth algorithm to mine 18 strong association rules linking five accident types to four violation categories; Wang et al. [23] constructed a causation analysis network using complex network topology, extracted dimension reduction factors from network parameters, and applied four machine learning algorithms to achieve precise classification of accident severity. Machine learning methods excel at capturing nonlinear relationships in high-dimensional spaces and boosting predictive performance, but remain limited in model interpretability and multi-source information fusion.

(3) Knowledge Graph Methods

More recently, knowledge graph approaches have emerged as powerful tools for deep fusion and inferential analysis of multi-source heterogeneous accident data. Zhang et al. [24] built a traffic accident knowledge graph that provides visual analyses across accident profiling, classification, statistics, and association pathways—seamlessly blending human and machine cognition and greatly improving the interpretability of massive, complex datasets; Wang et al. [25], for structured accident records, employed a Bi-LSTM + Bi-CRF pipeline for label extraction, and for unstructured text used a piecewise-convolutional

neural network (PCNN) to identify accident elements, then applied a relation-inference algorithm to complete implicit links among labels and visualized the unified data via the knowledge graph. Knowledge graph methods leverage semantic network construction and reasoning to achieve deep integration of multi-source heterogeneous data, thus laying a solid foundation for subsequent embedding-based representation learning and risk-prediction tasks.

As summarized in Table 1, traditional statistical methods offer strong interpretability and causal inference capabilities but are limited when handling high-dimensional heterogeneity and complex nonlinear interactions. Machine learning approaches can capture nonlinear patterns and deliver high-precision predictions in large feature spaces, yet they often lack sufficient interpretability and are sensitive to hyperparameter choices and data quality. Knowledge graph methods, by building semantic networks of multi-source heterogeneous information, explicitly model relationships between entities—enabling the integration of diverse accident factors and supporting graph-based reasoning and visualization—but their ontology design and triple-extraction processes require extensive manual annotation, and their embedding models still face parameter redundancy and insufficient expressive capacity when representing complex many-to-one and many-to-many relations. In light of these limitations of existing translation-based embeddings on complex traffic accident knowledge graphs, this paper proposes an enhanced TransD model; it incorporates entity–attribute projection vectors into the dynamic mapping mechanism to explicitly encode fine-grained domain attributes such as "accident severity" and "road conditions", and introduces a dual-limit scoring loss function to independently regulate positive and negative sample boundaries. This design aims to balance representational power with computational efficiency, thereby improving both embedding quality and discriminative performance in scenarios involving many-to-one and many-to-many relations.

Table 1. Summary of current research status.

Method Category	Representative Works	Main Methods	Advantages	Disadvantages
Traditional Statistical Methods	[1,16,17]	Improved DBSCAN clustering; Decision tree analysis; Aggregate clustering and injury-severity analysis	Strong interpretability and causal inference capability; Well-established parameter estimation frameworks	Poor scalability to high-dimensional, heterogeneous data; Limited ability to model complex nonlinear relationships
Machine Learning Methods	[18–23]	SVM classification; ConvLSTM + BCDU-Net for spatiotemporal feature extraction; k-means clustering; Time-series ensemble forecasting; FP-growth association mining; Complex-network topology analysis	Excellent at capturing nonlinear patterns; High predictive accuracy on large feature sets; Flexible to different data modalities	Model interpretability often low; Requires careful hyperparameter tuning; Risk of overfitting without sufficient data or regularization
Knowledge Graph Methods	[24,25]	Semantic network construction and visualization; Bi-LSTM + Bi-CRF and PCNN for entity/relation extraction; Rule-based and embedding-based relation inference	Deep integration of multi-source heterogeneous data; Explicit modeling of semantic relationships; Enables graph-based reasoning and explainability	Knowledge engineering overhead (ontology design, data curation); Computationally intensive for large graphs; Performance sensitive to quality of extracted triples

3. Construction Methodology of the Traffic Accident Knowledge Graph

A Traffic Accident Knowledge Graph is a semantic network-based data structure designed to systematically integrate and represent multi-source, heterogeneous data related to traffic incidents. By modeling entities—such as accident events, vehicles, road infras-

tructure, and meteorological conditions—and the semantic relationships among them, the graph captures the complex interactions underlying accident mechanisms and contributing factors [26]. This structured representation not only facilitates unified data management and information fusion but also provides a robust data foundation and decision-support framework for accident pattern recognition, causal analysis, and risk prediction.

3.1. Structure of the Traffic Accident Knowledge Graph

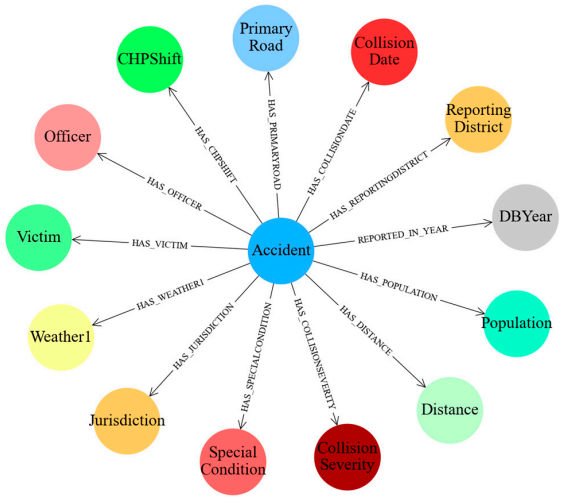
The architecture of a Traffic Accident Knowledge Graph is centered on the accident event, employing multi-layered, cross-dimensional semantic relations to articulate both the intrinsic logic of the incident and its external influencing factors. The accident node is directly and explicitly linked to entities representing spatiotemporal context, environmental conditions, and accident outcomes, thereby capturing detailed accident circumstances and impact levels [27]. Concurrently, two complementary subgraphs—one modeling participant-related factors and the other representing victim-centric injury characteristics—are each directly connected to the central accident node, forming an integrated and harmonized network. This design not only preserves the completeness and diversity of information representation but also provides a robust semantic foundation for in-depth investigation of accident mechanisms and for accurate risk prediction.

Figure 1a illustrates the knowledge graph constructed around the central concept of a traffic accident event (Accident), forming a tightly interlinked semantic network. The accident node is connected to temporal-attribute nodes that encode the event's spatiotemporal context. From this central node extend edges to entities representing environmental and situational factors—such as judicial jurisdiction, law enforcement officer, reporting region, work shift, regional population density, special environmental conditions, primary road information, accident severity, distance from roadway, weather conditions, and date of occurrence. To model accident impact, the accident node directly links to victim entities, which in turn are associated with role and age attributes to specify the nature of injuries. The resulting graph presents clear, direct relationships among over 120 elements, capturing the interplay of multidimensional factors and providing a solid semantic foundation for in-depth analysis of accident mechanisms and risk prediction.

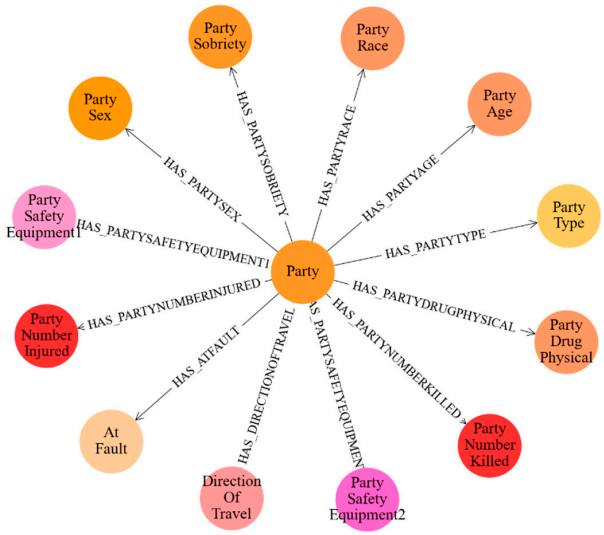
Figure 1b shows the subgraph centered on the party involved (Party), a key component of the overall Traffic Accident Knowledge Graph. This subgraph focuses on participant characteristics, constructing a semantic network that covers identity, behavioral patterns, and physiological attributes. Specifically, the Party node is directly linked to attribute nodes—party type, fault assignment, gender, age, alcohol/drug influence, direction of travel, safety equipment usage, numbers killed and injured, and race—forming a compact, semantically explicit subnetwork. This design not only highlights the role of human factors in accidents but also integrates seamlessly with other subgraphs that capture spatiotemporal context and environmental conditions, jointly enriching the graph with semantic cues crucial for fault attribution and risk assessment.

Figure 1c presents the subgraph centered on victims (Victim), providing another essential perspective of the knowledge graph. It details injury-related attributes of individuals involved in accidents. The Victim node tightly associates with key properties—victim role, age, injury severity, seating position, safety equipment usage, and ejection status. To complement individual injury data, this subgraph also incorporates nodes for external environmental factors and overall accident severity, such as weather conditions, road surface, primary road, and intersections—thereby forming a cohesive, richly connected subnetwork. Together with the other subgraphs, this structure intuitively reveals complex interactions between injury outcomes and environmental factors, constituting a comprehensive traffic

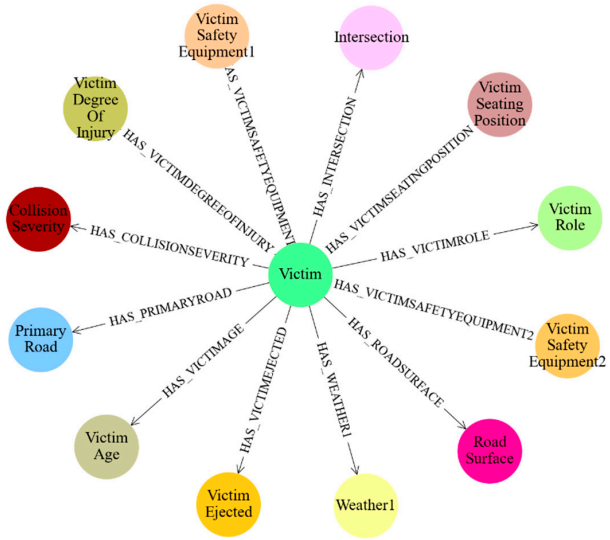
accident information system that underpins in-depth studies of injury mechanisms and the development of targeted rescue strategies.



(**a**) Illustration of Partial Entities and Relations Related to the "Accident" Entity



(**b**) Illustration of Partial Entities and Relations Related to the "Victim" Entity



(**c**) Illustration of Partial Entities and Relations Related to the "Victim" Entity

Figure 1. Schematic representation of the partial structure of the Traffic Accident Knowledge Graph.

3.2. Construction of the Traffic Accident Knowledge Graph

A knowledge graph fundamentally adopts a graph structure and typically represents real-world entities and their interrelations in the form of triples (head entity, relation, tail entity). In this context, a traffic accident can be intuitively modeled as a set of such triples—for example, "drunk driving causes vehicle collision" is represented as (Drunk Driving, causes, Vehicle Collision), where "Drunk Driving" and "Vehicle Collision" serve as the head and tail entities, respectively, and "causes" denotes the relation. This formalization transforms voluminous and redundant accident information into a standardized, structured knowledge network, enabling a comprehensive depiction of road traffic accident data within a single graph and laying the foundation for subsequent knowledge retrieval and deep semantic mining.

(1) Data Sources

The dataset used in this study is an open-source collection from Kaggle that compiles approximately 120 accident-related factors leading to traffic collisions in California—encompassing accident time, location, cause, severity, weather and lighting conditions, road characteristics, site attributes, regional demographics, victim profiles, and participant details. This rich, multidimensional, and fine-grained description framework enables in-depth analysis of interdependencies among diverse data elements. Specifically, the California SWITRS (Statewide Integrated Traffic Records System) data—provided by the California Highway Patrol (CHP)—covers every reported collision from 1 January 2001 to mid-December 2020. The dataset contains roughly 9,460,000 records, totaling 5.78 GB. Alex Gude transformed the original SWITRS CSV files into a consolidated SQLite database by merging four snapshots (2016, 2017, 2018, and 2020) and retaining, via the process_date field in the collisions table, only the most recent entry for each collision, thereby preserving update continuity and accuracy. The database comprises four primary tables:

Caseids: Stores the unique case_id and reporting year for referential integrity and traceability.

