Water Resources Management
https://doi.org/10.1007/s11269-025-04312-5

# A Multi-Agent and GraphRAG-Based Framework for Operation and Management Decision-Making in Hydraulic Projects

Yangrui Yang[1] · Pengfei Wang[1] · Xuemei Liu[1,3,4] · Wenyu Luo[2] · Libo Yang[1]

Received: 9 April 2025 / Accepted: 2 July 2025


**Abstract**
Inspection of hydraulic projects is a core management task to ensure the safe and stable operation of major infrastructure. The current traditional inspection model has significant shortcomings, such as high reliance on manual labor, low accuracy in identifying potential risks, and insufficient dynamic decision-making capabilities. This paper innovatively constructs a multi-agent collaborative intelligent decision-making framework that integrates large language models (LLMs) and graph retrieval-augmented generation (Graph RAG) technologies. Through a modular architecture encompassing perception, cognition, and decision, the framework achieves full-process automated inspection. The study employs multi-source inspection data from the past three years to construct a multimodal dataset. Domain-adaptive fine-tuning enhances the F1 scores of the multimodal large model in equipment recognition and defect detection by 7.2% and 6.9%, respectively. Furthermore, a dynamic knowledge graph system based on Graph RAG is established. Knowledge injection techniques compensate for gaps in domain-specific knowledge, while entity-relation reasoning mechanisms effectively mitigate model hallucination phenomena. Experimental results demonstrate that the hydraulic engineering inspection reports generated by this method, evaluated by both experts and operations personnel, accurately reflect professional knowledge and technical depth in the field of hydraulic engineering maintenance. This research provides a new technical paradigm with strong explainability and high reliability for the intelligent operation and maintenance of hydraulic engineering infrastructure, offering significant engineering application value to promote digital transformation within the industry.

**Keywords** Multimodal large language models · Multi-agent · Operation and maintenance · Graph retrieval-augmented generation

Extended author information available on the last page of the article

Published online: 17 July 2025

## 1 Introduction

The inspection of hydraulic engineering projects is a fundamental process to ensure the safe and stable operation of infrastructure, with its effectiveness playing a crucial role in the quality of lifecycle management. However, conventional inspection approaches still rely heavily on manual on-site surveys, where variations in inspectors' expertise often result in inconsistent identification of potential hazards and considerable human-induced uncertainty. Although modern monitoring technologies–such as unmanned aerial vehicles (Sibanda et al. 2021), satellites (Golestani et al. 2024), and remote sensing (Baronian et al. 2024; Faramarzzadeh et al. 2023)–have significantly improved the efficiency of data acquisition, the subsequent analysis and decision-making stages still depend largely on human interpretation. This dependency restricts the level of automation and intelligence in current inspection workflows. Moreover, existing intelligent inspection systems often exhibit limited capabilities in cross-modal data fusion and domain-specific knowledge reasoning, making them inadequate for supporting dynamic decision-making under complex operational conditions. As a result, traditional inspection methods continue to face bottlenecks in key areas such as manual dependency, risk detection accuracy, and adaptive responsiveness. To address these limitations, it is imperative to develop system-level upgrades by integrating intelligent frameworks, cross-modal data processing techniques, and automated decision-making mechanisms into hydraulic inspection processes.

The development of agent technology provides a new pathway to address the above challenges. Meng (2024) proposed a leadership model for ecological water resource governance based on evolutionary game theory and multi-agent collaboration, which reveals the interactions and equilibrium pathways among stakeholders in controlling water pollution and addressing resource scarcity. Bahrami et al. (2023) developed an agent-based system integrating Conditional Value-at-Risk optimization and the NSGA-II evolutionary algorithm to simulate the resource allocation behavior of various stakeholders under uncertainty, achieving a coordinated optimization of fairness and risk minimization. Motlaghzadeh et al. (2023) introduced a hierarchical multi-agent decision-making framework incorporating game theory to support the evaluation and optimization of water and environmental management scenarios under climate change. Aydin and Keleş (2021) further combined multi-agent systems with optimal planning and control models to develop a multi-agent dispatching system tailored for freshwater supply management in metropolitan areas. However, despite demonstrating strong collaborative decision-making capabilities in specific scenarios, these multi-agent systems generally rely on predefined rules or fixed game strategies, which limit their adaptability to complex and dynamic environments, thereby restricting their practical application in the field of hydraulic engineering.

In recent years, breakthroughs in LLMs (Chang et al. 2024) have created pivotal opportunities for advancing agent capabilities. By endowing agents with semantic understanding and dynamic reasoning abilities through LLMs, the flexibility and effectiveness of multi-agent collaboration can be significantly enhanced. LLMs are deep learning-based systems trained on massive datasets and equipped with large-scale parameter architectures, enabling reasoning and predictive performance that approaches human-level intelligence (Guo et al. 2024). This capability continues to improve as research in the field progresses (Wang et al. 2024). Consequently, LLM-based decision-making and planning agent systems are rapidly evolving (Ren et al. 2024; Wang et al. 2025). Multi-agent frameworks empowered by LLMs enable collaborative agents to operate within shared environments, facilitating the execution of complex and context-aware tasks.Altermatt et al. (2025) investigated the performance of single-agent and multi-agent language models in answering questions from EUNACOM, the Chilean national medical licensing exam. Qian et al. (2024) simulates the software development lifecycle by dividing the process into design, coding, testing, and documentation phases, assigning different prompts and roles to agents to mimic human collaboration and achieve consensus. Hong et al. (2023) integrates structured workflows into an LLM-based multi-agent programming framework to facilitate effective coordination across different agent roles. Similarly, Li et al. (2023) introduces an innovative collaborative agent framework that enables agents to autonomously complete tasks with minimal human intervention.

A review of the existing literature on multi-agent systems reveals that traditional multi-agent frameworks predominantly rely on predefined rules or shallow models to construct coordination strategies. These approaches generally lack the ability to comprehend high-level semantics and exhibit limited adaptability to dynamic and complex tasks. Meanwhile, with the continuous advancements in LLMs in general reasoning, natural language understanding, and complex task planning, LLM-based agent systems have demonstrated remarkable performance across various domains–especially in multi-agent collaboration. However, when applied to highly specialized and structurally complex tasks such as the operation and maintenance of hydraulic engineering projects, current LLM-enhanced multi-agent systems still face several critical challenges. On one hand, they exhibit insufficient deep semantic understanding of hydraulic entities and their interrelationships, and lack knowledge modeling mechanisms that integrate contextual domain logic. This often leads to suboptimal accuracy in hazard identification and risk diagnosis. On the other hand, due to inadequate incorporation of domain-specific knowledge, such systems are prone to "hallucination" effects, significantly increasing the risk of misjudgments during critical decision-making processes. Furthermore, current multi-agent frameworks generally lack mechanisms for dynamic knowledge updating and evolution, rendering them ineffective in adapting to emerging or unforeseen defects in hydraulic infrastructure maintenance.

