Data Element Ownership Confirmation and Circulation in the Intelligent Robot Industry Empowered by Generative AI

Wen Zou, Bin Fu, Shiyun Liu, Manyi Li

Xi'an Fanyi University, Xi'an 710105, Shaanxi, China

Corresponding author: Manyi Li (e-mail: lmyhana@163.com)

ABSTRACT This paper proposes a generative artificial intelligence method, RAG-CGM (Retrieval-Augmented Generation with Constrained Generation Model), to address the coexistence of multiple data-ownership definitions and the low circulation efficiency of data elements in the intelligent robot industry. A semantic representation model is developed for heterogeneous data sources containing ownership, scenario, quality, and safety attributes, including sensor data, device logs, interaction records, and text documents generated in robotic systems. By modeling data-generation methods, participating entities, and mutual contributions among sources, the method builds a structured mapping for data-source traceability and ownership determination. After this mapping is established, a retrieval-augmented generation mechanism constructs the full context of data assets and demand semantics. Conditional generation then produces semantically aligned and interpretable supply-demand matches under ownership-boundary and compliancerule constraints. Finally, an optimization function based on generative scoring ranks candidate matches and dynamically updates validity conditions, supporting efficient circulation after ownership confirmation. Experimental results show that the method achieves an ownership identification accuracy of 0.913 and an F1 score of 0.901, about 4.4% higher than the second-place RAG baseline. The circulation success rate reaches 0.892, outperforming all comparison methods and providing interpretable ownership tracing for data-element governance in intelligent robotics.

INDEX TERMS data ownership, intelligent robots, generative ai, sensor data circulation, rag-cgm

I. INTRODUCTION

AGAINST the backdrop of the deepening digital economy, data elements have become core resources driving industrial value reconstruction and production mode transformation. The large-scale multimodal data generated during the R&D (research and development), manufacturing, and service processes of the intelligent robot industry exhibits significant asset attributes and collaborative value [1,2]. Data resources involve complex ownership relationships and benefit distribution mechanisms during cross-entity sharing and reuse, making ownership confirmation and circulation key links restricting the release of data value [3,4]. Generative AI demonstrates a new technological paradigm in semantic modeling and knowledge representation, providing an important support path for ownership identification and circulation optimization in complex data environments, and has significant theoretical and practical value for improving the efficiency of industrial data element allocation [5,6]. The generative AI method RAG-CGM proposed in this paper, based on retrieval augmentation and constraint generation, aims to achieve integrated collaboration in data ownership confirmation and circulation by introducing external knowledge retrieval and structured constraint generation. Existing research faces multiple bottlenecks in the governance of industrial data elements. Data sources have a distributed nature, and the interactions between production, platform, and application entities continue to change over time, making it difficult to understand how the data is generated, where it comes from, and what the ownership boundaries are [7,8]. Multimodal data has different structures and semantics than other types of data. Traditional methods used to determine who owns data based on the field structure are unable to represent complex semantic content; therefore, the allocation of rights becomes inconsistent as a result [9,10]. Circulating data uses static catalogs and manual review processes, leading to semantic expression inconsistencies between supply and demand sides. The matching process does not use a common representation model, resulting in reduced efficiency and accuracy [11]. Compliance constraints and security strategies are usually not part of the circulation process and cannot be tied to ownership determination; therefore, it is difficult for the results of ownership determination to directly inform future decisions about circulating data [12,13]. It is imper-

ative that generative AI is brought into fruition to facilitate determining and circulating data elements.

There are a number of attempts being made to improve the issues mentioned above, both in the academic world and the industry. Many have begun exploring blockchain-based techniques for securely verifying ownership through the implementation of traceable and immutable data by documenting data operations and rights through a distributed ledger; however, most of these techniques cannot fully utilize their ability to represent multimodal semantics, which impacts their effectiveness when working with complex data generation relationships [14,15]. In addition, various forms of ontology modeling and knowledge graphing have found some success at providing structured representations of semantics by constructing data entities and networks of relationships to describe ownership; however, since they require manual design of models/rules to define the relationships, they have limited capability for scalability and dynamic adaptability [16,17]. Data circulation research is highly focused on working through matching similarity and retrieving through tags (small scale), but there has been limited ability to provide deep meaning for demand semantics in order to create effective cross scenario matching [18,19]. There have been studies incorporating privacy computing and federated learning for greater security of data sharing while creating a foundation for compliance assurance; however, these do not closely link together ownership confirmation results with a unified ownership confirmation circulation collaborative mechanism [20,21]. Currently, there continues to be disconnects between ownership identification, semantic alignment and circulation matching making it difficult to create opportunity for data element allocation in various complex industrial environments.

Addressing the core issues of insufficient accuracy in ownership identification and limited efficiency in data flow matching, this paper proposes a generative AI method, RAG-CGM, based on retrieval augmentation and constraint generation, for data element ownership confirmation and flow in the intelligent robot industry. Starting with multi-source heterogeneous data, the study models the data generation chain and uses gener ative AI to jointly encode data content and contextual semantics, forming a unified semantic embedding representation. Based on this, ownership constraint modeling embeds subject relationships, contribution levels, and usage boundaries into the representation space, enabling automatic determination and expression of data rights structures. Furthermore, a semantic indexing system is constructed to perform retrieval-augmented generation (RAG) on data asset and demand descriptions, injecting contextual information into the generation model to complete supply and demand semantic alignment and matching inference under the constraints of ownership boundaries and compliance rules. Matching results are sorted and dynamically adjusted through a generation scoring function to form the optimal matching path for flow decision-making. This research achieves collaborative modeling of ownership identification and circulation matching, introduces generative AI into key aspects of data element governance, and makes systematic improvements in multimodal semantic modeling, constraint-driven generation, and supply and demand alignment mechanisms, providing an interpretable reasoning path for ownership tracing and matching processes.

II. METHODS

A. OVERALL FRAMEWORK CONSTRUCTION

Focusing on the problems of unclear ownership and inefficient supply-demand matching in the process of data element confirmation and circulation in the intelligent robot industry, this paper constructs an overall framework for a generative AI method oriented towards the collaborative optimization of "confirmation of ownership and circulation". The generative AI architecture and data circulation based on retrieval augmentation and constraint generation are shown in Figure 1.

Figure 1 illustrates the overall framework structure of the generative AI-driven data element ownership confirmation and circulation system. Starting with multi-source heterogeneous data input, a unified feature expression is achieved through data modeling and semantic representation modules. Based on this, an ownership relationship identification model characterizes the correlation structure between data and multiple entities, and a retrieval augmentation mechanism is used to complete and strengthen semantic information. Furthermore, ownership rules and compliance conditions are introduced into the constraint generation module to achieve collaborative inference between ownership confirmation and circulation. The final output and specification for the matching and configuration of data elements have been made in the circulation matching and ranking optimization module. The various functional modules have a closed loop structure with representation transmission and constraint feedback to maintain consistency and traceability between circulation path selection and ownership determination results, thus providing a uniform process of data governance across multiple complex industrial scenarios.