Collisions: Contains 74 fields detailing collision time, geographic coordinates, environmental conditions, vehicle attributes, and severity ratings.

Parties: Includes 31 fields covering involved parties' demographic characteristics, vehicle information, and fault assignment.

Victims: Consists of 11 fields describing injury severity, seating position, and safety equipment usage for each casualty.

To facilitate downstream analysis, the SWITRS-to-SQLite conversion script normalized missing values and converted all date/time fields into SQLite-compatible formats. Hosted under an open license on Kaggle, this relational dataset can be imported directly into analysis platforms such as Pandas or GIS for further processing and visualization. Its comprehensive schema and extensive field set provide a robust foundation for constructing our California Traffic Accident Knowledge Graph (TAKG) and performing embedding-based representation learning.

By integrating and mining these data within a knowledge graph framework, our study can deeply explore complex interactions among factors, thereby delivering actionable insights for traffic safety management and policy planning. Prior to analysis, we ensured data integrity and reliability by manually cleaning the dataset—removing duplicate records, standardizing field formats, and verifying consistency across all entries.

(2) Knowledge Graph Construction

As illustrated in Figure 2, this study employs a top-down approach [28] to construct the Traffic Accident Knowledge Graph. This methodology first defines a domain ontology and data schema—establishing class hierarchies and constraint rules—and then populates

the graph by extracting and instantiating entities, attributes, and relationships from the identified data sources.

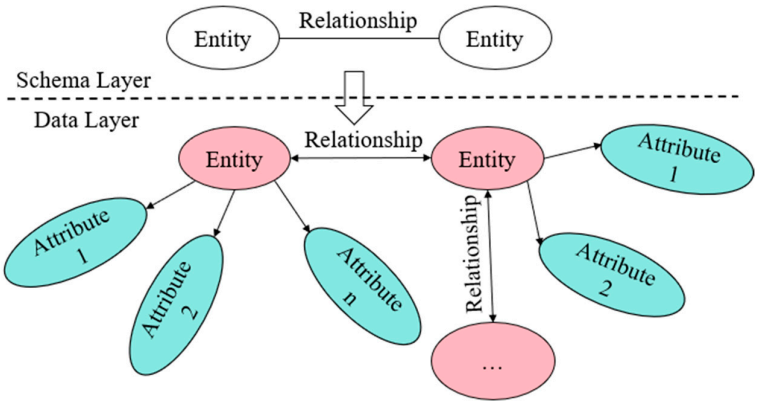


Figure 2. Schematic diagram of the top-down knowledge graph construction process.

In a knowledge graph, relationships between entities are represented as triples of the form (head entity–relation–tail entity), where the head and tail entities denote traffic accident elements or information and are depicted as nodes. Relations, which connect two entities, are depicted as edges.

In this study, we regard the constructed knowledge graph as the triple set G = {(h, r, t)}, and define the following matrices over its entities and attributes:

1. Relation matrix R

$$R \in \{0,1\}^{n\times n},\ R_{ij} = \begin{cases} 1, & \exists r \text{ s.t. } (s_i, r, s_j) \in G, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Here, n denotes the total number of entities, and R_{ij} = 1 indicates that there exists a semantic edge between entities s_i and s_j.

2. Attribute matrix S

$$S \in \mathbb{R}^{n\times N},\ S_{i,q} = v_{i,q} \tag{2}$$

N denotes the total number of attribute categories, and $v_{i,q}$ represents the value of entity S_i for the qth attribute (numerical attributes are used directly, whereas categorical attributes are encoded via one-hot encoding or integer mapping).

3. Attribute co-occurrence matrix A

$$A \in \mathbb{R}^{N\times N},\ A_{l,k} = \sum_{i=1}^{n} \mathbb{I}(S_{i,l} \neq 0 \wedge S_{i,k} \neq 0) \tag{3}$$

Here, $\mathbb{I}$ $(\cdot)$ denotes the indicator function, and $A_{l,k}$ measures the number of times attribute l and attribute k co-occur in the same entity, thus capturing the strength of association between these attributes.

The detailed construction workflow is depicted in Figure 3.

Table 2 presents a subset of the key entities and relations in the constructed Traffic Accident Knowledge Graph, integrating multidimensional information—such as accident identifiers, spatiotemporal localization, environmental factors, and damage assessment—into a graph-structured fusion of multi-source heterogeneous data. By designating Accident.case_id as the primary indexing node, reference consistency is maintained across different temporal and spatial contexts, thus enabling cross-period and cross-domain analyses. The REPORTED_IN_YEAR relation links accident nodes to DBYear.year nodes, supporting temporal distribution and trend analysis of accident occurrences by year. The HAS_JURISDICTION relation connects ac-

cident nodes to Jurisdiction.value nodes, facilitating the generation of administrative-area heat maps and comparative studies. Environmental factors are incorporated via relationships such as HAS_SPECIAL_CONDITION → SpecialCondition.condition and HAS_WEATHER1 → Weather1.condition, effectively capturing special road conditions and meteorological variables that serve as critical covariates in modeling accident causation and severity. Road spatial information is accurately encoded through HAS_PRIMARY_ROAD and HAS_DIRECTION relations and their corresponding nodes, pinpointing the road name and direction of travel for each incident. The HAS_LIGHTING → Lighting.condition node characterizes diurnal lighting conditions, which are essential for analyzing daytime-versus-nighttime accident patterns. Distance measures, introduced via HAS_DISTANCE → Distance.value, support micro-spatial offset analyses and buffer-zone studies. Finally, victim impact is quantified by HAS_KILLED_VICTIMS → KilledVictims.count and HAS_INJURED_VICTIMS → InjuredVictims.count relations; when combined with causal relations such as HAS_PRIMARY_COLLISION_FACTOR and HAS_TYPE_OF_COLLISION, they collectively profile accident outcomes and provide a robust data foundation for downstream tasks like data mining and link prediction [29].

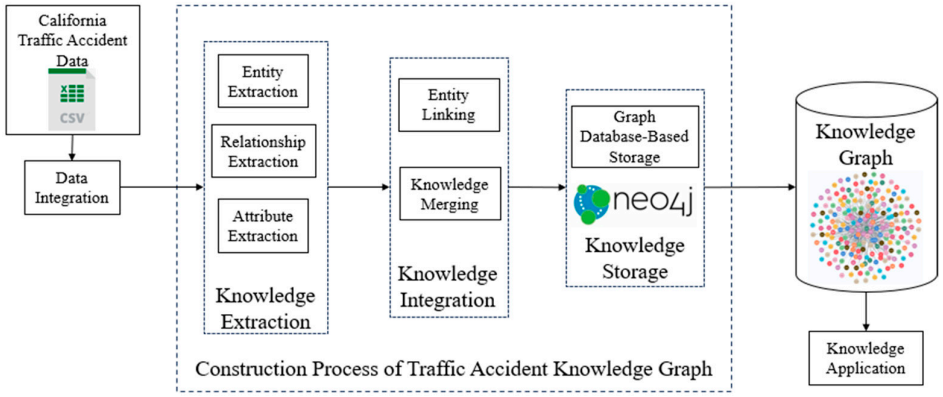


Figure 3. Schematic diagram of the Traffic Accident Knowledge Graph construction workflow.

Table 2. Description of key entities and relations in the Traffic Accident Knowledge Graph.

Type	Field Name	Description and Example
Entity	Accident.case_id	Unique identifier for the accident case; e.g., 21900517
Relationship	REPORTED_IN_YEAR	Relation linking an Accident to its reporting year; e.g., 21900517 REPORTED_IN_YEAR 2019
Entity	DBYear.year	Year when the accident was reported; e.g., 2019
Relationship	HAS_JURISDICTION	Connects Accident to Jurisdiction; e.g., 21900517 HAS_JURISDICTION Los Angeles
Entity	Jurisdiction.value	Name of the legal jurisdiction; e.g., Los Angeles
Relationship	HAS_SPECIAL_CONDITION	Links Accident to special road condition; e.g., 21900517 HAS_SPECIAL_CONDITION Vista Point or Rest Area
Entity	SpecialCondition.condition	Description of special condition; e.g., Vista Point or Rest Area
Relationship	HAS_PRIMARY_ROAD	Links Accident to primary road; e.g., 21900517 HAS_PRIMARY_ROAD US 101
Entity	PrimaryRoad.name	Name or designation of primary road; e.g., US 101
Relationship	HAS_DISTANCE	Associates Accident with distance off road; e.g., 21900517 HAS_DISTANCE 0.5
Entity	Distance.value	Distance from reference point (miles); e.g., 0.5
Relationship	HAS_DIRECTION	Links Accident to traffic direction; e.g., 21900517 HAS_DIRECTION Northbound
Entity	Direction.value	Traffic direction at accident location; e.g., Northbound
Relationship	HAS_WEATHER1	Connects Accident to primary weather condition; e.g., 21900517 HAS_WEATHER1 Clear
Entity	Weather1.condition	Primary weather during accident; e.g., Clear

Table 2. *Cont.*

Type	Field Name	Description and Example
Relationship	HAS_LIGHTING	Associates Accident with lighting condition; e.g., 21900517 HAS_LIGHTING Daylight
Entity	Lighting.condition	Lighting at time of accident; e.g., Daylight
Relationship	HAS_KILLED_VICTIMS	Links Accident to count of fatalities; e.g., 21900517 HAS_KILLED_VICTIMS 0
Entity	KilledVictims.count	Number of people killed; e.g., 0
Relationship	HAS_INJURED_VICTIMS	Links Accident to count of injuries; e.g., 21900517 HAS_INJURED_VICTIMS 2
Entity	InjuredVictims.count	Number of people injured; e.g., 2
Relationship	HAS_PRIMARY_COLLISION_FACTOR	Associates Accident with primary collision factor; e.g., 21900517 HAS_PRIMARY_COLLISION_FACTOR Unspecified
Entity	PrimaryCollisionFactor.factor	Description of the primary collision factor; e.g., Unspecified
Relationship	HAS_TYPE_OF_COLLISION	Links Accident to collision type; e.g., 21900517 HAS_TYPE_OF_COLLISION Sideswipe
Entity	TypeOfCollision.type	Type of collision; e.g., Sideswipe

This study employs the Neo4j graph database as the storage and visualization platform for the knowledge graph. Leveraging Neo4j's native graph storage engine and ACID-compliant transactions, the system can efficiently manage tens of thousands of entities and relations. Neo4j supports the Cypher query language, enabling flexible execution of path retrieval, graph algorithms, and neighborhood analysis, and provides comprehensive visualization tools for intuitively displaying the complex semantic associations among traffic accident entities, thereby facilitating subsequent knowledge retrieval and decision support. As shown in Figure 4, only a subset of the constructed knowledge graph is displayed due to the volume of data. The California Traffic Accident Knowledge Graph (TAKG) comprises 33,755 entity nodes and 115 relation types, encompassing over 120 categories of accident-related elements—such as accident events, participants, victims, and environmental factors. For downstream embedding, the training, validation, and test sets contain 460,005, 57,495, and 57,500 triples, respectively, thus fully reflecting the multi-source heterogeneity and spatiotemporal dynamics of real-world traffic accident data.

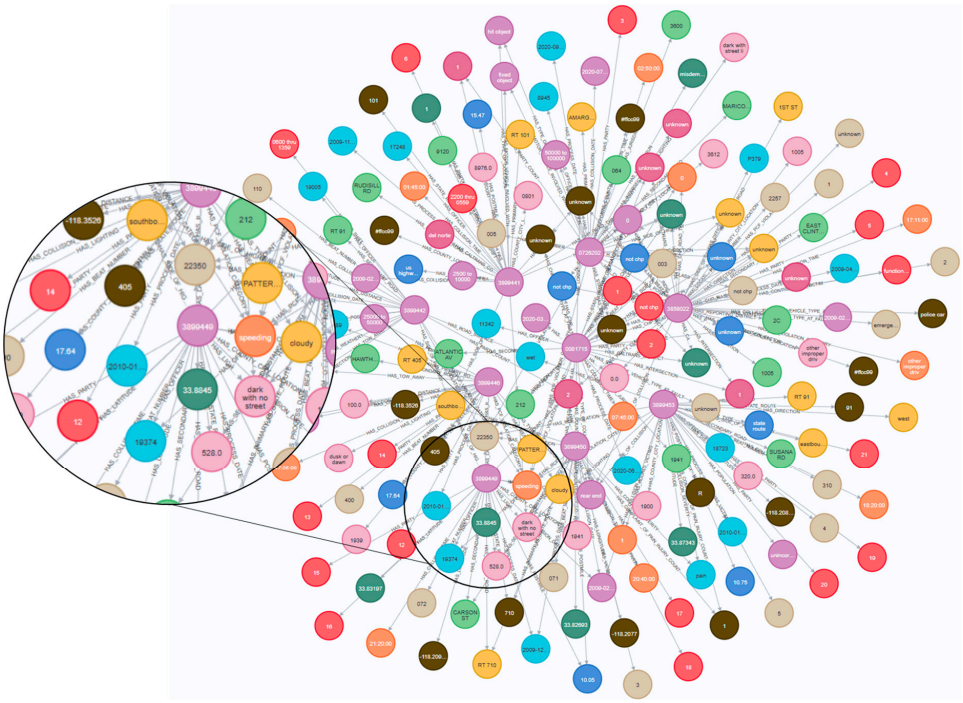


Figure 4. Partial view of the Traffic Accident Knowledge Graph.