To address the aforementioned challenges, this study proposes an intelligent decision-making framework that integrates multi-agent systems with knowledge-enhanced mechanisms to tackle key issues in hydraulic operation and maintenance, including low accuracy in hazard identification, limited dynamic decision-making capabilities, and high dependence on manual inspection. The proposed framework aims to improve the intelligence, autonomy, and knowledge-driven capabilities of hydraulic inspection tasks. The main research contributions and innovations are as follows:

(1) To improve the accuracy of hazard identification, this work fine-tunes a multimodal large language model using a domain-adaptive training set constructed from three years of multisource inspection data. This enhances the model's ability to recognize engineering risks more precisely.

(2) To address model hallucinations caused by insufficient domain knowledge and the lack of dynamic knowledge updating in changing environments, a knowledge enhancement approach is designed. By constructing scene-specific knowledge graphs based on unstructured sources such as historical inspection reports, equipment operation records, and technical manuals from the South-to-North Water Diversion Project, the framework enhances

LLM performance through domain knowledge injection. This not only mitigates hallucination effects but also improves adaptability to emerging infrastructure maintenance issues.

(3) A hierarchical multi-agent collaboration framework is proposed, grounded in the business logic of hydraulic inspection scenarios and the defined functional roles of agents. The framework comprises three layers: the perception layer, the cognition layer, and the decision layer, enabling structured coordination across heterogeneous agents.The modular decoupled design allows for flexible adaptation to various types of hydraulic projects and other infrastructure systems. Users can customize workflows and datasets according to specific scenarios, enabling rapid deployment and improving generalization in dynamic environments.

## 2 Methodology

### 2.1 Multi-Agent Collaborative Decision-making Framework

The multi-agent system framework proposed in this study is designed according to the business logic of hydraulic inspection tasks and consists of multiple specialized agents. Among these, the data collection agent, data processing agent, maintenance plan execution agent, inspection report authoring agent, and inspection report archiving agent are classified as simple reflex agents. In contrast, the defect identification agent, risk assessment agent, maintenance plan generation agent, and maintenance plan approval agent are categorized as learning agents (Ye et al. 2025; Russell and Norvig 2021). These agents collaborate through a shared memory structure and utilize message-passing mechanisms to achieve modular execution and closed-loop decision-making. The system follows a task-chain paradigm, where downstream agents depend on the outputs of upstream tasks, enabling automated workflow integration and intelligent feedback. This modular collaboration mechanism significantly enhances the automation level and flexibility of complex inspection processes. To clarify system functionality and enable hierarchical management, the agents are organized into three functional layers: perception, cognition, and decision-making. These layers cover the entire intelligent workflow from data acquisition to maintenance validation. Figure 1 illustrates the systemic implementation of this three-layer architecture and the collaboration paths among modules.

Perception Layer(Fig. 1_I): A multi-modal perception agent group is deployed at this layer to acquire real-time raw data from hydraulic engineering operations. Through an internet of things sensor network, various engineering parameters are continuously collected to monitor structural changes in real time. Simultaneously, UAV inspection images, satellite remote sensing data, and textual records are integrated to enrich the system's knowledge of operational risks. A previously developed multimodal large language model is employed to achieve spatial and temporal alignment of multi-source data. This layer consists of the data collection agent, data processing agent, and defect identification agent. The data collection agent is responsible for acquiring raw multimodal data from multiple sources, including drones, remote sensing imagery, Internet of Things sensors, and manual inputs. The data processing agent performs preprocessing on the collected data, which includes tasks such as image cleaning, video frame extraction, and text normalization. The defect identification

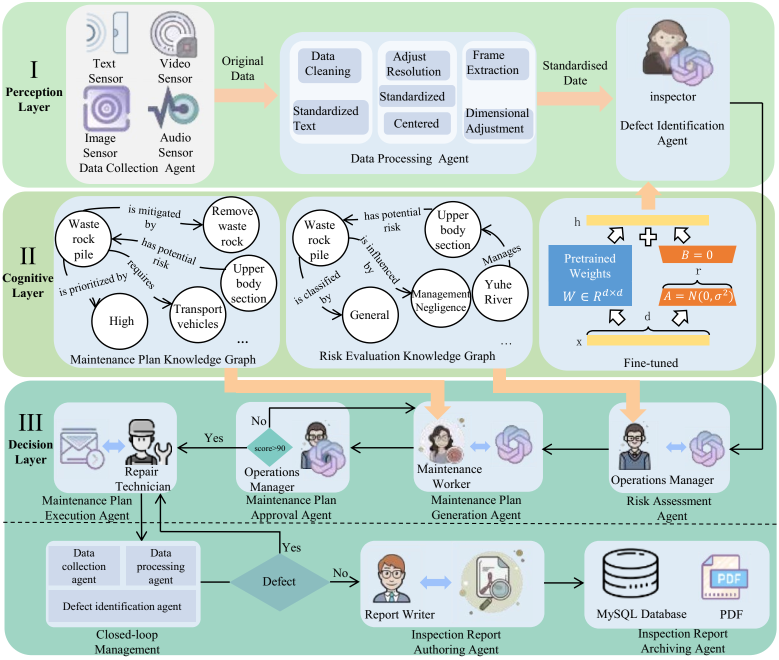


**Fig. 1** The multi-agent collaborative intelligent decision-making framework is divided into three layers: the perception layer (I), the cognition layer (II), and the decision layer (III), illustrating the closed-loop process of hydraulic engineering inspection tasks from data acquisition to intelligent decision-making

agent utilizes a domain-adapted multimodal large language model to detect defects in equipment and engineering structures.

Cognition Layer(Fig. 1_II): A dual knowledge enhancement mechanism is implemented at this layer. (1) Domain-adaptive fine-tuning strategy. By aligning multimodal representations and employing prompt fine-tuning strategies, the multimodal LLM is optimized using multimodal inspection data. This improves accuracy in key tasks such as defect identification in the operation and management of the South-to-North Water Diversion Project. (2) GraphRAG-based dynamic knowledge base. By integrating unstructured knowledge such as historical inspection reports, equipment operation records, and technical manuals from the South-to-North Water Diversion Project, knowledge graphs are constructed for different scenarios. Vector-based retrieval techniques extract highly relevant knowledge from the knowledge graph, which is then combined with role-specific prompt templates to generate precise responses. This mitigates LLM hallucinations and enhances system adaptability to new engineering maintenance defects through continuous knowledge updates.