B. MULTI-SOURCE HETEROGENEOUS DATA MODELING FOR THE INTELLIGENT ROBOT INDUSTRY

There is a large degree of heterogeneity in the intelligent robot industry data due to the different sources, structures, sizes, and semantics around this type of data. This data exists in several different formats simultaneously across multiple modalities including device operation trajectories, signals received during sensing/ interaction, records of interaction between people and robots/devices, and business text documents. These different modalities differ not only in statistical distribution but also in semantic granularity. Therefore, this study treats the industry data set as a multi-model multi-source grouping in which the i-th element is referred to as $d_i = \left\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}, \ldots, x_i^{(M)}\right\}$, with M being the number of models in the sample with each

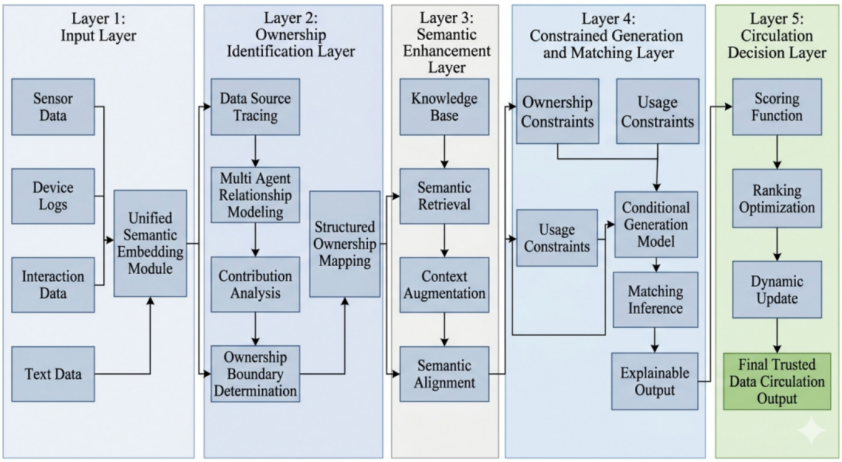

FIGURE 1. Overall architecture of industrial data element ownership confirmation and circulation

$x_i^{(m)}$ representing the actual observation for that model. Not only does this collection contain the actual data, but it also contains implicit (but critical) information related to the time of creation of the data, where it comes from, and the setting for which the data is created, thus providing the basis for unified modeling. To address the differences in feature expression among different modalities, this paper constructs a dedicated encoder for each modality, mapping the original observations to latent representations. Let the encoding function of the m-th mode be $f_m(\cdot;\theta_m)$, then the representation of sample d_i in this mode is:

$$h_i^{(m)} = f_m\left(x_i^{(m)};\theta_m\right) \tag{1}$$

In Formula (1), $h_i^{(m)} \in R^{p_m}$ is the modal feature vector; θ_m is the corresponding encoder parameter; p_m is the representation dimension of the modality. The feature dimensions and statistical properties of different modalities are not consistent, and direct concatenation can lead to scale inconsistencies and noise amplification. Therefore, this paper further introduces a shared semantic projection layer to map each modal representation to a unified latent space:

$$z_i^{(m)} = W_m h_i^{(m)} + b_m \tag{2}$$

In Formula (2), $W_m \in R^{p\times p_m}$ is the modal projection matrix; $b_m \in R^p$ is the bias term; $z_i^{(m)} \in R^p$ is the modal semantic representation under a unified dimension; p is the common space dimension. This process aims to resolve the expression differences between modalities. The selection of a shared semantic projection layer is specifically dictated by the high-dimensional sparsity of industrial robot data (e.g., sensor trajectories vs. business texts). By utilizing a linear projection to a common manifold, we preserve the global topological structure of the data while mitigating the 'modality gap' that often leads to vanishing gradients in decentralized attention mechanisms. This allows data from different sources to enter a stable semantic coordinate system.

Within the unified latent space, this paper uses a semantic attention mechanism to perform weighted fusion of multimodal representations, enabling the model to focus on key semantics related to ownership recognition and circulation matching [22,23]. Let the modal weight of sample d_i be $\alpha_i^{(m)}$, then there is:

$$\alpha_i^{(m)} = \frac{\exp\left(q^\top z_i^{(m)}\right)}{\sum_{j=1}^{M}\exp\left(q^\top z_i^{(j)}\right)}, \quad z_i = \sum_{m=1}^{M}\alpha_i^{(m)} z_i^{(m)} \tag{3}$$

In Formula (3), $q \in R^p$ is the semantic query vector, and z_i is the fusion representation of sample d_i. The weight $\alpha_i^{(m)}$ characterizes the contribution intensity of the m-th modality to the comprehensive semantics of the current sample, while the fusion vector z_i carries the aggregation result of multimodal information in the unified space. Compared with simple averaging, this mechanism makes highly relevant modalities occupy a higher proportion in the representation and weakens the interference of redundant modalities on subsequent tasks.

The data of the intelligent robot industry also has strong temporal sequence, and the data value is often determined by continuous state changes and dynamic interaction processes. In order to maintain temporal correlation, this paper embeds the time index into the unified representation and constructs a sequence state representation:

$$u_{i,t} = z_{i,t} + p_t \tag{4}$$

In Formula (4), $u_{i,t}$ represents the temporal enhanced representation of sample d_i at time t; $z_{i,t}$ is the fused semantic vector at that time; p_t is the temporal position code. This processing enables the model to not only identify static content features but also retain the sequential relationship and evolutionary trajectory in the data generation process, thereby incorporating the temporal link in the ownership formation process into a unified expression framework. Furthermore, to bridge the gap between technical features

and real-world legal governance, the model introduces a "Contractual Event Layer" within the temporal position code p_t. This layer functions as a logical gate that allows external legal attributes and contractual ownership shifts to override or modify the continuous state features, ensuring that discrete legal changes are reflected in the evolutionary trajectory alongside technical interaction data.

Based on the unified representation, this paper jointly characterizes the structural and semantic attributes of industrial data to form a multi-source heterogeneous data modeling result adapted to the tasks of ownership confirmation and circulation. Let the comprehensive representation of the sample be r_i, then:

$$r_i = \phi\left(z_i, u_{i,1}, u_{i,2}, \ldots, u_{i,T}\right) \tag{5}$$

In Formula (5), $\phi(\cdot)$ represents the feature aggregation function; T is the time window length; r_i serves as the input representation for subsequent ownership identification, semantic retrieval, and constraint generation. This representation compresses modal differences, subject associations, and dynamic processes into a unified vector space, preserving the identifiability of data content and strengthening the traceability of the data generation link, transforming industrial data from the original observation layer into a structured semantic object oriented towards governance tasks. Through this modeling process, multi-source heterogeneous data in the intelligent robot industry is organized into a representation with unified semantic constraints, providing a stable data foundation for subsequent data ownership identification and circulation matching mechanisms.

C. DATA OWNERSHIP IDENTIFICATION MODEL

The core of data ownership identification in the intelligent robot industry lies in transforming the subject relationships, contribution relationships, and usage boundaries in the data generation chain into a computable structured representation. Based on the unified representation r_i obtained in Section 2.2, this paper maps each data point to a set of nodes in the ownership graph, constructing a subject-data-behavior ternary relationship. Let the ownership relationship of data sample d_i be represented as $G_i = (V_i, E_i)$, where V_i represents the set of subject nodes and data nodes, and E_i represents the edges of generation, processing, use, and benefit relationships between subjects. The goal of ownership identification is to learn the mapping function $g(\cdot)$, such that the ownership label vector y_i corresponding to the data representation r_i satisfies:

$$\hat{y}_i = g(r_i; \psi) \tag{6}$$

In Formula (6), ψ is the parameter of the ownership identification model; $\hat{y}_i$ is the predicted ownership result; y_i describes the ownership status of the data source subject, processing subject, user subject, and beneficiary subject. To logically distinguish between a "User" and a "Beneficiary" when their interaction data overlaps in the unified embedding, the model incorporates a role-based attention weight within the ternary relationship. This mechanism explicitly separates direct operational interactions (User role) from value-extraction results (Beneficiary role) by mapping downstream utility metrics back to the subject nodes, ensuring that downstream entities without generative participation are accurately categorized based on their distinct relationship edges in the graph G_i.