4. Representation Learning Methods for Traffic Accident Knowledge Graphs

4.1. Translation-Based Embedding Methods

Translation-based embedding methods for knowledge graphs map entities and relations into a continuous vector space by treating each relation as a translation operator. Specifically, for any valid triple (h, r, t), they enforce that the vector of the head entity plus the relation vector should lie close to the tail entity vector. During training, the distance induced by this translational mapping for positive triples is minimized, while negative samples are generated to maximize incorrect translation distances. This process effectively captures both semantic connections and relational structure, enabling the model to learn the latent semantics and logical patterns within the graph, and providing strong support for downstream tasks such as link prediction and entity classification. Common translation models in knowledge graph embedding include TransE [30,31], TransH [32], TransR [33], and TransD [34].

Bordes et al. [31] introduced the TransE model, which embeds entities and multi-relation data into a low-dimensional vector space. The core idea is to interpret a relation as a translation from the head entity embedding to the tail entity embedding. As illustrated in Figure 5, given a triple (h, r, t), the relation vector r represents the translation $h + r \approx t$. In other words, if (h, r, t) holds in the knowledge graph, then the embeddings should satisfy $h + r \approx t$.

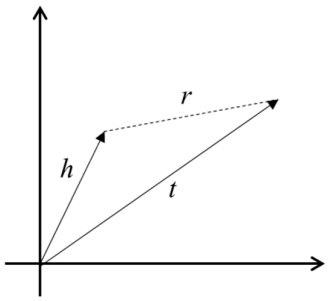


Figure 5. Schematic illustration of the TransE model.

The TransE model is notable for its simplicity and intuitive design; however, its basic translational operation struggles to distinguish entities within one-to-many, many-to-one, or many-to-many relationships, leading to insufficient representational capacity. To address these shortcomings, subsequent models such as TransH, TransR, and TransD have been proposed to extend and refine TransE. TransH, in particular, assigns each relation a dedicated hyperplane and projects entity embeddings onto this hyperplane before applying the translation operation. Concretely, for a triple (h, r, t), the head and tail embeddings h and t are projected to h_p and t_p on the relation-specific hyperplane, and the model enforces $h_p + r \approx t_p$. As illustrated in Figure 6, this mechanism allows entities to adopt distinct representations depending on the relation context, thereby better differentiating one-to-many and many-to-one mappings. Nonetheless, TransH still faces challenges in fully capturing very complex relations and introduces extra computational overhead due to the projection step.

TransR further extends TransH's idea that entities and relationships should exist separately in different semantic spaces. As shown in Figure 7, each relationship r has an exclusive projection matrix $M_{(r)}$ used to map entities from a unified entity space to a relationship-specific space, and then perform translation operations within that space. The optimization objective is $\|M_{(r)} \cdot h + r - M_{(r)} \cdot t\|_2$. Here, $\|\cdot\|_2$ notes the L2 norm, to measure the Euclidean distance in the embedding space. This approach separates entity and relation spaces, enabling the model to capture relation-specific characteristics at a finer granularity

and thus better accommodate more complex mappings. However, it necessitates learning a distinct projection matrix for each relation, which significantly increases the number of parameters, raises computational complexity, and prolongs training time.

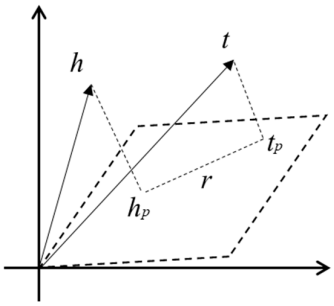


Figure 6. Schematic illustration of the TransH model.

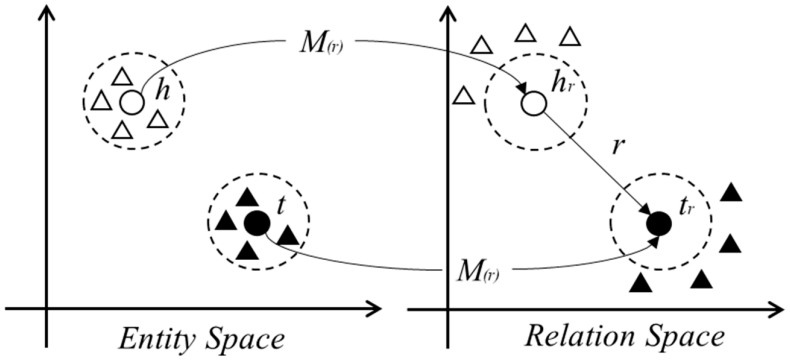


Figure 7. Schematic illustration of the TransR model.

TransD addresses the parameter redundancy of projection matrices in TransR—caused by the heterogeneous types of relations and entities—by dynamically generating mapping matrices. As illustrated in Figure 8, for each entity–relation pair, TransD introduces two vectors—one representing the semantics of the entity (or relation) and another for constructing the projection matrix—thereby dynamically computing the mapping matrix and performing the translation operation. This approach retains the ability to model relation-specific semantics while markedly reducing parameter count and computational complexity, thus enhancing training efficiency.

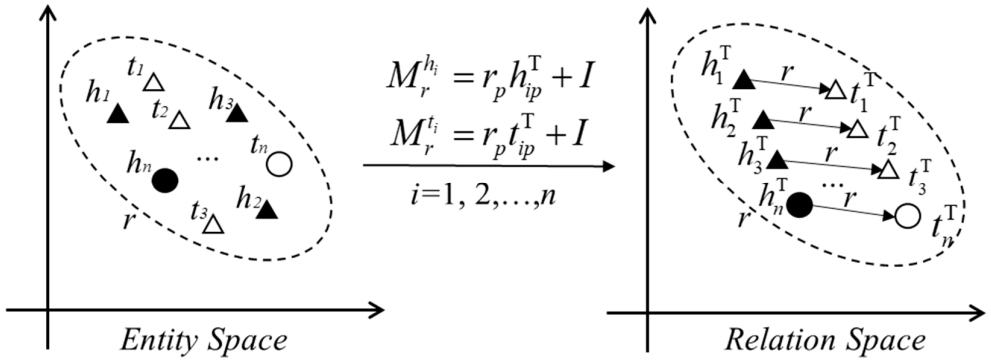


Figure 8. Schematic illustration of the TransD model.

The traffic accident knowledge graph constructed in this study encompasses heterogeneous entities—such as vehicles, drivers, accident locations, and environmental factors—and their intricate, diverse interrelations, resulting in frequent one-to-many, many-to-one, and many-to-many mapping scenarios. Although TransE is simple and efficient, it struggles

to distinguish entities under such complex mappings, causing representational ambiguity. TransH mitigates this by introducing relation-specific hyperplanes, yet it incurs higher computational overhead and still lacks fine-grained semantic modeling. TransR further separates entity and relation spaces to enhance expressiveness, but it demands a distinct projection matrix per relation, significantly increasing parameter counts and computational complexity. In contrast, TransD dynamically generates mapping matrices for each entity–relation pair, preserving fine-grained semantic nuance while markedly reducing parameter redundancy and computational cost, thus flexibly and efficiently capturing the complex relationships inherent in traffic accident data. Therefore, this study adopts the TransD model for knowledge graph embedding to strike an optimal balance between expressive power and computational efficiency.

4.2. Representation Learning of the Traffic Accident Knowledge Graph Based on the Improved TransD Model

In constructing and embedding the traffic accident knowledge graph, the standard TransD model leverages dual projections of entities and relations to capture fundamental semantics. However, because its projection matrices depend solely on the intrinsic vectors of entities and relations, TransD struggles to model the fine-grained influence of domain-specific attributes—such as accident severity or road surface conditions—on entity representations. Moreover, its single-margin loss function is inadequate for effectively separating positive and negative samples within complex, noise-laden accident data. To remedy these shortcomings, this study augments TransD's dynamic mapping mechanism with attribute projection vectors, thereby strengthening the coupling between inherent entity features and relation semantics. Simultaneously, we introduce a dual-limit scoring loss to independently adjust the margins for positive and negative samples, resulting in enhanced representational capacity and discriminative performance on the traffic accident knowledge graph.

(1) Entity and Relation Representation

For the traffic accident knowledge graph, entities of various types (e.g., Accident, Officer, Weather, Road Surface) exhibit distinct feature distributions in their initial embeddings, while relations (e.g., HAS_COLLISION_SEVERITY, HAS_VICTIM, HAS_REPORTING_DISTRICT) capture semantic linkages among different attributes. The core of the TransD model lies in creating dedicated projection vectors for each entity and each relation, which dynamically generate mapping matrices to enable adaptive projections of entities into relation-specific spaces, thereby fully modeling the complex interactions among entities in the traffic accident domain.

Consider an arbitrary triple (h, r, t) in the traffic accident knowledge graph, where $h \in \mathbb{R}^d$ is the embedding of the head entity (e.g., accident event, location, or jurisdiction), $t \in \mathbb{R}^d$ is the embedding of the tail entity (e.g., collision severity, victim, or party), and $r \in \mathbb{R}^d$ is the embedding of the relation (e.g., HAS_COLLISION_SEVERITY, HAS_VICTIM, HAS_PARTY). To capture each entity's intrinsic attributes and their diverse role under different relations, TransD assigns each entity h and t its own projection vector $h_p \in \mathbb{R}^d$ and $t_p \in \mathbb{R}^d$, and each relation r has its projection vector $r_p \in \mathbb{R}^d$, where d represents the embedding dimension. These projection vectors construct dynamic mapping matrices that project h and t into a relation-specific space before applying the translational operation.

(2) Construction of Dynamic Mapping Matrices

For each triple (h, r, t) in the traffic accident knowledge graph, the TransD model constructs dynamic mapping matrices to obtain relation-specific entity representations. For the head entity h and relation r pair, the dynamic mapping matrix is defined as [35]

$$M_r^h = r_p \otimes h_p^{\mathrm{T}} + I \tag{4}$$

where I is the $d \times d$ identity matrix. Similarly, for the tail entity t and relation r pair, the mapping matrix is

$$M_r^t = r_p \otimes t_p^{\mathrm{T}} + I \tag{5}$$

Each mapping matrix is constructed by taking the outer product of the entity (or attribute) projection vector with the relation projection vector, then adding an identity component. This design yields a full-rank transformation that remains efficient in parameter usage: the identity term preserves the original embedding's core features, while the low-rank outer-product term injects relation- or attribute-specific adjustments. As a result, each entity embedding undergoes a stable baseline shift together with a flexible, context-dependent transformation, enabling the model to capture complex one-to-many and many-to-one semantics without the overhead of fully dense matrices.

In the traffic accident domain, inherent attributes of entities—such as Collision Severity or Road Surface—are also critical for representation learning. To explicitly incorporate these attributes into the projection process, this study extends the baseline TransD model by introducing an attribute projection vector $a_p \in \mathbb{R}^d$ for each entity. The mapping matrices are thus expanded to

$$M_{r,a}^h = r_p \otimes h_p^{\mathrm{T}} + a_p \otimes a_p^{\mathrm{T}} + I \tag{6}$$

$$M_{r,a}^t = r_p \otimes t_p^{\mathrm{T}} + a_p \otimes a_p^{\mathrm{T}} + I \tag{7}$$

Here, the attribute projection vector a_p is derived via a dedicated mapping network that encodes supplemental attributes (e.g., accident severity levels). This outer-product term creates a low-rank adjustment that complements the identity component, enabling the mapping matrix to capture fine-grained, attribute-specific transformations while preserving the core features of the original embedding. These enriched mapping matrices enable the model to capture both the relational and attribute-level nuances of entities within the traffic accident knowledge graph.