Decision Layer(Fig. 1_III): A divide-and-conquer strategy is adopted to decompose the inspection process into a sequence of logically related, modular sub-tasks. Each module operates independently, enabling process decoupling (Zhang et al. 2023). A multi-stage task chain is designed to enable closed-loop management decisions, comprising risk assessment, maintenance plan generation, maintenance plan approval, maintenance execution, and effectiveness verification.First, The risk assessment agent interacts with the dynamic knowledge base using equipment and defect information to determine the corresponding risk levels and categories. Based on the assessment results, the maintenance plan generation agent formulates a maintenance plan pending validation by leveraging a domain-adapted fine-tuned model combined with GraphRAG dynamic knowledge base prompts. Subsequently, the maintenance plan approval agent performs automated evaluation of the plan content through an embedded scoring rubric integrated with LLM techniques. The maintenance plan execution agent then automatically sends notification emails to repair technician to initiate the implementation of the approved plan. Upon successful execution, a self-validation mechanism is triggered, utilizing a multimodal large language model to verify the maintenance effectiveness. The inspection report authoring agent employs a structured report template, as shown in Eq. 1, where the final decision output is formalized as:

$$\text{Report} = \langle E, D, C, G, M \rangle \tag{1}$$

where $E$ denotes the equipment code, $D$ describes the defect, $C \in \{\text{Structural Safety}, \text{Operational Status}, \text{Quality Defects}, \text{Management Violations}\}$ specifies the category of the issue, $G \in [\text{I}, \text{IV}]$ represents the risk level, and $M$ encompasses the maintenance plan, including resource constraints and estimated timelines.

Finally, the inspection report archiving agent automatically stores the finalized inspection report into a MySQL database.

## 2.2 Domain-Adaptive Fine-Tuning

### 2.2.1 Multimodal Fine-Tuning Strategy

To address the lack of domain knowledge in the LLaVA (Liu et al. 2023) model, this study proposes an optimized framework based on multimodal data construction and parameter-efficient fine-tuning, as illustrated in Fig. 2. First, engineering images and text are extracted from hydraulic engineering operation and maintenance documents. A Python script is used to generate multi-turn dialogue data annotated with engineering maintenance defects (Fig.2_I). The dataset is then filtered for noise using ChatGLM (GLM et al. 2024) and validated by domain experts to form a high-quality, image-text-aligned corpus.

Based on this corpus, Low-Rank Adaptation (LoRA) (Hu et al. 2022) is applied for parameter-efficient fine-tuning. To prevent catastrophic forgetting of general knowledge, the original model parameters are frozen (Fig. 2_II), and trainable matrices are injected

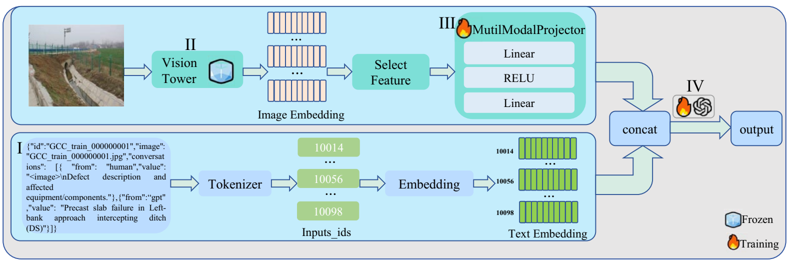


**Fig. 2** The LLaVA fine-tuning framework includes: (I) construction of image-text alignment datasets; (II) freezing of the Vision Tower; (III) injection of LoRA modules into the multimodal projector layers; and (IV) injection of LoRA modules into the attention layers of the LLM, enabling effective adaptation to multimodal tasks

```
prompt="As {Role Name},"
"Your core task is {Specific Task Content},"
"You must strictly adhere to {Constraints},"
"Refer to {Few-Shot Example Type} cases,"
"And ensure that the requirements of {Quality Standards} are met,"
"Output:"
```

**Fig. 3** Design of structured prompt template

**Table 1** Role assignments and skills

| Role | Skill |
|---|---|
| Inspector | Identify engineering maintenance defects |
| Operations Manager | Responsible for supervising and managing daily operational tasks |
| Maintenance Worker | Responsible for executing daily maintenance tasks |
| Repair Technician | Responsible for equipment repair |
| Report Writer | Responsible for drafting and submitting various reports |

only into the mutilmodalprojector layer (Fig.2_III) and the attention layer (Fig. 2_IV). A progressive learning rate scheduler and dynamic batch processing strategy are employed to efficiently integrate domain knowledge while preserving the model's general capabilities.

The LLaVA model utilizes CLIP (Radford et al. 2021) as the visual encoder $V(\cdot)$ to process engineering images with a resolution of $1024 \times 683$. The extracted visual features are then transformed via a learnable projection matrix. For the instruction text $X_q$, an embedding layer encodes the input, which is concatenated and fused with the visual features before being fed into the language model $f_\phi(\cdot)$ to generate the final response $X_a$,as shown in Eq. 2.:

$$X_a = f_\phi(\text{concat}([W(V(X_v)), E(X_q)])) \tag{2}$$

### 2.2.2 Prompt Strategy Fine-Tuning

To address the complexity of hydraulic engineering inspection tasks, this study designs a cognition-enhancing mechanism based on multi-role collaboration (Wang et al. 2023) for fine-tuning LLMs, using structured prompt templates to enable process-driven execution, as illustrated in Fig. 3.

This mechanism integrates few-shot learning (Brown et al. 2020) and decomposes the inspection workflow into modular execution stages. Each stage corresponds to specific skills assigned to distinct roles, as detailed in Table 1.

The hydraulic engineering prompt template adopts a two-tier instruction architecture:(1) Role-specific sub-instructions define operational standards (e.g., "As an inspector, please analyze the distribution characteristics of concrete cracks in the image."). (2) Dynamic examples construct knowledge anchors. The template incorporates parameterized placeholders (e.g., Role Name, Specific Task Content) to support automated adaptation. By coupling role-based expertise with process nodes, the model can collaboratively invoke multiple expert capabilities, significantly enhancing reasoning robustness in complex scenarios.

### 2.3 GraphRAG-based Dynamic Knowledge Base Construction

Inspection tasks involve multimodal data resources, including dynamic monitoring information, equipment operation records, maintenance history documents, and technical specification manuals. These data sources contain both structured data collected from real-time sensors and a large volume of unstructured textual data. Traditional data analysis methods face significant limitations in information integration and knowledge extraction, given the increasing volume and complexity of data relationships. While LLMs excel in general-domain text processing, they inherently lack domain sensitivity and dynamic updating mechanisms, which creates technical bottlenecks when generating accurate inspection decision recommendations. To address these challenges, this study proposes a GraphRAG-based domain knowledge enhancement method, which consists of two key components: operation and maintenance knowledge graph construction and hierarchical retrieval mechanism.