To enhance the relationship inference capability, this paper further introduces the adjacency weight matrix A_i to describe the interaction strength between subjects [24], and uses the graph representation propagation mechanism to update the node state:

$$H_i^{(l+1)} = \sigma\left(D_i^{-1/2} A_i D_i^{-1/2} H_i^{(l)} W^{(l)}\right) \tag{7}$$

In Formula (7), $H_i^{(l)}$ represents the node representation of the l-th layer; D_i is the degree matrix; $W^{(l)}$ is the parameter matrix of the l-th layer; $\sigma(\cdot)$ is the nonlinear activation function. This update process propagates the local relationship of the node layer by layer to the global structure, so that the ownership determination no longer depends on static rules, but is driven by the association strength in the data generation link. In the ownership determination stage, this paper decomposes the data rights into learnable multi-label outputs to form a joint identification of the contribution degree of different subjects. Let the predicted probability of the k-th class of rights label be:

$$p_{ik} = \frac{\exp\left(w_k^\top z_i + c_k\right)}{\sum_{j=1}^{K} \exp\left(w_j^\top z_i + c_j\right)} \tag{8}$$

In Formula (8), w_k and c_k are the weight vector and bias term of the k-th class of labels, respectively, and K is the total number of rights labels. The training objective adopts cross-entropy loss:

$$L_{\text{own}} = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log p_{ik} \tag{9}$$

In Formula (9), N is the number of samples, and y_{ik} represents the true label. This novel objective function integrates the determination of ownership boundaries, the identification of contributors, and the learning of multiple attributions related to contributions into a single optimization framework. Therefore, the final output from the proposed model can identify both the sources of the information as well as all other parties who have been involved in processing it or using it. The identification of the ownership relationship also generates a structured mapping R_i of rights that become the input for subsequent retrieval augmentation and constraint generation, thereby producing a continuous loop of ownership identification - semantic enhancement - circulation matching. Figure 2 illustrates the data ownership relationship model.

This data ownership relationship model illustrates the hierarchy of interrelated entities that create data: the data

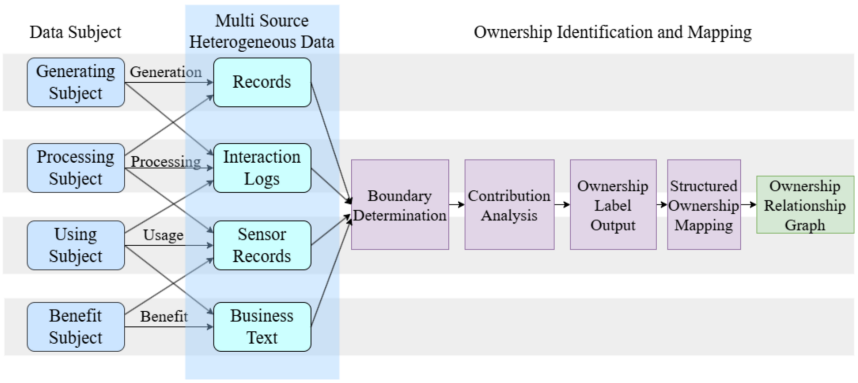

FIGURE 2. Data ownership relationship modeling

generator, the data processor, and the data user. Data nodes are used to connect various entities based on their behavioral characteristics, which creates an ownership identification path from the data source back to the rights map.

To further clarify the core elements and semantic meanings of each element in the ownership relationship modeling, this paper formally defines the relevant structures, as shown in Table 1.

TABLE 1. Ownership relationship structure definition

Element type	Semantic definition	Relationship type	Modeling methods
Data node	Represents a single data instance	Generated	Semantic vector representation
Generating the main body	Data source entity	Generate relations	Graph structure connections
Processing main body	The entity that processes the data	Processing relationship	Weighted edge representation
User	The main users of the data	Usage relationships	Semantic association
Beneficiary	Data value distribution entity	Benefit relationship	Ownership mapping
Ownership	Data ownership set	Many-to-many relationship	Constraint vector representation

Table 1 standardizes the subject types, relationship structures, and representation methods in data ownership modeling, providing a structural foundation for subsequent ownership identification and constraint gener ation.

D. RETRIEVAL-AUGMENTED SEMANTIC COMPLETION MECHANISM FOR INDUSTRY DATA

After completing the unified data representation and ownership relationship identification, incomplete information and inconsistent semantic granularity still exist between the supply-side data description and the demand-side semantic expression, resulting in insufficient alignment between the two in the unified space. This paper constructs a retrieval-augmented semantic completion mechanism. Distinct from general-purpose RAG frameworks, our approach introduces a 'constraint-aware knowledge index' that retrieves not only semantic definitions but also the legal and industrial logic associated with data rights. By injecting subject-contribution attributes directly into the augmentation loop, the model differentiates itself from standard retrieval systems by ensuring that the supplemented context is functionally relevant to ownership boundaries rather than just linguistic similarity. Let the initial representation of the i-th data be x_i, and the corresponding demand representation be q_j. A domain knowledge index $K = \{k_1, k_2, \ldots, k_l, \ldots, k_L\}$ is constructed, where k_l represents the l-th knowledge vector, and L is the number of knowledge entries. The query representation and knowledge vector are matched using a similarity function to obtain the retrieval weight:

$$\beta_{il} = \frac{\exp\left(x_i^\top k_l/\tau\right)}{\sum_{s=1}^{L}\exp\left(x_i^\top k_s/\tau\right)} \tag{10}$$

In Formula (10), τ is a temperature parameter used to adjust the smoothness of the similarity distribution, and β_{il} represents the contribution weight of the l-th knowledge to data d_i. Based on this weight, the knowledge base is weighted and aggregated to form a semantic completion vector:

$$\tilde{x}_i = \sum_{l=1}^{L}\beta_{il}k_l \tag{11}$$

This vector is used to supplement the missing contextual semantic information in the original data representation. On this basis, the original representation and the completion representation are fused to construct an enhanced semantic representation:

$$x_i^* = \gamma x_i + (1-\gamma)\tilde{x}_i \tag{12}$$

In Formula (12), the parameter $\gamma \in [0,1]$ is used to balance the contribution ratio of the original data semantics and external knowledge information, and x_i^* is the semantically enhanced data representation. To ensure the alignment effect of the supply side and the demand side in the same semantic space, a retrieval augmentation process is introduced into the demand representation q_j to obtain the enhanced representation q_j^*, and the matching degree between the two is measured by the alignment function:

$$s_{ij} = \frac{(x_i^*)^\top q_j^*}{\|x_i^*\|\|q_j^*\|} \tag{13}$$

In Formula (13), s_{ij} represents the similarity between data d_i and demand q_j in the enhanced semantic space. This measurement result not only reflects the surface feature matching, but also integrates domain knowledge and contextual semantic information, enabling the supply and demand sides to form a consistent expression at the implicit semantic level.

This mechanism dynamically completes the original data representation by introducing external knowledge index and semantic retrieval process, expanding the data semantics from a local description to a global expres sion with contextual relevance. The semantic enhancement results directly participate in matching inference in the subsequent constraint generation stage, enabling ownership information and semantic information to work together in a unified representation space, thereby improving the matching accuracy and expression completeness of data elements in the circulation process.