This construction fully leverages entity and relation projection vectors: the low-rank matrices generated by their outer products perform relation-specific transformations on entity embeddings, dynamically capturing interactions among diverse traffic accident attributes—such as jurisdiction, road conditions, and collision severity. Consequently, inherent entity attributes are explicitly modeled within the projection matrices, enabling embeddings to reflect fine-grained factors like "accident severity" and "road slipperiness".

(3) Projected Entity Representations and Scoring Function

Using the dynamic mapping matrices defined above, the original entity embeddings can be projected into the relation-specific subspaces. Concretely, for a given relation r, the projected representations of the head and tail entities are computed as:

$$h\prime = M_{r,a}^h h \tag{8}$$

$$t\prime = M_{r,a}^t t \tag{9}$$

In the constructed traffic accident knowledge graph, this implies, for example, that the embedding of an Accident node is infused with features pertinent to the "Collision Severity" relation, thereby more accurately reflecting the various factors involved in each incident.

For any triple (h, r, t), the TransD-based scoring function is then defined as:

$$f_r(h,t) = \left\| h' + r - t' \right\|_2^2 = \left\| \left(r_p h_p^{\mathrm{T}} + a_p a_p^{\mathrm{T}} + I \right) h + r - \left(r_p t_p^{\mathrm{T}} + a_p a_p^{\mathrm{T}} + I \right) t \right\|_2^2 \tag{10}$$

This scoring function measures the distance between the relation-transformed head embedding (plus the relation vector) and the relation-transformed tail embedding. In the traffic accident knowledge graph, each triple—such as (Accident, HAS_COLLISION_SEVERITY, Collision Severity)—is evaluated using this function, enabling the model to separate true triples from corrupted (negative) ones in the embedding space as effectively as possible.

To compute the distance between the relation-transformed head embedding and the relation-transformed tail embedding, we adopt the Euclidean (L2) norm. In practice, this means that for any difference vector, we take each of its components, square them, sum all of those squares, and then extract the square root of that sum—thus obtaining the straight-line distance in the embedding space. This choice ensures a smooth, differentiable measure that is widely used for geometric similarity and yields stable gradients during optimization. Moreover, to guard against over-fitting and to keep embedding magnitudes under control, we apply independent L2 regularization on all embedding and projection vectors by adding the sum of their squared lengths (multiplied by a small weight factor) to the overall loss. This penalty encourages the model to learn more compact representations without interfering with the primary ranking objective.

(4) Loss Function and Training Objective

In the baseline model, training employs a margin-based ranking loss. Specifically, let S denote the set of positive triples and construct a negative sample set S′ by randomly corrupting head or tail entities. The training objective minimizes the following loss:

$$\mathcal{L} = \sum_{(h,r,t)\in S} \sum_{(h\prime,r,t\prime)\in S\prime} [\gamma + f_r(h,t) - f_r(h\prime,t\prime)]_+ \tag{11}$$

where γ is the margin hyperparameter and $[x]_+ = \max(0, x)$ denotes the hinge function.

To ensure that positive triples score sufficiently low while negative triples score sufficiently high, this study introduces a dual-limit loss [36]. Let $\mathbb{C}$ be the positive set and $\mathbb{C}'$ the negative set, with upper limit m_2 lower limit m_3 ($m_3 > m_2$), and balance coefficient λ. The optimized objective becomes:

$$\mathcal{L}_{DL} = \sum_{(h,r,t)\in\mathbb{C}} [f_r(h,t) - m_2]_+ + \lambda \sum_{(h\prime,r,t\prime)\in\mathbb{C}\prime} [m_3 - f_r(h\prime,t\prime)]_+ \tag{12}$$

In this study, both the upper and lower scoring thresholds were regarded as hyperparameters and determined by conducting a systematic grid search on the validation set. For each threshold, a discrete grid of candidate values spanning multiple scales was defined, and the pair satisfying the constraint that the lower threshold exceeds the upper—while yielding the highest validation performance—was selected for final evaluation. In the same manner, the balance coefficient governing the weight of the negative-limit term was treated as a tunable hyperparameter: a discrete range of values was evaluated by grid search, and the coefficient achieving peak validation metrics was adopted. By imposing this dual-limit framework, the score distributions of positive and negative samples are independently constrained, ensuring a minimum separation margin between them and accelerating convergence during training.

By leveraging dynamic mapping matrices, integrating attribute projection vectors for each entity, and introducing a dual-limit loss, the enhanced model fully accounts for semantic transformations of entities across diverse relations, thereby achieving precise modeling of multi-factor interactions in complex traffic accident scenarios. Ultimately, the resulting low-dimensional embeddings not only preserve the structural information of the knowledge graph but also capture fine-grained semantic interactions among various

entities within the traffic accident domain, providing a robust representational foundation for downstream tasks such as accident prediction and risk assessment.

5. Experiments

5.1. Model Validation

(1) Datasets

This study selects FB15K-237 and WN18RR as benchmark datasets for evaluation, effectively mitigating the inverse-relation leakage issues present in their predecessor datasets and thus ensuring the rigor and reliability of the evaluation results. FB15K-237 and WN18RR are derived from FB15K and WN18, respectively, by removing redundant inverse relations, thereby reducing performance inflation in the test sets caused by duplicated information. This combined dataset configuration not only reflects the model's capability to generalize across sparse and de-noised knowledge graphs but also demonstrates its potential for application in high-complexity domains. Detailed statistics of these datasets are provided in Table 3.

Table 3. Statistics of the validation datasets for model evaluation.

Data Set	Number of Entities	Number of Relationships	Size of Training Set	Size of Validation Set	Size of Test Set
FB15K-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3034	3134

(2) Experimental Design

In the data preprocessing stage, we uniformly extract all original triples and each entity's intrinsic attributes, constructing separate mapping dictionaries for entities, relations, and attributes to convert textual information into corresponding integer indices. A full-set triple filter is also generated to automatically mask known positive samples during subsequent training and validation. In this stage, we standardize all field formats and impute or discard missing values in critical attributes according to domain-specific rules. We then normalize date/time fields to a uniform format and clean categorical strings by trimming whitespace, applying consistent casing, and removing special characters. Subsequently, each unique entity name, relation label, and attribute value is mapped to a corresponding integer index via dedicated lookup tables. Numerical attributes are scaled to facilitate convergence during training, and referential integrity is verified across the head–relation–tail triples. Finally, the fully cleaned and indexed triple set is randomly partitioned into training, validation, and test subsets using a fixed seed. Additionally, a full-set filter is applied to mask known positive triples during subsequent model fitting and evaluation.

During model construction, dropout layers are incorporated into the entity embeddings, relation embeddings, and newly introduced attribute projection vectors, and independent L_2 regularization is applied to each of these three vector types to mitigate overfitting while maintaining a balance between attribute and structural information. To address the skewed distribution of head and tail entities, negative samples are generated according to a Bernoulli sampling strategy that respects the relation-specific corruption ratios. We replace the single-margin loss with a dual-limit loss to achieve finer control over the score distributions of positive and negative samples. Training proceeds in mini-batches, with block-wise candidate entity search to ensure that, even under GPU memory constraints, scores and gradients for all entities within each batch are correctly computed and updated. To assess the effect of embedding dimensionality on the generalization and representational capacity of the enhanced TransD model, we conducted unified training

and evaluation at dimensions of 50, 100, 150, 200, 250, and 300. Comparative experiments against TransE, TransH, TransR, and the original TransD were performed to comprehensively validate the advantages of our enhancement. Additionally, ablation studies isolating the attribute projection vector and the dual-limit loss modules were carried out to quantify each component's independent impact on link prediction and various evaluation metrics, thereby ensuring the reliability and modular robustness of the conclusions. Model performance variations were recorded for each experimental condition.

(3) Evaluation Metrics

To quantitatively assess model performance, this study employs Mean Rank (MR), Mean Reciprocal Rank (MRR), Hits@K, triple classification accuracy (Accuracy), and F1-score as evaluation metrics.

The Mean Rank (MR) is calculated as:

$$MR = \frac{1}{N}\sum_{i=1}^{N} rank_i \tag{13}$$

where N is the total number of test samples and $rank_i$ denotes the rank position of the correct entity for the i-th sample within the list of candidate entities. A lower MR indicates that, on average, the model ranks correct entities closer to the top.

The Mean Reciprocal Rank (MRR) is defined as:

$$MRR = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{rank_i} \tag{14}$$

This metric emphasizes the cases where the correct entity appears at high positions, as higher reciprocal ranks (i.e., lower $rank_i$) yield larger contributions to MRR, reflecting the model's ability to assign higher scores to true answers.

The Hits@K metric measures the proportion of test cases in which the correct entity appears among the top K candidates. It is defined as:

$$Hits@K = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(rank_i \leq K), (K = 1, 3, 10) \tag{15}$$

where $\mathbb{I}(\cdot)$ is the indicator function. A higher *Hits@K* value indicates that, in most cases, the model ranks the correct answer within the top K, thereby improving retrieval effectiveness.

Triple classification accuracy (Accuracy) reflects the overall correctness of distinguishing positive and negative triples, computed as:

$$\text{Accuracy} = \frac{\text{Number of Currectly Classified Triples}}{\text{Total Number of Triples}} \tag{16}$$

A higher accuracy implies that the model correctly classifies a larger proportion of triples, directly indicating enhanced overall discriminative capability.

The F1-score harmonizes precision and recall and is defined as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

and TP, FP, and FN denote the counts of true positives, false positives, and false negatives, respectively. The F1-score balances precision and recall; a higher F1 indicates that the model

maintains high precision while also achieving strong recall, yielding a more balanced and robust performance in classifying both positive and negative triples.

(4) Results Analysis

The experimental results demonstrate that, across varying embedding dimensions, the TransD model achieves significant performance gains on both FB15K-237 and WN18RR datasets, as shown in Figure 9 for FB15K-237. Experimental results on the FB15K-237 dataset demonstrate that the enhanced TransD model (hereafter Enhanced TransD), which integrates attribute projection vectors and a dual-limit loss mechanism, yields substantial improvements in representational capacity and retrieval performance over classical translation-based embedding methods. The FB15K-237 dataset contains approximately 15,000 entities and 237 relation types, whose diversity and sparsity pose significant challenges for embedding models. Unlike TransE, TransH, and TransR—which typically exhibit a "dimension–performance" trade-off—Enhanced TransD maintains the original dynamic mapping mechanism while explicitly encoding entity attributes into projection vectors and employing a dual-threshold scoring loss to reinforce positive–negative sample separation. This design leads to superior performance across most evaluation metrics.

For the Mean Rank (MR) metric, Enhanced TransD achieves a continuous decline from ~320 at 50 dimensions to ~230 at 300 dimensions. In contrast, the original TransD fluctuates between 240 and 290 over the same range and only slightly outperforms the enhanced model at the lowest dimensions (50 and 100). Beyond 150 dimensions, Enhanced TransD rapidly surpasses the baseline and sustains its lead; above 200 dimensions, MR reductions plateau, indicating convergence toward optimal performance. Compared with TransE, TransH, and TransR, Enhanced TransD attains lower average MR in medium-to-high dimensions ($\geq$150), reducing MR by approximately 15–20% in the 250–300 range, which underscores the efficacy of attribute projection in augmenting entity feature representation in high-dimensional spaces.

Regarding Mean Reciprocal Rank (MRR), Enhanced TransD exhibits a steady increase from ~0.18 to ~0.28 as dimensions grow from 50 to 300. By contrast, the original TransD only rises marginally from ~0.19 to ~0.20, with noticeable oscillations. Enhanced TransD's advantage emerges as early as 100 dimensions and maintains a 0.03–0.05 gain thereafter. TransE, TransH, and TransR show early MRR improvements but often suffer declines or stagnation above 200 dimensions, suggesting overfitting or optimization instability. The dual-limit loss effectively balances positive and negative gradient updates, securing consistent ranking accuracy across all dimensions.