#### 2.3.1 Operation and Maintenance Knowledge Graph Construction

This study collected unstructured textual data from historical inspection reports, equipment operation records, and technical specification manuals related to the South-to-North Water Diversion Project. After a series of text preprocessing steps, a standardized corpus was constructed, covering 12 categories of hydraulic engineering objects, such as sluice gates and pumping stations. With the assistance of domain experts in hydraulic engineering, a conceptual model for the operation and maintenance knowledge graph was developed. This model comprises 10 types of entities (e.g., “organization”, “site”, and “component”) and 9 types of relationships (e.g., “manages”, “contains”, and “is composed of”), as defined and illustrated in Fig. 4.

Figure 5 presents the technical workflow for constructing a knowledge graph and generating community reports to support operation and maintenance in hydraulic projects. As shown in Fig. 5(I), standardized corpora obtained from processed text sources are initially segmented into discrete text blocks. Figure 5(II) illustrates the extraction of entities and semantic relations. This process adopts a two-stage parsing strategy based on LLMs, guided by an expert-defined conceptual schema and prompt engineering. An iterative optimization mechanism is introduced to regulate the coverage and consistency of extracted entities, ensuring convergence toward a predefined completeness threshold. Figure 5(III) depicts the knowledge graph construction phase. In this stage, the extracted entities and relations are structured into a graph-based representation, where nodes denote domain-specific entities and edges encode semantic or functional relationships. Graph consistency and connectivity are validated through rule-based constraints and ontology alignment to ensure semantic integrity and usability in downstream tasks. Finally, Fig. 5(IV) describes the community generation and summarization phase. Using the Leiden community detection algorithm, graph nodes are partitioned into high-cohesion clusters to facilitate modularized management. Each community is further enriched with an automatically generated summary via LLMs and prompt engineering, providing a concise representation of the global semantic landscape.

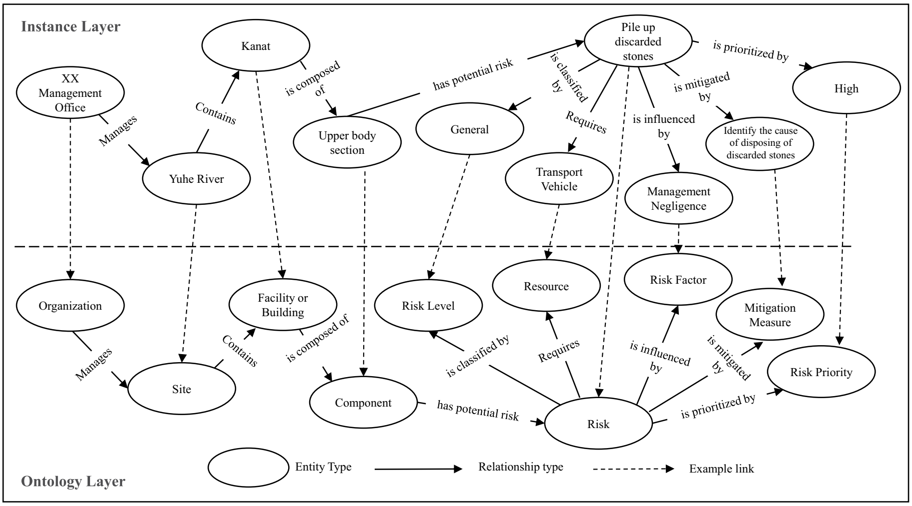


**Fig. 4** Conceptual model of the knowledge graph for operation and maintenance of hydraulic engineering projects

### 2.3.2 Hierarchical Retrieval Mechanism

The user query retrieval mechanism consists of two components: local query and global query. As shown in Fig. 6.

Local Query: The embedding model converts user input into vector representations and performs an initial match using cosine similarity within the entity vector database. Based on the match, the system locates the corresponding entity nodes in the knowledge graph and retrieves their associated semantic relationships. The entity attributes, semantic associations, and contextual information are then structured and processed by LLMs, which integrate multiple features to generate the final answer $A$. As shown in Eq. 3, the local query mechanism is formally represented as follows:

$$\mathrm{A} = f_{\mathrm{LLMs}}\left(\mathrm{concat}\left(Q, f_{\mathrm{search}}\left(\left\{e_i \mid e_i \in \mathrm{top}\text{-}k\left(\frac{E_Q \cdot E_e}{\|E_Q\|\|E_e\|}\right)\right\}, G\right)\right)\right) \tag{3}$$

where $E_Q$ is the embedding vector of the query $Q$, and $E_e$ is the embedding vector of entity $E$. $\mathrm{top}\text{-}k$ refers to selecting the top-$k$ most relevant entities. $G$ denotes the knowledge graph, $f_{\mathrm{search}}$ is the search operation, $\mathrm{concat}$ represents the concatenation of query and retrieved content, and $f_{\mathrm{LLMs}}$ is the LLM-based response generation function.

Global Query: Mapping-Reduction Architecture for Knowledge Retrieval

Mapping Stage: This stage evaluates the semantic relevance between the query $Q$ and various communities $\{C_i\}_{i=1}^n$ in the knowledge graph. As shown in Eq. 4, the specific process involves providing a question $Q$ and a set of communities $\{C_i\}_{i=1}^n$, and using prompt-based LLMs to compute the semantic relevance score $S_i$ between $Q$ and each community $C_i$:

$$S_i = f_{\mathrm{map}}(Q, C_i) \tag{4}$$

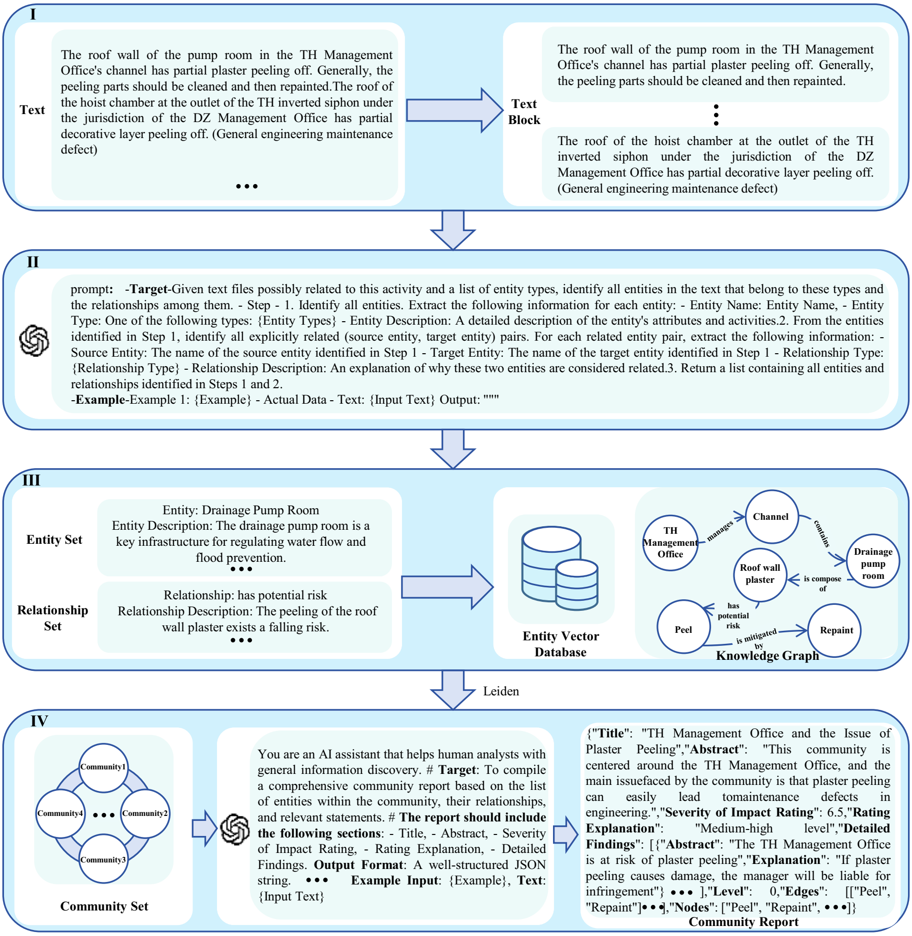


**Fig. 5** Workflow of constructing the operation and maintenance knowledge graph and generating community reports. The process includes: (I) segmenting standardized text into blocks; (II) extracting entities and relations using large language models; (III) building the knowledge graph for hydraulic engineering operation and maintenance; and (IV) detecting semantic communities and generating corresponding summaries

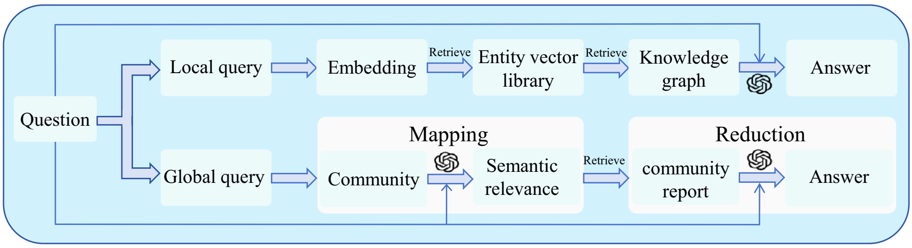


**Fig. 6** User queries, comprising both local and global queries

where $f_{\text{map}}$ is the semantic relevance scoring function powered by LLMs. $S_i$ denotes the matching degree between the query and community $C_i$.

Reduction Stage: Core entities and related knowledge are extracted and integrated from highly relevant communities. Through collaborative reasoning, The final answer is generated. As shown in Eq. 5, a relevance threshold $\theta$ is set to select the high-relevance community set $C^*$ that satisfies $S_i \geq \theta$:

$$C^* = \{C_i \mid S_i \geq \theta\} \tag{5}$$

Then, knowledge fusion and collaborative reasoning are performed to generate the final answer $A$, as shown in Eq. 6:

$$A = f_{\text{reduce}}(Q, E_i, R_i) \tag{6}$$

where $f_{\text{reduce}}$ represents the reasoning function that integrates retrieved entities and relationships using LLMs.

This Mapping-Reduction paradigm enables cross-community knowledge selection and integration at a global scale, overcoming the limitations of single-community boundaries. As a result, it facilitates the generation of more comprehensive and in-depth responses.

## 3 Experiments

### 3.1 Evaluation Metrics

To validate the performance of the fine-tuned LLaVA model and GraphRAG technology in hydraulic engineering inspection tasks, this study selects Precision (P), Recall (R), and F1-score (F1) as the key evaluation metrics. The corresponding formulas are presented in Eqs. 7–9:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{9}$$

Where True Positive (TP) denotes the number of correctly identified positive samples by the model; False Positive (FP) represents the number of samples incorrectly identified as positive; and False Negative (FN) indicates the number of positive samples that the model failed to identify.

Given the complexity of evaluating the multi-agent framework, this study compares the inspection reports generated by the framework with historical human reports to validate its applicability and technological advantages in hydraulic engineering operations.

To systematically assess the quality and effectiveness of the generated reports, this study establishes a four-dimensional evaluation system based on a five-point Likert scale (Uhm et al. 2025). Additionally, a paired *t*-test (Cheng et al. 2024) is employed to examine whether there are statistically significant differences in evaluation scores across different user groups. The four evaluation dimensions include:

- **Professionalism**: The extent to which the report reflects professional knowledge of hydraulic engineering operations and engineering standards.
- **Relevance**: The semantic consistency between the report conclusions and the problem context.
- **Usefulness**: The value of the information in supporting operational decision-making and its practical applicability.
- **Efficiency**: The end-to-end response time from task initiation to report generation.

## 3.2 Benchmark Experiment Setup

### 3.2.1 Experimental Setup for Fine-tuning the LLaVA Model

To systematically evaluate the impact of domain-adaptive fine-tuning on multimodal model performance, This study employs the LoRA fine-tuning strategy to adapt the LLaVA model to the target domain. The fine-tuning process is conducted on a GPU cluster composed of four NVIDIA RTX 4090 cards, leveraging the DeepSpeed (Rasley et al. 2020) framework to implement ZeRO-2 level distributed training. The detailed fine-tuning hyperparameters are listed in Table 2. Subsequently, performance evaluations were carried out on five models: LLaVA, LLaVA fine-tuned, GPT-4o, Gemini 1.5 Pro, and Yi-VL. Detailed information of three comparison models is provided in Table 3.