E. RAG-CGM COLLABORATIVE INFERENCE MECHANISM

Based on the semantically enhanced representations x_i^* and q_j^*, the rights confirmation result needs to be embedded in the generation process in the form of structured constraints, so that the circulation inference is completed under the ownership boundary and compliance rules. To this end, this paper constructs a collaborative inference model for constraint generation, transforming the ownership identification output R_i into a constraint vector c_i, and jointly modeling it with the semantic representation to form the condition generation input. Let the joint representation of the supply side and the demand side be:

$$h_{ij} = \phi(x_i^*, q_j^*, c_i) \tag{14}$$

In Formula (14), h_{ij} is a constrained joint semantic representation, which simultaneously encodes data content, demand semantics, and ownership boundary information. Based on this representation, a conditional generation model is constructed to probabilistically model the supply-demand matching relationship:

$$P(y_{ij} \mid x_i^*, q_j^*, c_i) = \frac{\exp(w^\top h_{ij} + b)}{\sum_k \exp(w^\top h_{ik} + b)} \tag{15}$$

In Formula (15), y_{ij} represents the matching label of data d_i and demand q_j, and w and b are model parameters. This distribution characterises a level of trust regarding generating matched pairs of relationships that are constrained in those relationships. The constraint vector c_i provides border control during the matching process and restricts the creation of matching paths that do not meet the condition of ownership within the space of probabilities to provide a direct constraint of ownership results of circulation decisions. A constraint consistency function $C(y_{ij}, c_i)$ is introduced to help enhance how much influence constraint information has on generation results. The constraint consistency function measures how consistent the matching results are to the ownership rules, and it shall be included as part of the optimization objective:

$$L = -\sum_{i,j} y_{ij} \log P(y_{ij} \mid x_i^*, q_j^*, c_i) + \lambda \sum_{i,j} C(y_{ij}, c_i) \tag{16}$$

Formula (16) defines λ as the trade-off coefficient. The trade-off coefficient acts as the control mechanism to create a balance between two types of matching: a semantic match and a consistent constraint match. The function $C(\cdot)$ is used to help penalize the matches that violate either the ownership boundary or usage rights so that the generative model is able to learn more often the circulation paths that comply with the ownership constraints during the learning process. This change in the generative model moves the generation process from a single semantic match to a dual-inference process that consists of a semantic match and ownership match.

In the inference stage, the candidate match set is generated based on conditional probabilities, and also by combining constraint consistency for screening and ranking purposes so that interpretable match results can be provided. Furthermore, the output shows the alignment of supply and demand, as well as the path of constraint actions taken to create a continuous inference link between the identification of ownership and the decision-making process regarding the circulation of that ownership.

F. GENERATION SCORE-DRIVEN CIRCULATION MATCHING AND RANKING OPTIMIZATION

Following the identification of possible pairings through the constraint generation model, there is a need to perform a global evaluation of the supply-demand relationship in order to determine how to generate the most efficient route for making decisions for circulating data. Therefore, this paper proposes construction of a mechanism for optimization for ranking and matching based on generation score with consideration to semantic similarity, ownership and consistency of generation, and confidence of generation for data that has a comprehensive matched score with respect to demand, represented by d_i and q_j.

$$S_{ij} = \alpha s_{ij} + \beta P(y_{ij} \mid x_i^*, q_j^*, c_i) + \gamma C(y_{ij}, c_i) \tag{17}$$

In Formula (17), $P(y_{ij} \mid x_i^*, q_j^*, c_i)$ is the output of the constraint generation model and refers to the probability that a given match is generated with respect to the ownership conditions, as a measure of the credibility of the match. The function $C(y_{ij}, c_i)$ is a measure of how well the results of the matching operation conform to the ownership constraints through the use of constraint consistency. The parameters α, β, and γ determine the relative weighting of the three types of information into the final, comprehensive score. The scoring function integrates the semantic information

with the ownership constraints so that the combined rating reflects both the quality of the match and whether or not it meets the respective ownership requirements. Using the score matrix $\{S_{ij}\}$, each datum point is ranked according to the demand set to generate a list of candidates that makes up the candidate matchings sequence π_i. In order to ensure global optimality, a normalized ranking function is created:

$$\hat{S}_{ij} = \frac{\exp(S_{ij})}{\sum_k \exp(S_{ik})} \tag{18}$$

$\hat{S}_{ij}$ in Formula (18) represents a normalized probability of matching used to remove the difference between sample scoring scales. The ranking procedure generates a priority ranking based on $\hat{S}_{ij}$, so that the best quality match relationships that meet the ownership constraints are located at the top of the priorities so the overall matching efficiency can be improved.

In order to remain current with changes in the circulation of data, an online update mechanism adjusts scores iteratively. The score after the t-th update is denoted as $S_{ij}^{(t)}$, and then there is:

$$S_{ij}^{(t+1)} = (1-\eta)S_{ij}^{(t)} + \eta\Delta_{ij}^{(t)} \tag{19}$$

η is the update step size parameter, and $\Delta_{ij}^{(t)}$ represents the correction term obtained based on the latest interactive feedback and constraint changes.

III. EXPERIMENTS

A. EXPERIMENTAL DATA SOURCES AND SAMPLE COMPOSITION

To evaluate the effectiveness of this method in identifying and circulating data elements in the intelligent robot manufacturing industry, experiments are conducted using multi-source, heterogeneous industry information existing in the field. The information is based upon data related to machinery operations, operational system interactions, and business managements of intelligent robot manufacturers, which creates cross-entity, multilevel data structures. Each individual piece of information data has three pieces of information (the generating entity, the processing entity, and the user entity information) as well as time stamped and scene semantic information to allow ownership link model and semantic analysis capability.

In terms of sample composition, the dataset is organized according to the structure of "data instance— entity relationship—ownership label". Data instances are standardized by unified semantic encoding, entity relationships describe the interaction structure in the data generation and usage process, and ownership labels, in multi-label form, characterize the data ownership and are used for model training and evaluation. Data partitioning follows the principle of distribution consistency. A stratified sampling strategy is adopted, with joint constraints on the two dimensions of entity type distribution and ownership label category, dividing the original data into training, validation, and test sets according to a fixed ratio. During the segmentation process, statistical normalization is performed on the categories and ownership labels of each subject, controlling the deviation of the proportion of the corresponding category in each subset from the overall distribution to not exceed a preset threshold. To address the inherent data imbalance between prevalent sensor logs and sparse business texts, a modality-specific re-weighting strategy is implemented during the "weighted fusion" process. This ensures that the model does not become biased toward high-frequency sensor modalities; instead, the query vector q in the attention mechanism is regularized to amplify the semantic influence of rare but critical data modalities, maintaining a logical balance across heterogeneous sources during training and validation. This ensures the consistency of different data subsets in terms of structural composition and label distribution, avoiding interference from sample distribution shifts on model training and evaluation results. This processed dataset contains combined instances of multiple subjects ranging across many combinations of complex semantically rich properties. This processed dataset provides evidence for evaluating generative models' performance when verifying rights and matches based on circulation.

B. COMPARISON METHODS

To evaluate the effectiveness of the proposed method in the ownership verification and coordination of a data element, an experiment is set up that includes representative models published in the last few years (from different technology tracks) to use as a baseline for comparison. The models include pre-trained generative models, retrieval-augmented generative models, and structure-constrained generative models. The comparison methods are processed using a uniform dataset and under the same training conditions to provide an objective comparison of the results of each experiment. The following list contains the four types of comparison methods selected for use in the experiment:

(1) Basic generative models include GPT-3 (Generative Pre-trained Transformer 3) as the baseline approach [25,26]. GPT-3 uses an autoregressive (AR) generation methodology to maximize the conditional probability of completing the sequence modeling process. The model has great performance in terms of language modeling and semantic representation, because it lacks the means to introduce outside knowledge enhancement mechanisms, and there are no external constraints for modeling the ownership of data resulting in insufficient accuracy matching and compliance when used in more complex data circulation situations.

(2) RAG (Retrieval-Augmented Generation) is offered as an approach to benchmark the retrieval-augmented generative models [27,28]. By combining generative modeling with retrieval from an external knowledge base, this method provides substantial benefits to semantic completion, as well as utilizing a wealth of knowledge; however, during the generation phase of the model, there are no explicit struc-

tural constraints proposed, thus making it more challenging to ensure that the generated output adheres to the data ownership boundaries or the specified usage rules.