In terms of Hits@k metrics, Enhanced TransD outperforms baselines markedly: Hits@10 increases from ~0.30 to ~0.44 versus 0.37→0.36 for the original TransD; Hits@3 and Hits@1 rise from 0.20 to 0.32 and 0.10→0.20, respectively, achieving average gains exceeding 0.10. Relative to TransE, TransH, and TransR, the enhanced model's Hits@k improvements exceed 0.05 for all k values, with Hits@1—the strictest measure—demonstrating robust first-rank retrieval. Explicit encoding of domain attributes enables the model to more rapidly distinguish correct entity pairs amid a complex semantic space, thereby significantly boosting hit rates.

Finally, classification metrics (Accuracy and F1 score) also improve notably. Accuracy climbs from ~0.30 to ~0.46 and F1 from ~0.26 to ~0.42 as dimensions increase, plateauing beyond 250 dimensions. This trend indicates that the dual-limit loss not only enhances ranking performance but also strengthens binary classification capability. However, marginal gains diminish at very high dimensions, suggesting that further increases in classification precision may require finer regularization, sparsity constraints, or multi-task learning strategies. Overall, experiments on FB15K-237 affirm the synergistic benefits of attribute

projection and dual-limit loss, offering a novel optimization paradigm for translation-based knowledge graph embedding.

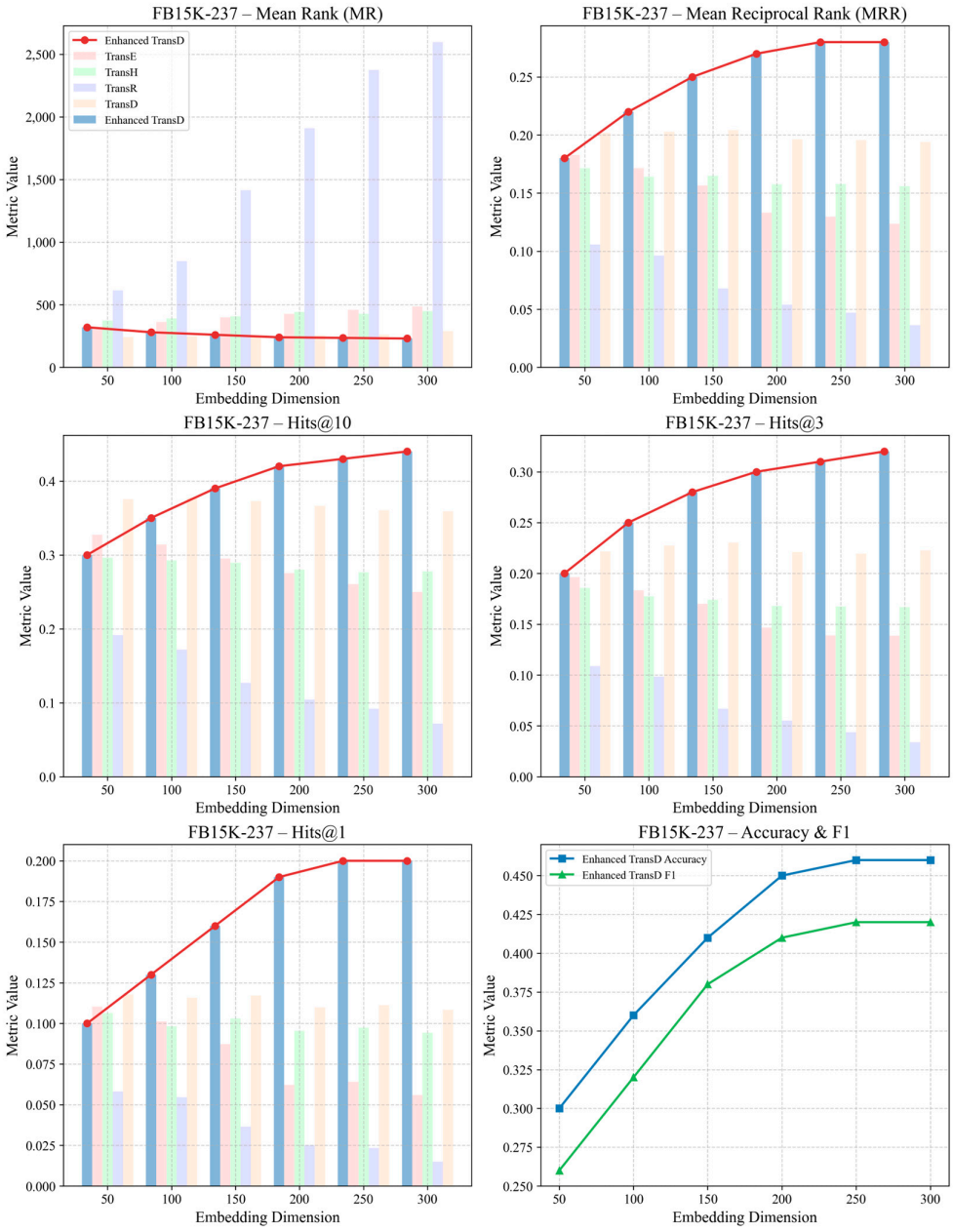


Figure 9. Performance of translation-based knowledge graph embedding models on the FB15K-237 dataset across varying embedding dimensions.

As illustrated in Figure 10 for the WN18RR dataset, experimental results on the WN18RR dataset further demonstrate the applicability and robustness of Enhanced TransD for large-scale relational inference tasks. WN18RR comprises nearly 40,000 entities and 11 relation types; compared with FB15K-237, it exhibits fewer relations but greater entity volume and path diversity, requiring models to capture long-range dependencies and multi-hop reasoning. Through the synergistic effects of attribute projection vectors and the dual-limit loss, Enhanced TransD significantly outperforms baseline models (TransE, TransH, TransR, and original TransD) across MR, MRR, and Hits@k metrics, confirming its adaptability to diverse knowledge graph structures.

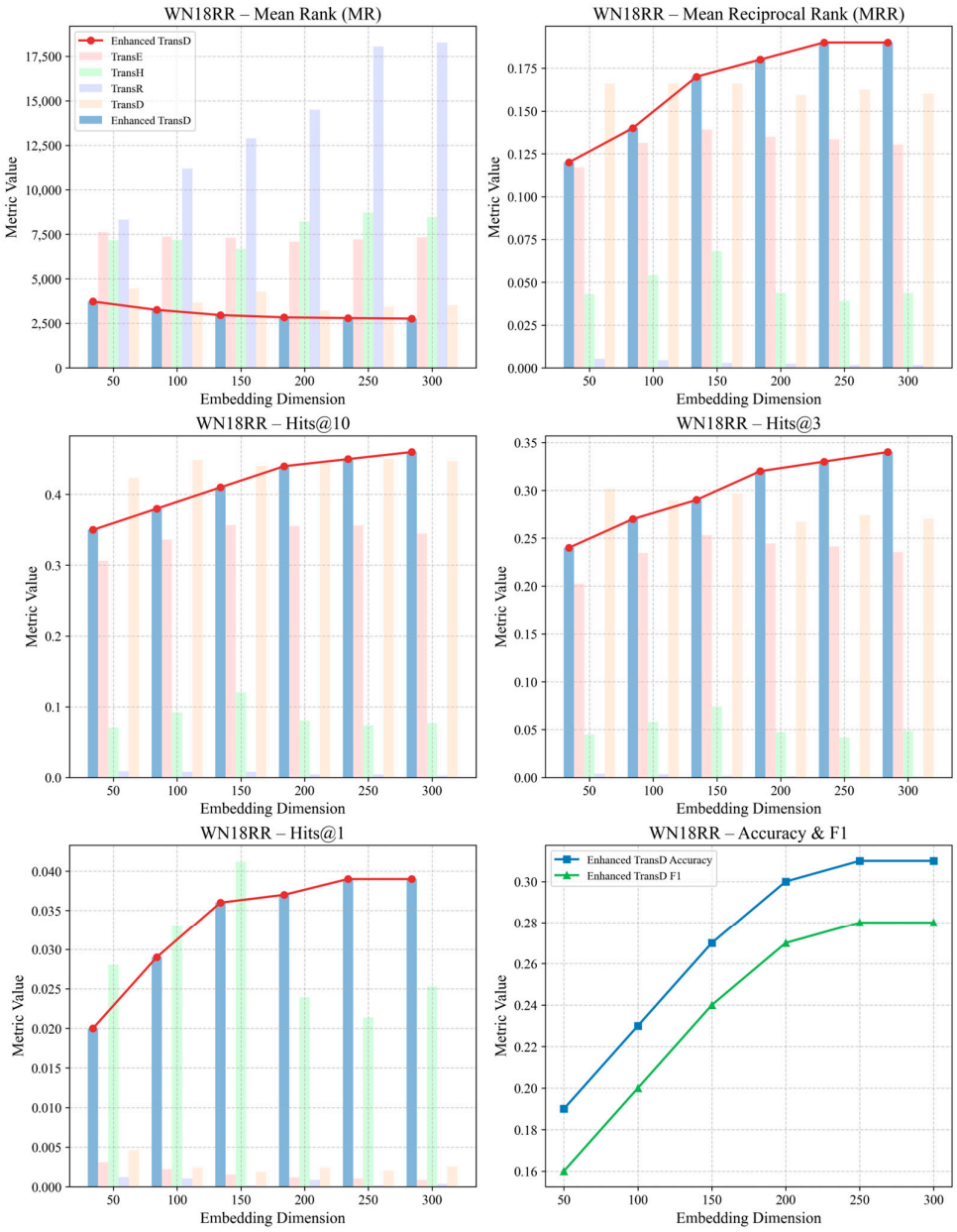


Figure 10. Performance of various translation-based embedding models on the WN18RR dataset across different embedding dimensions.

In terms of mean rank (MR), Enhanced TransD reduces MR from approximately 3726 at 50 dimensions to 2766 at 300 dimensions—a decline exceeding 25%. By contrast, the original TransD only lowers MR from about 4463 to 3532 over the same range. Enhanced TransD already surpasses TransE, TransH, and TransR at low dimensions (50 and 100) and widens its advantage in medium-to-high dimensions (150–300), achieving MR reductions exceeding one thousand ranks at each dimension. These results indicate that attribute projection enriches local and global semantic features for complex paths and multi-hop inference, while the dual-limit loss effectively constrains distance distributions between positive and negative samples, yielding improved ranking even for long-tail entities.

Mean Reciprocal Rank (MRR) also reflects the stability and enhancement brought by our approach; Enhanced TransD's MRR rises from ~0.12 to ~0.19 as dimensions increase, an improvement of 0.07. In contrast, the original TransD fluctuates from ~0.17 to ~0.16, showing an early rise followed by a decline. Although TransE, TransH, and TransR exhibit temporary MRR gains at medium dimensions, they fail to sustain growth and decline

significantly above 200 dimensions. The continuous upward trend of Enhanced TransD at higher dimensions demonstrates that the dual-limit loss maintains balanced gradient updates across complex relations and multi-hop paths, thereby improving sensitivity to correct head–tail pair ordering.

For Hits@k, Enhanced TransD again shows notable gains; Hits@10 increases from ~0.35 to ~0.46 (over 0.10 improvement), while Hits@3 and Hits@1 improve from 0.24 to 0.34 and 0.02→0.039, respectively, underscoring enhanced performance under strict matching criteria. Notably, the most pronounced improvement occurs for Hits@1, which demands precise first-rank retrieval in multi-hop scenarios. By incorporating entity attribute features, the model more effectively distinguishes semantically similar but path-divergent entity pairs, thus achieving superior first-rank accuracy.

Finally, classification metrics mirror ranking improvements; accuracy increases from 0.19 to 0.31 and F1 score from 0.16 to 0.28, with gains plateauing beyond 200 dimensions, indicating near-optimal expressive capacity for large-scale entities and complex multi-hop structures. Overall, on the WN18RR dataset, Enhanced TransD not only significantly surpasses traditional translation-based models but also validates the general effectiveness of attribute projection and dual-limit loss mechanisms across multiple dimensions and metrics, offering a novel paradigm for large-scale knowledge graph embedding and reasoning.