### 3.2.2 Evaluating the Enhancement Effect of GraphRAG

To systematically validate the impact of GraphRAG on improving the domain adaptability of LLMs, this study selects Qwen2-7B and DeepSeek_R1 as baseline models and designs three sets of comparative experiments:

1. **Baseline group:** No RAG technology is applied for solution generation. Experiments are conducted using Qwen2-7B and DeepSeek_R1.
2. **Traditional RAG enhancement group:** Uses traditional unstructured text data as an external knowledge base for augmentation. Implements vector embedding methods for solution generation. Experiments are conducted using Qwen2-7B_Naive and DeepSeek_R1_Naive.
3. **GraphRAG enhancement group:** Combines vector embedding with symbolic representation. Utilizes knowledge graphs to organize and structure knowledge. Provides structured contextual information for knowledge augmentation. Experiments are conducted using Qwen2-7B_Graph and DeepSeek_R1_Graph.

**Table 2** Fine-tuning hyperparameters of the LLaVA model

| Parameter | Value | Description |
|---|---|---|
| lora_r | 128 | Rank of LoRA |
| lora_alpha | 256 | Scaling factor of LoRA |
| bf16 | True | Use bfloat16 precision during training |
| deepspeed | `./scripts/zero2.json` | DeepSpeed configuration file path |
| learning_rate | 2e-4 | Initial learning rate |
| lr_scheduler_type | cosine | Learning rate scheduler |
| train_epochs | 100 | Total number of training epochs |
| per_device_train_batch_size | 128 | Training batch size per device |
| per_device_eval_batch_size | 32 | Evaluation batch size per device |

**Table 3** Multimodal models and institutions

| Model name | Development institution |
|---|---|
| GPT-4o | OpenAI |
| Gemini 1.5 Pro | Google |
| Yi-VL | 01-ai |

**Table 4** Experimental model parameter comparison table

| Model type | Temperature | Top-p | Top-k | Retrieval augmentation type |
|---|---|---|---|---|
| Qwen2-7B | 1.0 | 1.0 | 10 | None |
| Qwen2_7B_Naive | 1.0 | 1.0 | 10 | Traditional RAG |
| Qwen2-7B_Graph | 1.0 | 1.0 | 10 | GraphRAG |
| DeepSeek_R1 | 0.7 | 1.0 | 10 | None |
| DeepSeek_R1_Naive | 0.7 | 1.0 | 10 | Traditional RAG |
| DeepSeek_R1_Graph | 0.7 | 1.0 | 10 | GraphRAG |

To balance randomness and predictability, ensure the model considers the complete token probability distribution, and avoid unintended word repetition penalties, the experimental parameters are set as listed in Table 4.

### 3.2.3 Framework Evaluation

In this study, 20 hydraulic engineering inspection scenarios were selected based on the opinions of experts and maintenance personnel. These scenarios cover a variety of common hydraulic engineering issues, including equipment failures, structural damage, and emergency response handling, representing typical working conditions in hydraulic engineering inspections. These cases were used to evaluate the professionalism, relevance, efficiency, and usefulness of the generated reports. The evaluation framework was designed to cover a multi-tiered user group (with 1 to 15 years of industry experience). A dual-track verification mechanism, involving both domain experts and maintenance teams, was adopted to ensure the reliability of results: Domain experts provided structured scoring based on technical standards. Maintenance teams conducted utility validation based on actual workflows. Using Kappa consistency testing (K = 0.82), the evaluation system demonstrated both objectivity and engineering guidance value, effectively supporting the iterative optimization of the system.

# 4 Experiments Result

## 4.1 Performance Evaluation of Fine-tuned LLaVA

### 4.1.1 Data Collection and Processing

To systematically evaluate the performance improvement of domain-adaptive fine-tuning on the multimodal model, this study selected flight inspection weekly reports related to the operation and management quality issues of the South-to-North Water Diversion Project from the past three years as the data source. A total of 136 standardized inspection documents were collected. During the feature extraction stage, regular expressions and OpenCV were jointly employed to cover six categories of engineering maintenance defects, including channel engineering issues and facility equipment problems. In the data cleaning process, semantic compression was performed based on the ChatGLM-6B model to remove redundant information irrelevant to engineering maintenance defects, such as locations, timestamps, and numeric data, ensuring the data is concise and focused. For the fine-tuning data construction, the focus was placed on defects of equipment and its components. A dialogue format simulating human-machine interaction was adopted to accurately and comprehensively reflect potential risks of the equipment, generating a high-quality image-text instruction fine-tuning dataset. Finally, a panel of five experts was assembled to validate the cleaned data, resulting in 7,596 training samples, 1,889 validation samples, and 1,889 test samples.

### 4.1.2 Performance Evaluation

This study evaluates the performance of five models (LLaVA, fine-tuned LLaVA, GPT-4o, Gemini 1.5 Pro, and Yi-VL) on a test set of 1,889 samples, using precision (P), recall (R), and F1-score (F1) as evaluation metrics, as shown in Table 5.

**Table 5** Values of evaluation metrics for different models with different tasks

| Model | Equipment identification | | | Defect recognition | | |
|---|---|---|---|---|---|---|
| | P (%) | R (%) | F1 (%) | P (%) | R (%) | F1 (%) |
| LLaVA | 69.5 | 77.0 | 73.1 | 68.6 | 84.2 | 75.6 |
| GPT-4o | 78.1 | 73.3 | 75.6 | 75.8 | 78.7 | 77.2 |
| Gemini 1.5 Pro | 70.2 | 72.5 | 71.3 | 76.7 | 72.7 | 74.7 |
| Yi-VL | 70.5 | 74.7 | 72.5 | 72.3 | 75.0 | 73.6 |
| $\text{LLaVA}_{\text{fine-tuning}}$ | **81.5** | **79.2** | **80.3** | **80.6** | **84.5** | **82.5** |

**Table 6** Performance improvement of GraphRAG on various models

| Model | P | R | F1 |
|---|---|---|---|
| Qwen2_7B | 40.5% | 47.5% | 43.7% |
| Qwen2_7B-naive | 51.3% | 62.6% | 56.4% |
| Qwen2_7B-Graph | 63.2% | 67.9% | 65.5% |
| DeepSeek_R1 | 72.5% | 68.0% | 70.2% |
| DeepSeek_R1-naive | 78.9% | 78.7% | 78.8% |
| DeepSeek_R1-Graph | 81.5% | 90.8% | 85.9% |

Experimental results indicate that the fine-tuned LLaVA model achieves an F1-score of **80.3%** for equipment identification and **82.5%** for engineering maintenance defect recognition, ranking second only to GPT-4o. Compared to the non-fine-tuned LLaVA, the fine-tuned version improves the F1-score by **7.2%** for equipment identification and **6.9%** for defect recognition.