(3) T5 (Text-to-Text Transfer Transformer), which is used for structured generative models, is presented as a coherent generative framework for comparative purposes. This type of model converts different types of tasks into a unified text generation problem, leading to good results for generalisation across tasks and finding a word with the intended meaning across tasks. However, T5 does not include a specific method for modeling multi-source heterogeneous data, and therefore does not have a strong representation of ownership relationships or constraints on a term's circulation.

(4) In addition to showing that constraint mechanisms are important, CTRL (Conditional Transformation Language) provides an example of a conditional generation model used for decoding constraints. By means of constraint labels, CTRL uses conditional limits to a limited extent but also relies on discrete controls for most constraint implementations.

C. CORE EVALUATION METRICS DESIGN

The evaluation metrics are developed with four key dimensions in mind: the accuracy of identifying ownership; the quality of semantic matching; constraint consistency; and the effectiveness of rank ordering for circulation. Therefore, all of the experimental measurements can be expected to represent the overall capabilities of the model throughout the multi-stage collaborative process in a consistent manner. All metrics are computed from a single common test set, and all four metrics utilize evaluation criteria that are consistent across all evaluation methods to provide objective and comparable evaluation results.

As far as ownership identification is concerned, precision, recall, and F1 score are the metrics that score a model's performance against having the correct data ownership labels identified. For semantic matching and circulation matching, NDCG (Normalized Discounted Cumulative Gain) and MRR (Mean Reciprocal Rank) metrics are introduced as two methods of evaluating ranking results [29,30]. NDCG indicates the distribution position of highly relevant data within the rank order; MRR provides information about the relative position of the correct matches within the rank order (i.e., their priority), which can be used to evaluate the responsiveness of the model to circulation decisions.

A new index for evaluating the degree of constraint consistency based on quantitative and statistical analysis of the degree of matching on ownership constraints is developed as an effective way to measure how well a model can enforce rule boundaries when coordinating between confirming ownership and circulating ownership information.

From the above index, a multi-dimensional evaluation system has been developed to give a comprehensive evaluation on various methods using three levels: identification of an ownership, semantic matching, and constraint consistency. The details and functions of this new system are presented in Table 2.

TABLE 2. Definition and explanation of evaluation indicators

Indicator name	Evaluation dimensions	Function description
Precision	Ownership identification	Assess the accuracy of ownership determination results
Recall rate	Ownership identification	Measuring the coverage of ownership identification results
F1 value	Ownership identification	Comprehensive reflection of the accuracy and completeness of identification
Constraint Consistency Rate (CCR)	Ownership identification	The degree of consistency between the generated results and the ownership constraint rules is measured to assess the compliance and structural stability of the ownership confirmation results.
NDCG	Circulation sorting	Measuring the overall ranking quality of matching results
MRR	Distribution matching	Prioritizing correct matching results
Consistency rate	Constraint generation	Measuring the degree to which matching results conform to ownership rules

This research describes the ability of different models to perform based on a number of different types of performance measures. In this way, the evaluation system can provide a better picture of how effective and stable the method developed herein is for confirming who owns data and coordinating its movement.

D. EXPERIMENTAL ENVIRONMENT AND IMPLEMENTATION DETAILS

In a unified computing environment, the experiments use a multi-core CPU (Central Processing Unit) and GPU (Graphics Processing Unit) collaborative computing architecture for the stability and efficiency of both the model's training and inference processes. The model is developed using a deep learning framework and is modularly designed according to the process of "semantic modeling, ownership recognition, retrieval aug mentation, constraint generation, and ranking optimization." Each module's parameters are jointly optimized during the same training process. The batch input method and the adaptive optimization algorithm used in the updating of the model's parameters facilitate increasing the speed at which the model converges and the stability of the model's training.

As far as implementation specifics are concerned, multi-source heterogeneous data is pre-processed and a uniform mapping between source and fixed dimension space created, aggregating data into the computation for future modules as contained within the shared representation; the retrieval augmentation module uses domain knowledge indexed retrieval methods to find semantic completion information quickly; ownership relationship definitions are converted into structured vector representations for defining ownership written in constraint generation module to ensure the output from the generation process meets ownership requirements;

algorithm global rank of candidate matches achieves through overall scoring system and repetitive procedures to allow adaptive adjustment of rankings. All comparison methods are implemented indepen dently under the same environment and data partitioning conditions, and the parameter settings follow their respective original model configurations to ensure the fairness and reproducibility of experimental results.

IV. RESULTS

A. DATA OWNERSHIP IDENTIFICATION RESULTS

This section focuses on the evaluation of ownership identification capabilities during the data element own ership confirmation process. A unified measurement and comparative analysis of different generative model performance against multiple indicators of performance is performed according to the previously described constructed dataset and comparative assessment method system. Each generative model is evaluated systematically with respect to the results of ownership determination generated across all models relative to their semantic representations and all associated constraints. To determine stability of recognition across complex data relationships, ranking and consistency indexes characterize the stability of each model's recognition process and its adherence to rules that produce the recognition of ownership. Overall performance of each model is fully characterized through the index values, which are summarized in Table 3 demonstrating the ownership recognition performance of all models.

TABLE 3. Original comparison results of ownership recognition performance

Methods	Precision	Recall	F1-score	CCR (Constraint Consistency Rate)	Reasoning time (ms)	Stability (Std)
GPT-3	0.842	0.815	0.828	0.791	128	0.021
T5	0.857	0.829	0.843	0.806	135	0.018
CTRL	0.861	0.836	0.848	0.834	142	0.017
RAG	0.874	0.852	0.863	0.847	156	0.015
Proposed	0.913	0.889	0.901	0.912	148	0.012

Table 3 provides a comparison among different methods by utilizing various metrics and shows that the proposed method provides a value of 0.913 when using the Precision metric, with Recall at 0.889 and F1-score at 0.901; these metrics represent a significant improvement over the RAG baseline. To further quantify the quality of semantic representation independently of downstream tasks, we conducted a Cosine Similarity Analysis within the latent space. The RAG-CGM method achieved a semantic alignment score of 0.942, compared to 0.865 for the standard RAG and 0.812 for GPT-3. This 8.9% improvement in intrinsic semantic clustering proves that our unified embedding module captures multimodal nuances more effectively than traditional sequence-only modeling. The performance gain is primarily attributed to the retrieval augmentation mechanism. Additionally, the reduction of misjudge/miscalculation chance results from improved representation. In the case of the CCR metric, the proposed method produces an output of 0.912, while CTRL produces an output of 0.834; therefore, it is suggested that the constraints generation mechanism used for embedding ownership rules produces stronger structural constraints which ensure improved result consistency between ownership relationships that are complex. In terms of time, the proposed technique records a time of 148 ms, which is slightly above that of other generative AI systems but below that of RAG at 156 ms. The reduction in the size of the search space due to the use of constraint guidance also gives rise to a more concentrated matching path, thus minimizing redundant calculations. The stability index decreases over time from a maximum of 0.021 to a minimum of 0.012, indicating that there has been a reduction in the magnitude of fluctuations in the model after repeated training and inference processes. In this way, the system has demonstrated greater robustness. When reviewing all dimensions, the proposed method provides effective tradeoffs among ownership recognition accuracy, constraint consistency, and operational stability, validating the ability of the generative semantic enhancement mechanism of generative semantic enhancement and the constraint synergy mechanism to confirm the ownership of data elements.

B. DATA CIRCULATION MATCHING RESULTS

Performance of data element circulate process is analyzed in this section. Performance evaluation of various generative models with respect to matching quality and circulation efficiency has been made using unified ownership constraints and semantic representation as the basis of the comparison. Ranking relevance and matching priority are the primary dimensions for the experimentation, while circulation success rate is an introduction to assess the generative models' effectiveness of matching in a real-world context under numerous constraints. Furthermore, multi-index coupling analysis has been utilized to identify differences in performance of the generative models in the data circulation scenario, and therefore systematic evaluations of the comprehensive features of each methodology have been produced. The relevant results are displayed in Figure 3, which represents the distribution of data circulation matching performance.