In summary, the Enhanced TransD model demonstrates robust and consistent performance gains across two datasets with distinct semantic structures and relational complexities. In retrieval and ranking tasks over entity–relation pairs, its augmented dynamic mapping mechanism and differentiated loss function jointly improve discrimination between positive and negative samples. For classification and hit-rate metrics, the enhanced model significantly outperforms traditional translation-based baselines, with the attribute projection module offering pronounced compensatory benefits in low-dimensional, resource-constrained settings. Moreover, convergence curves on both datasets exhibit greater stability and generalization, indicating that the dual-limit loss not only deepens separation in the embedding space but also mitigates overfitting.

To assess the statistical significance of overall differences among paired model results, we conducted a Friedman test. All χ^2 statistics corresponded to $p < 0.05$, thereby rejecting the null hypothesis of "no difference in model performance". This confirms that observed variations are not due to random fluctuation and justifies subsequent Nemenyi post-hoc comparisons.

We then performed the Nemenyi test to identify pairwise significant differences across evaluation metrics. As illustrated by the heat map in Figure 11, Enhanced TransD achieves highly significant improvements ($p < 0.01$) over TransE and TransR on nearly all metrics, most notably in Mean Rank (MR) and Hits@10. Comparisons with TransH and the original TransD reveal consistent mild to moderate gains (small p-values), further validating the synergistic effect of attribute projection and dual-limit loss in enhancing discriminatory power. Notably, under the strict Hits@1 criterion on WN18RR—which poses challenges due to prevalent inverse relations—Enhanced TransD also significantly outperforms TransH and TransR ($p < 0.01$), underscoring its resilience. Overall, the Nemenyi analysis quantifies and visually substantiates the superior efficacy of Enhanced TransD, particularly in addressing complex inverse and symmetric relations, while also highlighting the intrinsic difficulty of low-k retrieval on inverse relation-dense datasets.

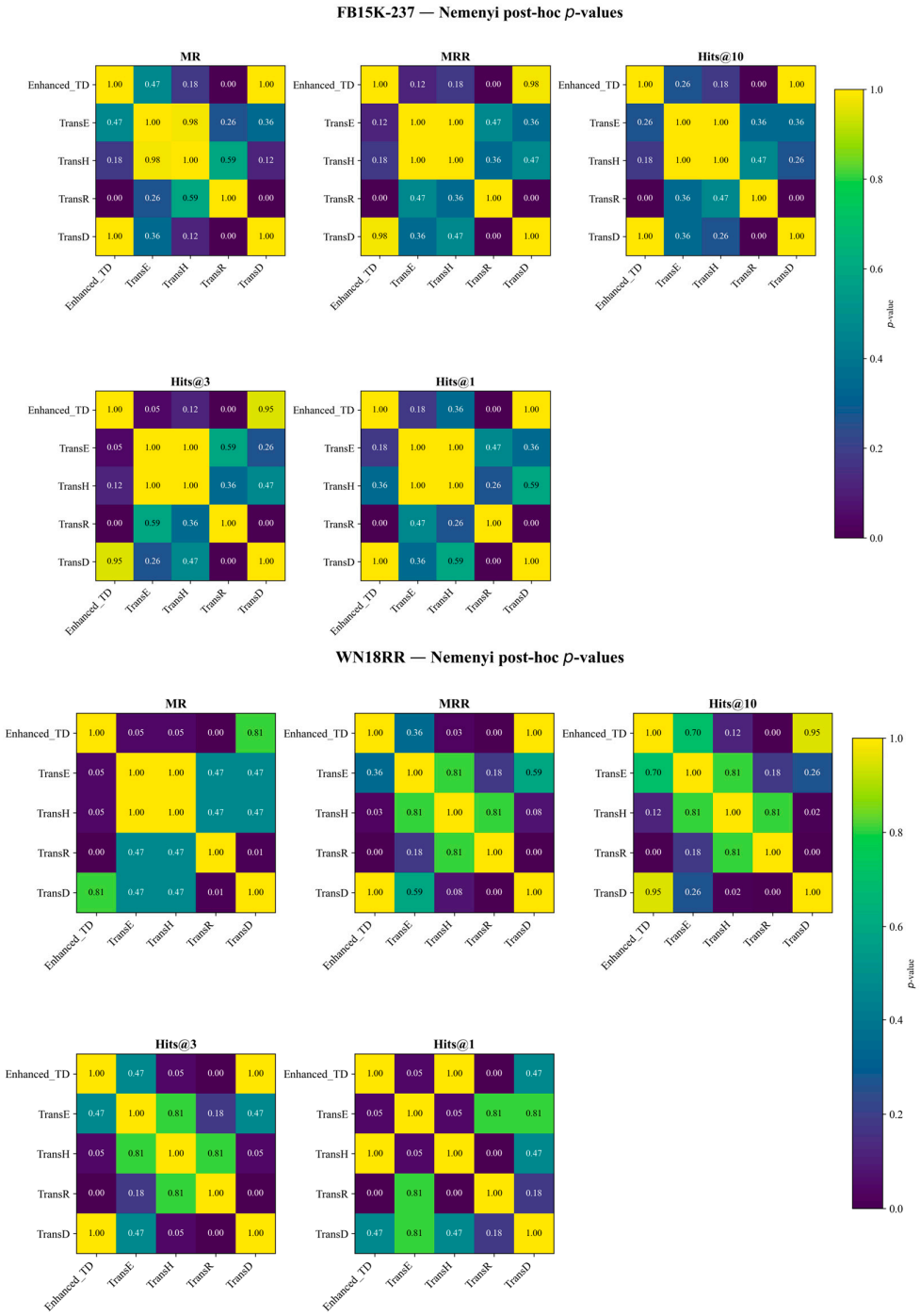


Figure 11. Heat map of Friedman test results across evaluation metrics.

To comprehensively assess the independent contributions and complementary effects of each enhancement module, we designed four ablation variants of the baseline TransD model: (1) the original TransD; (2) TransD with only the attribute projection module; (3) TransD with only the dual-limit loss module; and (4) the fully enhanced TransD combining both modules. All experiments were conducted on the FB15K-237 and WN18RR benchmarks. Observing that model performance stabilized around an embedding dimensionality of 200, we fixed this dimension for the ablation study. As shown in Figure 12, both the attribute projection and dual-limit loss modules yield significant gains over the original TransD, albeit with differing emphases: the attribute projection module—by explicitly encoding inherent entity attributes into the dynamic mapping matrices—enhances semantic

discrimination in the embedding space, leading to superior performance on ranking-quality metrics such as MRR and Hits@1; the dual-limit loss module—by imposing independent upper and lower bounds on positive- and negative-sample scores—optimizes mean rank and Hits@K, improving retrieval precision and recall. When combined, the fully enhanced TransD model achieves the best results across all evaluation metrics, thereby validating the complementarity and synergistic value of the two strategies.

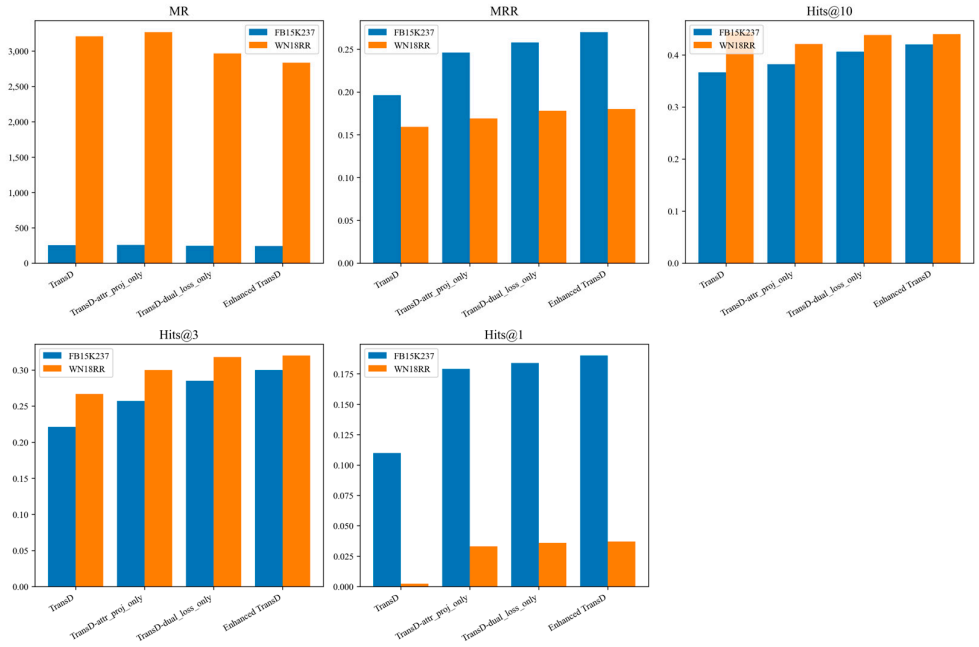


Figure 12. Results of the ablation experiments.

Notably, the attribute projection module provides fine-grained representations of domain attributes, reducing confusion among semantically similar triples; the dual-limit loss module enforces clearer separation at decision boundaries, mitigating misclassification in challenging cases and enhancing generalization. Their joint application yields a balanced improvement in both precision and recall, with stable, sustained performance gains and no signs of overfitting. Moreover, consistent results across datasets of varying scale and complexity further attest to the stability and generalizability of the proposed approach.

In summary, the enhanced TransD model achieves outstanding performance on the benchmark datasets FB15K-237 and WN18RR, delivering stable and superior results in entity ranking, accuracy, F1 score, and various Hits@k metrics. These findings not only validate the model's effectiveness in capturing complex semantic relationships among entities but also demonstrate its robustness against data sparsity and noise. Building on this foundation, we will next apply the enhanced TransD model to representation learning on the constructed California Traffic Accident Knowledge Graph, with the aim of uncovering latent semantic patterns in real-world accident data and thereby providing a solid theoretical and practical basis for optimizing accident prevention and safety management strategies.

5.2. Model Application

The improved TransD model constructs dynamic mapping matrices for each entity–relation pair, accommodating the diversity of entities and relations while enabling flexible projections that avoid the computational overhead and parameter bloat associated with large static matrices. By embedding nodes and edges into a low-dimensional space that preserves both structural and semantic information, it furnishes a reliable foundation for pattern discovery and relational inference. In the traffic accident domain, a knowledge graph can integrate heterogeneous data—such as accident location, time, weather, and road

conditions—into a semantic network encompassing entities (e.g., vehicles, drivers, environments) and their multifaceted relationships, thereby elucidating deep couplings among risk factors. Although graph neural networks excel at capturing spatiotemporal and topological features in traffic accident data, they often fall short in uniformly modeling multi-source semantic information. This shortcoming motivates the adoption of an enhanced knowledge graph embedding to boost generalization and inference capabilities. Compared to TransE, TransH, TransR, and the baseline TransD, the improved TransD is more robust when handling one-to-many, many-to-one, and many-to-many relations, accurately capturing subtle variations in complex associations such as "Accident–RoadSegment" and "Accident–Weather". Applying TransD embedding to the self-constructed Traffic Accident Knowledge Graph (TAKG) yields vector representations rich in both semantic and structural cues. These embeddings serve as highly discriminative inputs for downstream tasks—including clustering analysis, anomaly detection, and link prediction—thereby supporting accident root-cause analysis and intelligent safety decision-making. Detailed statistics for TAKG are provided in Table 4.

Table 4. Detailed statistics of the Traffic Accident Knowledge Graph (TAKG).

Data Set	Number of Entities	Number of Relationships	Size of Training Set	Size of Validation Set	Size of Test Set
TAKG	33,755	115	460,005	57,495	57,500

As shown in Figure 13, the multidimensional evaluation of the improved TransD model on the self-constructed Traffic Accident Knowledge Graph (TAKG) reveals that, as the embedding dimension increases from 50 to 250, all ranking metrics (MR, MRR, Hits@K) and classification metrics (Accuracy, F1) exhibit clear upward trends, achieving optimal performance at 250 dimensions. Beyond this threshold, slight declines in each metric indicate the onset of overfitting. At 250 dimensions, the model attains its lowest MR and highest MRR, with Hits@1, Hits@3, and Hits@10 reaching their peaks, while Accuracy and F1 achieve approximately 0.51 and 0.68, respectively. These results underscore the significant impact of incorporating attribute projection vectors and dual-limit scoring loss within the dynamic mapping mechanism to enhance the representational fidelity and discriminative power of TAKG embeddings.