These findings demonstrate that LoRA fine-tuning significantly enhances LLaVA's adaptability to hydraulic engineering inspection scenarios. Specifically, LoRA updates the original weight matrix $W_0$ in the attention layers and multimodal mapping layers, as shown in Eq. 10:

$$W = W_0 + BA \tag{10}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are low-rank matrices with rank $r \ll \min(d, k)$. This low-rank adaptation effectively adjusts model parameters while keeping most of the original weights frozen, thereby improving the model's ability to recognize both equipment and engineering maintenance defects.

The technique not only improves model performance in domain-specific tasks but also broadens the potential applications of multimodal large language models in specialized fields such as hydraulic engineering.

## 4.2 Impact of GraphRAG on LLM Performance

To assess the impact of GraphRAG's dynamic knowledge base on LLMs' final response generation, a dataset of 985 samples was used for evaluation, with results shown in Table 6.

For open-source models in the Qwen2 series, the base model achieved a precision of 40.5%, recall of 47.5%, and an F1-score of 43.7%. Introducing NaiveRAG significantly improved performance ($\Delta P = +10.8\%$, $\Delta R = +15.1\%$, $\Delta F1 = +12.7\%$), with a breakthrough in recall (62.6%). This suggests that NaiveRAG enhances domain-specific task handling through knowledge retrieval. Integrating GraphRAG further strengthened knowledge representation, leading to synergistic improvements in precision (+22.7%), recall (+20.4%), and F1-score (+21.8%).

For commercial models such as DeepSeek_R1, which already demonstrates industry-leading performance (P = 72.5%, R = 68.0%, F1 = 70.2%), GraphRAG still provided notable gains. NaiveRAG improved professional domain retrieval precision by 6.4% (78.9%). GraphRAG significantly boosted recall to 90.8% (+22.8%), while optimizing retrieval efficiency and response quality (P = 81.5%, R = 90.8%, F1 = 85.9%).

Overall, GraphRAG effectively enhances model performance, improving recall and F1-score while maintaining high precision. These results highlight GraphRAG's broad potential for improving knowledge retrieval and generation capabilities in LLMs.

### 4.3 Framework Evaluation Results

A total of 20 maintenance personnel and 10 hydraulic engineering experts evaluated the generated inspection reports across four key dimensions:

- **Q1:** How professional do you find the generated report?
- **Q2:** How relevant is the report to your work?
- **Q3:** How helpful is the report for your tasks?
- **Q4:** Does the report generation speed improve your daily workflow efficiency?

These questions assess the professionalism, relevance, usefulness, and efficiency of the inspection reports. Specifically:

- **Professionalism** evaluates the extent of domain-specific knowledge in hydraulic engineering embedded within the report.
- **Relevance** measures how well the report aligns with real-world tasks.
- **Usefulness** gauges the report's value in operational decision-making.
- **Efficiency** assesses whether automated reporting accelerates workflow.

By constructing a systematic evaluation framework across multiple dimensions, this study verifies the framework's overall performance and provides a clear direction for iterative optimization.

**Professionalism (Q1)** As shown in Fig. 7, 40% of experts selected "Highly Agree", indicating high recognition of the report's professionalism. Additionally, 65% of maintenance personnel chose either "Agree" or "Highly Agree", demonstrating broad acceptance from field-level stakeholders. The results of the paired $t$-test, as presented in Table 7, further support this finding. The expert group yielded $t$-statistic = 2.5861, $p$-value = 0.0147 ($p < 0.05$), and the operation and maintenance personnel group showed $t$-statistic = 2.545, $p$-value = 0.0099 ($p < 0.05$), indicating statistically significant differences in favor of the framework-generated reports over traditional manually written ones.In summary, both subjective evaluations and statistical significance testing validate the superior performance of the proposed framework in terms of professionalism. The generated reports comprehensively encompass critical technical components–such as defect descriptions, risk classifications, and mainte-

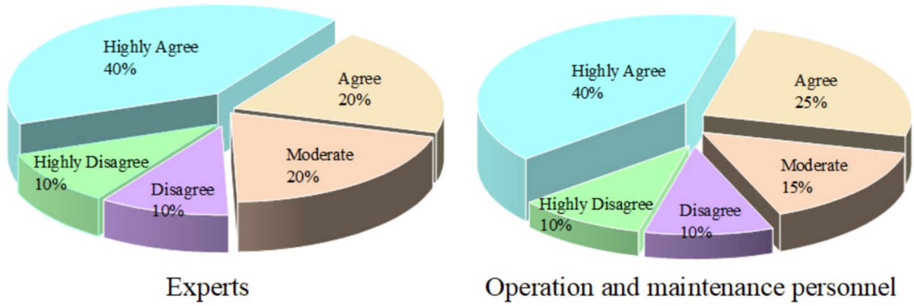


**Fig. 7** Experts and operation and maintenance personnel's evaluation results on the professionalism of the inspection report

**Table 7** Statistical analysis results of evaluation metrics by different participant groups

| Evaluation metric | Group | *T*-statistic | *P*-value |
|---|---|---|---|
| Professionalism | Experts | 2.5861 | 0.0147 |
| | Operation and maintenance personnel | 2.5450 | 0.0099 |
| Relevance | Experts | 2.3265 | 0.0225 |
| | Operation and maintenance personnel | 2.2078 | 0.0199 |
| Usefulness | Experts | 2.4495 | 0.0184 |
| | Operation and maintenance personnel | 2.3714 | 0.0142 |
| Efficiency | Experts | 4.7143 | 0.0005 |
| | Operation and maintenance personnel | 4.0978 | 0.0003 |

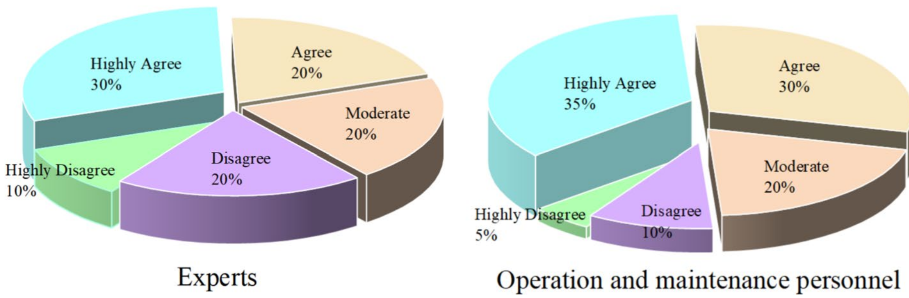


**Fig. 8** Evaluation results of inspection report relevance by experts and operation and maintenance personnel.

nance planning–thereby effectively meeting the operational and technical requirements of professional hydraulic engineering operation and maintenance.