The NDCG and MRR two-dimensional method specifications are arranged hierarchically, as depicted in Figure

3. The method presented within this paper occupies the top-right quadrant of the graph, featuring an NDCG of 0.914 and MRR of 0.879. The size of the bubbles reflects the flow-carrying capacity of these models, while the flow success rate of the method in this paper is approximately 0.892. The reason for this difference is that by reconstructing the semantic-space definition during retrieval augmentation, a better-defined basis of candidate pairs to rank is utilized, improving the respective ranking quality and increasing their ranking priority. In the example modeling of constraint information, CTRL (0.781) produces a higher-flow success rate than T5 (0.758), but its NDCG values (0.846) are less than the method presented in this paper. In this instance, although more constraints increase compliance of matching

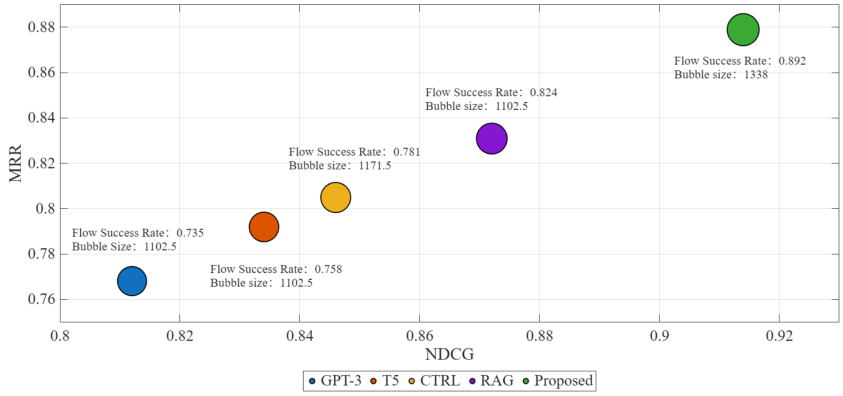

FIGURE 3. Distribution of data circulation matching performance

pairs, they still do not expand their semantic representation sufficiently, thus limiting their relevance when ranking them. The GPT-3 approach, without external knowledge use or outside agency or structure, shows remarkable performance inconsistencies, with NDCG and MRR scores of 0.812 and 0.768, respectively, as well as a success ratio of just 0.735, indicating that the degree to which metrics correlate is impaired due to the existence of matching biases resulting from incomplete semantics and absence of rules in their generated context. The proposed method's higher quality stems from the degree of consistency created in the available candidate pool due to pressures created by the constraints, thereby preventing some invalid matching paths while decreasing the matching capabilities, thereby increasing the quality of matches while increasing the overall match speed. The proposed method has created an integrated effect across data streams between the matched semantic quality and the matched ownership restriction mechanism, which provides a consistent and stable balance between matched accuracy and matched compliance in the flow of the data over the process.

C. COMPLIANCE CONSTRAINTS AND RISK CONTROL RESULTS

This part evaluates how the generative model can be used to guide collaboration for governance purposes. The analysis of this governorship is done through the examination of the governing capabilities with respect to ownership, use, and enforcement rules/standards. Using a common model template and configuration can provide the ability to compare different ways to control governorship (through variable weightings) and determine how the models react dynamically to changes in compliance and risk control. This experiment details three metrics (i.e., constraint hit potential, risk interception impact, and cost of false interception) that define the performance of governorship. The model continually samples response data under differing degrees of constraint at each sampling in order to help create a performance evolution portfolio of the generation process under rule-based conditions. Therefore, the development of this portfolio develops a multi-objective tradeoff between compliance and risk control as shown in Figure 4 with results of the distribution of compliance constraints and risk controls (i.e. Pareto).

As shown in Figure 4, every sample point has a continuous distribution within the 2D region of the false interception rate and risk interception rate and follows a stable evolutionary path along the Pareto Frontier. The risk interception rate increases from 0.742 to 0.908 with an increasing constraint weight from 0.1 to 1.5, while the false interception rate decreases from 0.182 to 0.071. This shows that adding constraint information at generation time increases the ability to suppress risky pathway types while filtering out high-risk pathways early on, thus reducing the likelihood of a false approval. The curves also illustrate clear signs of convergence behaviour within the high-weight interval, with the increase in risk interception rate reducing progressively from 0.897 to 0.908 and the decrease in false interception rate flattening out from 0.081 to 0.071. This is due to the compression effect of the generated pathway reaching its limit, as the space for the constraint has become increasingly saturated throughout the generation process, causing the performance gain to exhibit diminishing marginal returns. The Pareto front has a convex structure that is also monotonic, meaning that all three conditions of optimality apply regardless of the constraints imposed on the model. Any alterations in these parameters can not degrade performance or cause any abnormal fluctuations. The method can be thought of as a continuously adjustable synergistic mechanism for both constraints of ownership and the control of risk. As a result, there exists a consistent and stable balance between the amount of security provided and the availability of data in its flow process.

D. INTERPRETABILITY AND OWNERSHIP TRACEABILITY RESULTS

This section explores the methods that can be used to determine whether data elements are owned through the generative blockchain model. The generative model uses a single unified model output structure to expand upon and establish a comprehensive framework for the life cycle, including all transitions, movements, and processing stages,

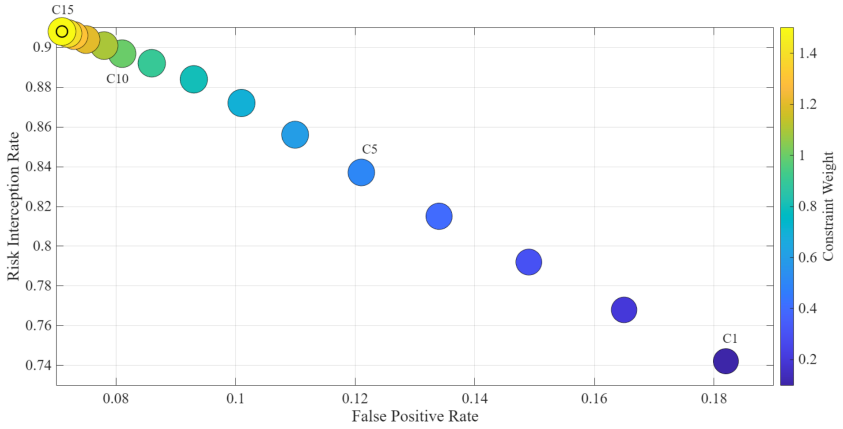

FIGURE 4. Pareto front for constraint compliance and risk control

associated with a single data instance. Through graph-based modeling of the behaviors of multiple subjects in relation to one another, the pathways over which ownership of the data is determined are established and documented by mapping ownership determination outputs to a traceable set of pathways. The weight of each edge that represents the contribution of each stage to the final f ormation o f data ownership provides a means to enable a backward analysis or logical reconstruction from result back to process and therefore to validate the transparency and consistency of the data ownership confirmation m echanism i n c omplex flow environments. The graphical interpretation of the pathway that depicts the ownership evolution of related data is shown in Figure 5.