As Table 5 illustrates, in the Mean Rank (MR) metric, the model constrained by low embedding dimensions (50–150) exhibits limited mapping capacity, with entity-ranking errors fluctuating around 985. When the dimension increases beyond 200, the embedding space can better accommodate the additional attribute information introduced by the attribute projection vectors, thereby enlarging the distances between accident and node embeddings. Consequently, MR declines sharply and reaches its minimum—approximately 950—at 250 dimensions. This pattern aligns with the behavior of the dynamic mapping matrices in the TransD framework, which effectively integrates inherent entity attributes and relational semantics. Beyond 250 dimensions, further dimensional increases lead to a steep rise in parameter count and heightened sensitivity to noise, causing a slight rebound in MR—consistent with the overfitting commonly observed in high-dimensional spaces.

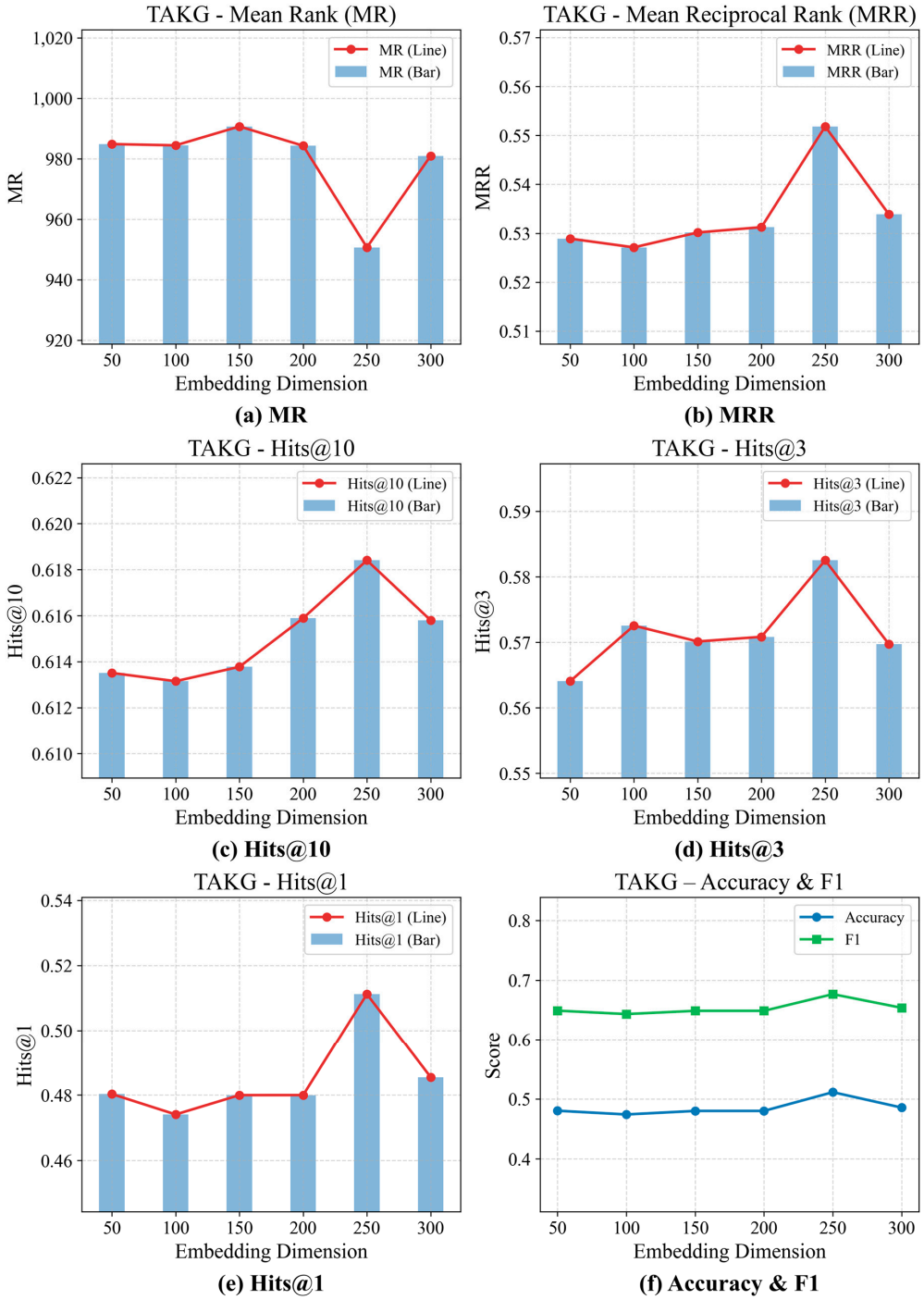


Figure 13. Performance of the enhanced TransD model on the Traffic Accident Knowledge Graph (TAKG) across different embedding dimensions.

The Mean Reciprocal Rank (MRR) reflects the model's effectiveness in ranking correct entities near the top of candidate lists. In the lower embedding range (50–150), due to indistinct margins between positive and negative samples, MRR remains between 0.528 and 0.530. As the dimension increases, the dual-limit loss independently enforces suitable margins for positive and negative samples, concentrating high-scoring positives toward the top. Consequently, MRR begins to rise after 200 dimensions and peaks at 0.552 at 250 dimensions—substantially above the low-dimensional stage. Although MRR slightly declines at 300 dimensions, it still exceeds its initial levels, indicating that the dual-limit loss maintains strong discriminative power in the expanded embedding space.

Table 5. Experimental results of representation learning on the Traffic Accident Knowledge Graph using the enhanced TransD model.

dim	MR	MRR	H@1	H@3	H@10	Acc	F1
50	984.8603	0.528903	0.480301	0.564064	0.613507	0.480301	0.648924
100	984.4376	0.527122	0.474052	0.572557	0.613148	0.474052	0.643196
150	990.6853	0.530165	0.479965	0.570133	0.613774	0.479965	0.648617
200	984.3071	0.53125	0.479977	0.570846	0.61589	0.479977	0.648627
250	950.6806	0.551829	0.511194	0.582533	0.618406	0.511194	0.676543
300	980.9584	0.533855	0.485461	0.569704	0.615797	0.485461	0.653617

Similarly, Hits@10, Hits@3, and Hits@1 show steady upward trends across dimensions: Hits@10 increases from 0.6135 to 0.6185 at 250 dimensions, Hits@3 rises from 0.5635 to 0.5830, and Hits@1 jumps from 0.4800 to 0.5105 at 250 dimensions. These gains primarily stem from the inclusion of attribute projection vectors, which draw true positives closer within the top candidate positions, and the dual-limit loss, which further separates positives from near-neighbor negatives—most prominently under the strict Hits@1 criterion. The slight declines observed at 300 dimensions further corroborate the general tendency toward overfitting and noise induction in excessively high-dimensional spaces.

In the node-classification tasks assessed by Accuracy and F1-score, the model likewise attains optimal performance at 250 dimensions: Accuracy improves from 0.48 at 50 dimensions to 0.51 at 250 dimensions, and F1-score rises from 0.65 to 0.68; at 300 dimensions, both metrics decline to approximately 0.49 and 0.655, respectively. The gains in classification accuracy stem primarily from the dual-limit loss, which simultaneously elevates the scores of negative samples and suppresses those of positive samples, thereby establishing clearer decision boundaries in the mid-dimensional embedding space. The improved F1-score further indicates enhanced detection of minority classes, such as rare accident types. Similar performance trends observed on general benchmarks like FB15K-237 and WN18RR suggest the broad applicability of this enhancement across diverse knowledge graphs.

In our experiments, we augment the TransD dynamic mapping with attribute projection vectors and adopt a dual-limit scoring loss. Each attribute projection vector is dedicated to critical accident attributes—such as time, location, weather, and severity—enabling entity embeddings to encode multidimensional attribute information alongside relational mappings. This design allows the model to extract richer semantic features even at low to mid embedding dimensions, yielding superior performance on fine-grained ranking metrics like Hits@1. The dual-limit loss applies distinct margins to positive and negative samples: on one hand, it ensures positive samples cluster without over-compression; on the other, it enforces a robust separation between negatives and positives. Together, these mechanisms synergize in the mid to high embedding space, driving MR, MRR, Hits@K, and classification metrics to peak at 250 dimensions. Below this threshold, the embedding space is too constrained to balance attribute encoding with relation mapping; above it, excessive parameters induce overfitting and performance degradation.

6. Discussion

In this study, we conducted a series of experiments on our Enhanced TransD model across two public benchmark datasets (FB15K-237 and WN18RR) as well as on a self-constructed Traffic Accident Knowledge Graph (TAKG). By combining ablation studies with statistical significance tests, we systematically evaluated the individual and joint contributions of the "attribute projection vector" and "dual-limit scoring loss" modules. The following analysis is organized around four dimensions: overall performance improvement, module functionality and complementarity, sensitivity to embedding dimensionality, and model robustness.

(1) Overall Performance Improvement.

On the FB15K-237 dataset, Enhanced TransD reduced the mean rank (MR) by approximately 30 positions compared to the original TransD, increased the Mean Reciprocal Rank (MRR) by 0.03–0.05, and achieved average gains of 0.10, 0.12, and 0.07 on Hits@1, Hits@3, and Hits@10, respectively. These results demonstrate that the attribute projection mechanism and the dual-limit scoring loss work synergistically to sharpen the model's ability to distinguish positive from negative triples. Similar improvements were observed on WN18RR: MR dropped by over 25%, MRR rose by about 0.07, and Hits@k metrics increased by more than 0.10 on average, indicating that our enhancements generalize well to large, path-complex knowledge graphs.

(2) Module Functionality and Complementarity.

Ablation experiments revealed distinct roles for the two modules. The variant with only the attribute projection vector excelled on ranking-quality metrics (MRR and Hits@1), suggesting that explicitly embedding fine-grained domain attributes (e.g., "accident severity", "road conditions") helps distribute entities more coherently in semantic space, thereby reducing confusion among neighboring entities. The variant with only the dual-limit scoring loss yielded greater gains in MR and Hits@K (K > 1), underscoring the value of independently regulating positive and negative score boundaries to boost recall. When both modules were combined, all evaluation metrics reached their peak values, confirming their complementary effects on ranking precision and overall retrieval efficiency.

(3) Sensitivity to Embedding Dimensionality.

Across embedding dimensions, performance followed a "gain-then-deteriorate" curve. In the low-dimensional range (50–150 dimensions), capacity constraints limited the model to capturing only basic dynamic mapping features. As the dimension increased to 200–250, the additional semantic information introduced by the attribute projection vector, coupled with the boundary-control benefits of the dual-limit loss, yielded peak performance across all ranking and classification metrics. Beyond 250 dimensions, however, a slight performance decline appeared, reflecting the onset of parameter redundancy and overfitting. This trend highlights the necessity of balancing embedding dimensionality against computational cost to avoid excessive capacity that undermines generalization.

(4) Model Robustness.

Using the Friedman test followed by Nemenyi post-hoc comparisons, we observed that all major metrics attained χ^2 statistics corresponding to $p < 0.05$, strongly rejecting the null hypothesis of no performance differences among models. This confirms the statistical significance and robustness of the Enhanced TransD's improvements. Moreover, the ablation gains were consistently reproducible across datasets of varying scale and attribute complexity, demonstrating our method's general applicability and reliability for traffic accident knowledge graph embedding.

Taken together, the experimental outcomes and statistical analyses indicate the following:

(1) The attribute projection vector enhances sensitivity to semantic differences by explicitly encoding domain attributes;
(2) The dual-limit scoring loss independently constrains positive and negative sample boundaries, improving recall and classification accuracy;
(3) Their combined use achieves a balanced optimization of representational capacity and discriminative power;
(4) Embedding dimensionality and hyperparameter settings critically affect model generalization and must be tuned in accordance with dataset characteristics.

7. Conclusions

The California Traffic Accident Knowledge Graph (TAKG) developed in this study, together with the Enhanced TransD-based representation method, provides multidimensional, interpretable tools for accident-risk prediction and causal analysis to traffic management authorities, emergency response agencies, and urban planners. Traffic agencies may leverage the knowledge graph to identify high-risk road segments and characteristic accident factors, thereby guiding targeted remediation and prevention measures. Emergency responders can use model outputs to optimize deployment and improve response efficiency. Urban planners can assess the safety of proposed roads or intersections and validate design alternatives via simulation platforms, facilitating intelligent traffic safety management and data-driven decision optimization.