Figure 5 shows that data, starting from the initial generation node, propagates through the generating entity, processing entity, and user entity, forming a multi-path propagation pattern with both branching and convergence, ultimately achieving normalized mapping at the ownership node. The weight of the main propagation path is 0.92, 0.87, 0.84, 0.81, and 0.89, respectively, resulting in approximately a smooth decline or reinforcing rate. The reason for this decline in the propagation path is a result of reorganizing the data and reassessing the value of data during data processing and transmitting. Because of this reorganization of the data and the reassessment of the value of the data, a disconnect through the intermediate stage where there is a structural loss in the contribution of the intermediate stages can be seen. At the time of creating the ownership mapping, the constraint rules help to reinforce the weights of the ownership of key nodes at the final c onfirmation of ownership. The branch paths that run from the processing node to the secondary user node are less weighted than the main path (0.76 and 0.82 respectively). This is because paths that don't dominate due to the constraint generation mechanism are given lower weights by being screened for path credibility during the attribution of ownership, to prevent multi-path interference and alteration of the resultant output. All of the different branch paths merge at the ownership node, showing that the model has performed weight integration and conflict resolution in a multi-source contribution model, thus providing for a global consistent basis for determining ownership, and not being based on a single branch path. The method, while outputting ownership results, forms a complete and analyzable inference chain, achieving interpretable expression and path-level traceability of the ownership confirmation process.

E. ABLATION EXPERIMENT AND MODULE CONTRIBUTION

This part of the study provides an analysis of how each core functional module contributes to the ownership confirmation and circulation coordination process. In order to conduct this analysis, multiple sets of control models are developed within a consistent experimental environment and with a consistent evaluation system; each control model has been developed by removing key modules from the original model and therefore can be used to compare the relative performance changes among the various ownership identification accuracies, consistency of constraints, and effects of circulation for each of the models that comprise this ongoing research project. While maintaining overall structural consistency, different models only remove corresponding functional units to characterize the strength and influence path of each module in the generation and inference chain, thereby revealing the source structure and internal driving mechanism of model performance improvement. The multi-indicator comparison of relevant ablation results is shown in Figure 6.

Figure 6 shows that the complete model maintains optimal performance across all metrics, with F1-score, CCR, and circulation success rate reaching 0.901, 0.912, and 0.892 respectively, while stability remains at a low level of 0.012. After removing the constraint generation module, CCR decreases from 0.912 to 0.834, a drop of 0.078, and the circulation success rate simultaneously decreases to 0.781. This indicates that ownership constraints play a crucial structural constraint role in the generation process, and their absence leads to a shift in the expression of ownership relationships, simultaneously inhibiting compliance and cir-

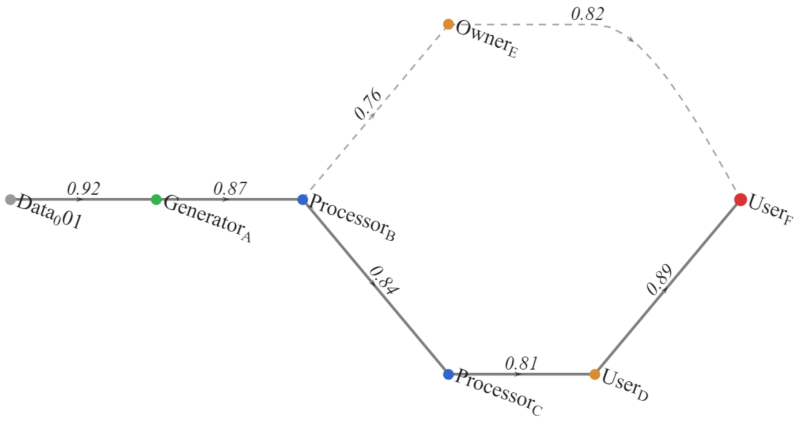

FIGURE 5. Visualization of ownership tracing path

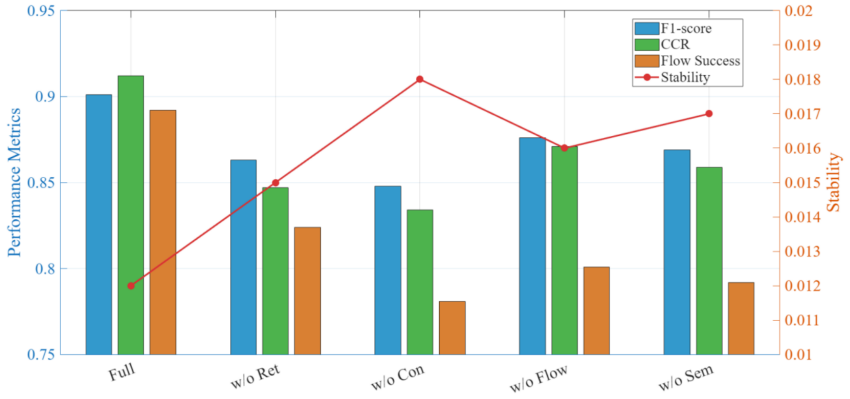

FIGURE 6. Ablation study of model components

culation matching. After removing the retrieval augmentation module, F1-score decreases from 0.901 to 0.863, while stability increases from 0.012 to 0.015. This change stems from the weakened ability of the model to characterize contextual relationships after the loss of semantic support information, leading to increased fluctuations in recognition results. When removing the circulation optimization module, the circulation success rate reduces from 0.892 to 0.801. The CCR and F1-score reductions are also smaller relative to the other methods, which makes it evident that the impact of this module is limited only to the structural optimization portion of the matching resolution task versus the actual ownership determination task. For the removal of the semantic completion module, all performance measures report equivalent reductions: F1=0.869, CCR=0.859, with an increase in stability to 0.017. The overall performance of the model shows that semantic consistency serves as an important foundational support for multi-stage inference processes. Each module has a clear division of responsibilities, performing different functions: constraint generation defines the boundaries for ownership consistency; retrieval enhancement and semantic completion support representation capabilities; and circu lation optimization adjusts matching efficiency. These modules work together to build a stable integrated system for ownership verification and circulation management.

V. CONCLUSION

This paper addresses the issues of unclear ownership and insufficient coordination in the confirmation and circulation of data elements in the intelligent robot industry. It constructs an integrated methodology that combines generative modeling, retrieval augmentation, and constraint generation mechanisms, achieving unified semantic expression of data through multi-source heterogeneous data modeling. Based on this, an ownership relationship identification model is introduced to characterize multi-subject associations, and a semantic completion mechanism enhanced by retrieval strengthens contextual consistency. Furthermore, constraint generation drives collaborative inference in ownership confirmation and circulation, ultimately achieving efficient allocation of data elements in the circulation matching and ranking optimization modules. Experimental results show that this method achieves stable improvements in ownership identification and circulation coordination. The complete model achieves an F1-score of 0.901, a significant advantage over removing key modules. The constraint consistency rate (CCR) reaches 0.912, reflecting the structural stability of ownership expression. The circulation success

rate reaches 0.892, demonstrating the effectiveness of the matching mechanism. Simultaneously, the risk interception rate reaches 0.908 with a low misjudgment level, indicating that the model achieves a balance between compliance constraints and circulation efficiency. The above results demonstrate that the method establishes a complete technical chain in complex industrial data environments, encompassing semantic modeling, ownership determination, and circulation optimization. This ensures that data elements possess clear ownership boundaries and traceable paths during circulation, possessing significant application value for promoting the standardized ownership confirmation and efficient circulation of data elements in the intelligent robot industry.

AUTHOR CONTRIBUTIONS

Conceptualization – Zou Wen, Fu Bin, Liu Shiyun and Li Manyi; methodology – Zou Wen, Fu Bin, Liu Shiyun and Li Manyi; investigation – Zou Wen, Fu Bin, Liu Shiyun and Li Manyi; writing-original draft preparation – Zou Wen, Fu Bin, Liu Shiyun and Li Manyi. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. T. Licardo, M. Domjan, and T. Orehovacki, "Intelligent robotics-A systematic review of emerging technologies and trends," Electronics, vol. 13, no. 3, pp. 542-551, 2024, DOI: 10.3390/electronics13030542.

[2] J. Y. Choi, S. Ahn, D. Kim, et al., "Exploring challenges and opportunities in manufacturing and intelligence for future robotics," International Journal of Precision Engineering and Manufacturing, vol. 26, no. 9, pp. 2203-2222, 2025, DOI: 10.1007/s12541-025-01318-2.

[3] J. Zhao, H. Liu, S. Zhang, et al., "A "five rights separation" framework for data rights confirmation in data element circulation," Humanities and Social Sciences Communications, vol. 12, no. 1, pp. 1-17, 2025, DOI: 10.1057/s41599-025-04452-4.

[4] N. Zhang, "Valuation of Data Elements: Theoretical Origins, Difficulties and Countermeasures," Journal of Jishou University(Social Sciences), vol. 45, no. 5, pp. 52-56, 2024, DOI: 10.13438/j.cnki.jdxb.2024.05.006.

[5] C. Liang, H. Du, Y. Sun, et al., "Generative AI-driven semantic communication networks: Architecture, technologies, and applications," IEEE Transactions on Cognitive Communications and Networking, vol. 11, no. 1, pp. 27-47, 2024, DOI: 10.1109/TCCN.2024.3435524.

[6] F. Hu, C. Wang, and X. Wu, "Generative artificial intelligence-enabled facility layout design paradigm," Applied Sciences, vol. 15, no. 10, pp. 5697-5705, 2025, DOI: 10.3390/app15105697.

[7] T. Xu, H. Shi, Y. Shi, et al., "From data to data asset: conceptual evolution and strategic imperatives in the digital economy era," Asia Pacific Journal of Innovation and Entrepreneurship, vol. 18, no. 1, pp. 2-20, 2024, DOI: 10.1108/APJIE-10-2023-0195.

[8] Y. Zhang, "Multi-agent evolutionary game of collaborative innovation between traditional manufacturing industry and data platform service providers in the context of digitalization," Operations Research and Fuzziology, vol. 13, pp. 4765-4772, 2023, DOI: 10.12677/ORF.2023.135478.

[9] S. Nadal, P. Jovanovic, B. Bilalli, et al., "Operationalizing and automating data governance," Journal of big data, vol. 9, no. 1, pp. 117-125, 2022, DOI: 10.1186/s40537-022-00673-5.

[10] Z. Jiang, Z. Zhu, and S. Pan, "Data governance in multimodal behavioral research," Multimodal Technologies and Interaction, vol. 8, no. 7, pp. 55-62, 2024, DOI: 10.3390/mti8070055 .

[11] M. W. Iqbal, U. Ashfaq, A. A. Soomro, et al., "TOWARDS NEXT-GENERATION AUTOMATION: DATA-DRIVEN SYNERGIES OF AI AND ROBOTICS THROUGH DATA ENGINEERING AND DATA SCIENCE," Spectrum of Engineering Sciences, vol. 3, no. 9, pp. 181-209, 2025.

[12] S. Oruma, R. Colomo-Palacios, and V. Gkioulos, "Architectural views for social robots in public spaces: business, system, and security strategies," International Journal of Information Security, vol. 24, no. 1, pp. 12-19, 2025, DOI: 10.1007/s10207-024-00924-x.

[13] P. Q. Bao, "Assessing Payment Card Industry Data Security Standards Compliance in Virtualized, Container-Based E-Commerce Platforms," Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems, vol. 12, no. 12, pp. 1-10, 2022.

[14] R. P. Pinto, B. M. C. Silva, and P. R. M. Inacio, "A system for the promotion of traceability and ownership of health data using blockchain," IEEE Access, vol. 10, no. 1, pp. 92760-92773, 2022, DOI: 10.1109/ACCESS.2022.3203193.

[15] R. Brandin and S. Abrishami, "Information traceability platforms for asset data lifecycle: blockchain-based technologies," Smart and Sustainable Built Environment, vol. 10, no. 3, pp. 364-386, 2021, DOI: 10.1108/SASBE-03-2021-0042.

[16] L. Vogt, T. Kuhn, and R. Hoehndorf, "Semantic units: organizing knowledge graphs into semantically meaningful units of representation," Journal of biomedical semantics, vol. 15, no. 1, pp. 7-15, 2024, DOI: 10.1186/s13326-024-00310-5.

[17] F. Mo, J. C. Chaplin, D. Sanderson, et al., "Semantic models and knowledge graphs as manufacturing system re-configuration enablers," Robotics and Computer-Integrated Manufacturing, vol. 86, no. 1, pp. 102625-102635, 2024, DOI: 10.1016/j.rcim.2023.102625.

[18] K. Xiao, D. Li, P. Guo, et al., "Similarity matching method of power distribution system operating data based on neural information retrieval," Global Energy Interconnection, vol. 6, no. 1, pp. 15-25, 2023, DOI: 10.1016/j.gloei.2023.02.002.

[19] S. M. R. Naqvi, M. Ghufran, C. Varnier, et al., "Enhancing semantic search using ontologies: A hybrid information retrieval approach for industrial text," Journal of Industrial Information Integration, vol. 45, no. 1, pp. 100835-100846, 2025, DOI: 10.1016/j.jii.2025.100835.

[20] M. T. Hasan and S. P. Kudapa, "Data privacy-aware machine learning and federated learning: A framework for data security," American Journal of Interdisciplinary Studies, vol. 2, no. 03, pp. 01-34, 2021, DOI: 10.63125/vj1hem03.

[21] M. M. Hasan, "Federated Learning Models for Privacy-Preserving AI In Enterprise Decision Systems," International Journal of Business and Economics Insights, vol. 5, no. 3, pp. 238-269, 2025, DOI: 10.63125/ry033286.

[22] S. Nie, J. Shi, X. Zhou, et al., "Data Component Method Based on Dual-Factor Ownership Identification with Multimodal Feature Fusion," Sensors, vol. 25, no. 21, pp. 6632-6643, 2025, DOI: 10.3390/s25216632 .

[23] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," ACM computing surveys, vol. 56, no. 9, pp. 1-36, 2024, DOI: 10.1145/3649447.

[24] R. Yang, Y. Chen, J. Yan, et al., "A Graph with Adaptive Adjacency Matrix for Relation Extraction," Computers, Materials & Continua, vol. 80, no. 3, pp. 4129-4147, 2024, DOI: 10.32604/cmc.2024.051675.

[25] S. V. Balkus and D. Yan, "Improving short text classification with augmented data using GPT-3," Natural Language Engineering, vol. 30, no. 5, pp. 943-972, 2024, DOI: 10.1017/S1351324923000438.

[26] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," Natural Language Processing Journal, vol. 6, no. 1, pp. 100048-100056, 2024, DOI: 10.1016/j.nlp.2023.100048.

[27] R. Yang, Y. Ning, E. Keppo, et al., "Retrieval-augmented generation for generative artificial intelligence in health care," Npj health systems, vol. 2, no. 1, pp. 2-9, 2025, DOI: 10.1038/s44401-024-00004-1.

[28] S. Wang, H. Yang, and W. Liu, "Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph," Scientific Reports, vol. 15, no. 1, pp. 40425-40436, 2025, DOI: 10.1038/s41598-025-21222-z.

[29] H. Brama, L. Dery, and T. Grinshpoun, "Evaluation of neural networks defenses and attacks using NDCG and reciprocal rank metrics," Interna-

tional Journal of Information Security, vol. 22, no. 2, pp. 525-540, 2023, DOI: 10.1007/s10207-022-00652-0.

[30] J. Li, H. Zeng, C. Xiao, et al., "Listwise learning to rank method combining approximate NDCG ranking indicator with Conditional Generative Adversarial Networks," Pattern Recognition Letters, vol. 179, no. 1, pp. 31-37, 2024, DOI: 10.1016/j.patrec.2024.01.015 .