This study addresses the challenges posed by the high-dimensional heterogeneity and complex mapping relationships inherent in traffic accident data by systematically developing a knowledge graph construction framework and an enhanced representation-learning method. The main conclusions are as follows:

1. Knowledge Graph Construction: Guided by a top-down ontology design principle, we constructed a multi-layered traffic accident knowledge graph that integrates entities such as accident events, involved parties, victims, and environmental factors. This graph efficiently consolidates over 120 accident attributes and supports intuitive visualization, thereby providing a unified data platform for causal relationship mining and risk prediction.
2. Improved TransD Model: By introducing attribute projection vectors, domain-specific features—such as "accident severity" and "road slipperiness"—are explicitly encoded into the dynamic mapping matrices. Furthermore, a dual-limit scoring loss enforces clearer separation between positive and negative samples in the embedding space. This enhancement balances representational expressiveness with parameter efficiency and overcomes performance bottlenecks in complex mapping scenarios.
3. Experimental Validation: Experimental validation was conducted on the FB15K-237 and WN18RR benchmarks, including comparative evaluations, statistical significance testing, and ablation studies, to comprehensively assess the performance of the enhanced TransD model. Upon application to our self-constructed California Traffic Accident Knowledge Graph (TAKG), Enhanced TransD achieved optimal performance across all evaluation metrics under medium-to-high embedding dimensions, markedly improving entity ranking and classification accuracy and confirming the synergistic benefits of attribute projection and dual-limit loss. These results demonstrate that the proposed approach possesses strong generalization and discrimination capabilities in multi-source heterogeneous knowledge graph representation learning.

By integrating entity–attribute projection vectors and a dual-limit scoring loss into the TransD framework, this work significantly advances the modeling of complex one-to-many, many-to-one, and many-to-many relations, thereby enriching the ecosystem of translation-based knowledge graph embeddings. Compared to prior studies, it introduces fine-grained domain attribute encoding and explicit positive–negative boundary control. The resulting interpretable embeddings can be deployed within intelligent traffic simulation and early-warning systems to enhance the accuracy and responsiveness of accident-risk predictions, providing traffic management authorities with robust, data-driven decision support.

Moreover, this research opens several avenues for future innovation. First, the manual selection of entity attributes relies heavily on domain expertise; subsequent work should explore automated attribute screening and the integration of multimodal data. Second, although this study focuses on translation-based embeddings, the joint application of graph neural networks and knowledge graph embeddings remains underexplored; future

GNN–KG hybrid models may further improve spatiotemporal feature extraction. Finally, extending the proposed framework to support online incremental learning and real-time accident-warning systems would meet the pressing demands of intelligent transportation safety management.

Author Contributions: Formal analysis, X.L.; Funding acquisition, X.L.; Methodology, X.L., H.W. (Haopeng Wu), D.Y. and Y.C.; Resources, X.L.; Software, H.W. (Haopeng Wu); Supervision, D.Y. and H.W. (Hao Wu); Validation, H.W. (Haopeng Wu); Writing—original draft, X.L. and H.W. (Haopeng Wu); Writing—review and editing, H.W. (Haopeng Wu), D.Y., Y.C. and H.W. (Hao Wu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Innovation Strategy Research Project of the Fujian Provincial Department of Science and Technology [2024R0124].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors would like to thank Jimei University for providing equipment support for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Depaire, B.; Wets, G.; Vanhoof, K. Traffic Accident Segmentation by Means of Latent Class Clustering. *Accid. Anal. Prev.* **2008**, *40*, 1257–1266. [CrossRef] [PubMed]
2. Fan, Z.; Liu, C.; Cai, D.; Yue, S. Research on Black Spot Identification of Safety in Urban Traffic Accidents Based on Machine Learning Method. *Saf. Sci.* **2019**, *118*, 607–616. [CrossRef]
3. Yu, X. Analysis of injury and death characteristics of residents in Haidian District, Beijing, 2010–2015. *Dis. Surveill.* **2019**, *34*, 166–170.
4. Mao, Y.-P.; Yu, F.-Q.; Sun, Y.-Y.; Tang, Z.-G. Road traffic accident data mining and its application research. *Traffic Transp.* **2020**, *33*, 106–111.
5. Xu, P.; Jiang, K.; Wang, Z.-H.; Zhu, Z. Significant analysis of objective factors of road traffic accidents based on rough set. *J. East. China Jiaotong Univ.* **2017**, *34*, 66–71.
6. Zhi, Y.; Wang, D.-S.; Cong, H.-Z.; Rao, Z.-B. Deep Data Mining Technology and Application of Road Traffic Accident Data: A Case Study of Shenzhen. *Urban. Transp. China* **2018**, *16*, 28–32,61.
7. Ma, Z.; Mei, G.; Cuomo, S. An Analytic Framework Using Deep Learning for Prediction of Traffic Accident Injury Severity Based on Contributing Factors. *Accid. Anal. Prev.* **2021**, *160*, 106322. [CrossRef]
8. Hadjidimitriou, N.S.; Lippi, M.; Dell'Amico, M.; Skiera, A. Machine Learning for Severity Classification of Accidents Involving Powered Two Wheelers. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 4308–4317. [CrossRef]
9. Peng, C.; Xia, F.; Naseriparsa, M.; Osborne, F. Knowledge Graphs: Opportunities and Challenges. *Artif. Intell. Rev.* **2023**, *56*, 13071–13102. [CrossRef]
10. Fu, L.-J.; Cao, Y.; Bai, Y.; Leng, J.-W. Development Status and Prospects of Knowledge Graphs in Domestic Vertical Domains. Comput. *Appl. Res.* **2021**, *38*, 3201–3214.
11. Zhang, Z.-J.; Ni, Z.-N.; Liu, Z.-H.; Xia, S.-D. Research on Dynamic Relation Prediction Method for Financial Knowledge Graphs. Data Anal. *Knowl. Discov.* **2023**, *7*, 39–50. [CrossRef]
12. Pu, T.-J.; Tan, Y.-P.; Peng, G.-Z.; Xu, H.-F.; Zhang, Z.-H. Construction and Application of Knowledge Graph in the Power Domain. *Power Syst. Technol.* **2021**, *45*, 2080–2091.
13. Wang, C.-Y.; Zheng, Z.-L.; Cai, X.-Q.; Huang, J.-H.; Su, Q.-M. A Review of the Application of Knowledge Graphs in the Medical Field. *J. Biomed. Eng.* **2023**, *40*, 1040–1044.
14. Lin, J.; Zhao, Y.; Huang, W.; Liu, C.; Pu, H. Domain Knowledge Graph-Based Research Progress of Knowledge Representation. *Neural Comput. Appl.* **2021**, *33*, 681–690. [CrossRef]
15. Zhang, Z.; Li, Z.; Liu, H.; Xiong, N.N. Multi-Scale Dynamic Convolutional Network for Knowledge Graph Embedding. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 2335–2347. [CrossRef]

16. Qiu, C.; Xu, H.; Bao, Y. Modified-DBSCAN Clustering for Identifying Traffic Accident Prone Locations. In *Proceedings of the Intelligent Data Engineering and Automated Learning–IDEAL 2016, Yangzhou, China, 12–14 October 2016*; Yin, H., Gao, Y., Li, B., Zhang, D., Yang, M., Li, Y., Klawonn, F., Tallón-Ballesteros, A.J., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 99–105.
17. Li, J.; He, J.; Liu, Z.; Zhang, H.; Zhang, C. Traffic Accident Analysis Based on C4.5 Algorithm in WEKA. *MATEC Web Conf.* **2019**, *272*, 01035. [CrossRef]
18. Lv, Y.; Tang, S.; Zhao, H.; Li, S. Real-Time Highway Accident Prediction Based on Support Vector Machines. In Proceedings of the 2009 Chinese Control and Decision Conference, Guilin, China, 17–19 June 2009; pp. 4403–4407.
19. Hu, Z.; Zhou, J.; Zhang, E. Improving Traffic Safety through Traffic Accident Risk Assessment. *Sustainability* **2023**, *15*, 3748. [CrossRef]
20. Shokry, S.; Rashwan, N.K.; Hemdan, S.; Alrashidi, A.; Wahaballa, A.M. Characterization of Traffic Accidents Based on Long-Horizon Aggregated and Disaggregated Data. *Sustainability* **2023**, *15*, 1483. [CrossRef]
21. Wang, S.; Yan, C.; Shao, Y. LSTM Road Traffic Accident Prediction Model Based on Attention Mechanism. In Proceedings of the 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 26–28 April 2023; pp. 215–219.
22. Wang, H.; Liang, G. Association Rules Between Urban Road Traffic Accidents and Violations Considering Temporal and Spatial Constraints: A Case Study of Beijing. *Sustainability* **2025**, *17*, 1680. [CrossRef]
23. Wang, Y.; Zhai, H.; Cao, X.; Geng, X. Cause Analysis and Accident Classification of Road Traffic Accidents Based on Complex Networks. *Appl. Sci.* **2023**, *13*, 12963. [CrossRef]
24. Zhang, L.; Zhang, M.; Tang, J.; Ma, J.; Duan, X.; Sun, J.; Hu, X.; Xu, S. Analysis of Traffic Accident Based on Knowledge Graph. *J. Adv. Transp.* **2022**, *2022*, 3915467. [CrossRef]
25. Wang, C.-H.; Ji, Y.-T.-S.; Ruan, L.; Luhwago, J.; Saw, Y.-X.; Kim, S.; Ruan, T.; Xiao, L.-M.; Zhou, R.-J. Multisource Accident Datasets-Driven Deep Learning-Based Traffic Accident Portrait for Accident Reasoning. *J. Adv. Transp.* **2024**, *2024*, 8831914. [CrossRef]
26. Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; Yuan, N.J. Multi-Modal Knowledge Graph Construction and Application: A Survey. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 715–735. [CrossRef]
27. Zhu, J.; Han, X.; Deng, H.; Tao, C.; Zhao, L.; Wang, P.; Lin, T.; Li, H. KST-GCN: A Knowledge-Driven Spatial-Temporal Graph Convolutional Network for Traffic Forecasting. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15055–15065. [CrossRef]
28. Nickel, M.; Murphy, K.; Tresp, V.; Gabrilovich, E. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* **2016**, *104*, 11–33. [CrossRef]
29. Zhu, H.; Xu, D.; Huang, Y.; Jin, Z.; Ding, W.; Tong, J.; Chong, G. Graph Structure Enhanced Pre-Training Language Model for Knowledge Graph Completion. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2697–2708. [CrossRef]
30. Zhang, Z.-H.; Qian, Y.-R.; Xing, Y.-N.; Zhao, X. A Review of Representation Learning Methods Based on TransE. *Comput. Appl. Res.* **2021**, *38*, 656–663.
31. Bordes, A.; Usunier, N.; Garcia-Durán, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2, Lake Tahoe, NV, USA, 5–10 December 2013*; Curran Associates, Inc.: New York, NY, USA, 2013; pp. 2787–2795.
32. Chang, P.; Cao, Y. Research on Improved TransH Model in Knowledge Representation and Reasoning. *J. Guangxi Univ.* **2020**, *45*, 321–327.
33. Zhang, Z.; Jia, J.; Wan, Y.; Zhou, Y.; Kong, Y.; Qian, Y.; Long, J. TransR*: Representation Learning Model by Flexible Translation and Relation Matrix Projection. *IFS* **2021**, *40*, 10251–10259. [CrossRef]
34. Li, Z.; Huang, R.; Zhang, Y.; Zhu, J.; Hu, B. Two Flexible Translation-Based Models for Knowledge Graph Embedding. *IFS* **2023**, *44*, 3093–3105. [CrossRef]
35. Ji, G.; He, S.; Xu, L.; Liu, K.; Zhao, J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015*; Zong, C., Strube, M., Eds.; Association for Computational Linguistics: Beijing, China, 2015; pp. 687–696.
36. Zhou, X.; Niu, L.; Zhu, Q.; Zhu, X.; Liu, P.; Tan, J.; Guo, L. Knowledge Graph Embedding by Double Limit Scoring Loss. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 5825–5839. [CrossRef]