**Relevance (Q2)** As shown in Fig. 8, 50% of experts selected "Agree" or "Highly Agree", while 65% of maintenance personnel expressed similar agreement. These results indicate a high degree of recognition from both groups, with maintenance personnel demonstrating slightly higher approval than experts.According to the paired $t$-test results presented in Table 7, The expert group showed $t\text{-statistic} = 2.3265$, $p\text{-value} = 0.0225(\ p < 0.05)$, and the operation and maintenance personnel group yielded $t\text{-statistic} = 2.2078$, $p\text{-value} = 0.0199(\ p < 0.05)$. These results confirm that the framework-generated reports exhibit a statistically significant advantage in terms of alignment between report content and task context, compared to traditional manually written reports.The observed strength in relevance is primarily attributed to the integration of a dynamic context-awareness mechanism, which enables the model to effectively capture the semantic characteristics of maintenance tasks. By adjusting the report content accordingly, the framework demonstrates enhanced scenario adaptability and contextual accuracy, particularly in sections such as case-specific recommendations and resource planning. This significantly improves the practical value of the generated reports.

**Usefulness (Q3)** As shown in Fig. 9, 70% of experts and 60% of maintenance personnel selected "Agree" or "Highly Agree," reflecting a strong consensus between both groups on the usefulness of the framework-generated reports.As presented in Table 7, the paired $t$-test results further verify this finding:The expert group showed $t\text{-statistic} = 2.4495$, $p\text{-value} = 0.0184(\ p < 0.05)$, while the operation and maintenance personnel group exhibited $t\text{-statistic} = 2.3714$, $p\text{-value} = 0.0142(\ p < 0.05)$, confirming the statistically significant advantage of the proposed framework in enhancing report utility.Specifically, the structured outputs generated by the framework provide actionable guidance for critical processes such as emergency plan formulation and routine maintenance management. The high level of agreement between expert and user evaluations highlights the practical utility of the generated reports in facilitating operational decision-making for the maintenance of hydraulic engineering.

**Efficiency (Q4)** As shown in Fig. 10, 40% of experts and 55% of maintenance personnel selected "Highly Agree," indicating a strong endorsement from both groups regarding the framework's rapid response capabilities, with experts demonstrating particularly high recognition.As indicated in Table 7, the paired $t$-test results show significant differences favoring the framework: The expert group reported a $t\text{-statistic}$ of 4.7143 ($p\text{-value} = 0.0005 < 0.05$), while the operation and maintenance personnel group had a $t\text{-statistic}$ of 4.0978 ($p\text{-value} = 0.0003 < 0.05$). These results confirm that the efficiency of reports generated by the proposed framework is significantly higher than that of manually written reports.Combining subjective evaluations with statistical significance analysis, It

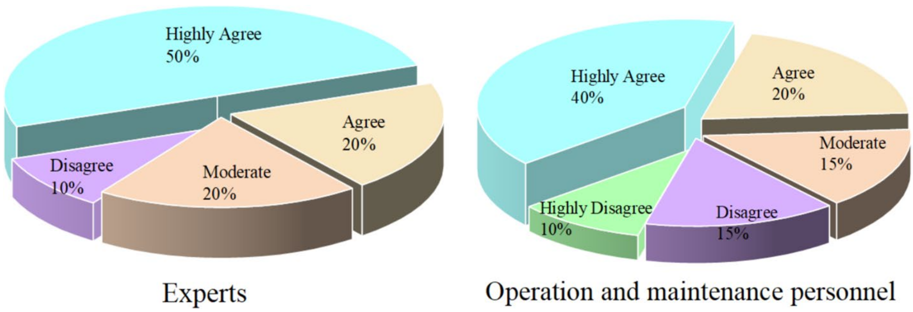


**Fig. 9** Evaluation results of experts and operation and maintenance personnel on the usefulness of inspection reports.

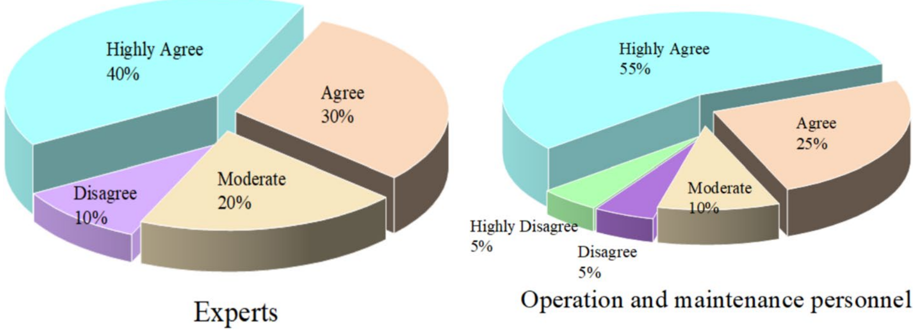


**Fig. 10** Evaluation results of experts and operation and maintenance personnel on the efficiency of the framework in generating inspection reports

can be concluded that the proposed framework significantly shortens the end-to-end process cycle while maintaining high report quality, thereby providing strong technical support for the efficient execution of operation and maintenance tasks in hydraulic engineering.

Overall, the comprehensive evaluation validates the technical advantages of the proposed framework in enhancing the intelligence level of maintenance operations in hydraulic engineering. Future research will focus on optimizing multimodal data processing mechanisms, strengthening knowledge reasoning capabilities in complex scenarios, and establishing a dynamic feedback-driven model iteration system to continuously improve the system's engineering applicability and domain universality.

## 5 Discussion

Although the dataset constructed in this study exhibits certain advantages in terms of scale and domain expertise, several limitations remain (Hao et al. 2024). First, sampling bias is present because the data predominantly originate from the South-to-North Water Diversion Project. Consequently, samples related to facility equipment faults and automated devices are somewhat overrepresented, whereas representative samples from other categories, such as cross-building issues and channel engineering, are relatively scarce. Second, the diversity of the data requires improvement. Most existing image samples were collected during daytime under clear weather conditions, with a lack of data from more complex meteorological scenarios such as nighttime or rainy and cloudy weather, which constrains the model's adaptability in extreme environments. Lastly, a minor portion of the data contains noise; some images suffer from blurriness or occlusion, reducing their clarity and overall usability. To address these challenges, future work will focus on expanding the dataset's coverage, enriching the variety and diversity of samples, and incorporating data augmentation techniques as well as multimodal generative models to synthesize high-quality synthetic samples (Qu et al. 2024; Zhao et al. 2025). These efforts aim to enhance the model's generalization capability and robustness under complex environmental conditions.

Furthermore, the proposed intelligent operation and maintenance framework demonstrates strong scalability and generalizability. Designed with a modular architecture and empowered by knowledge-enhanced reasoning mechanisms, the framework allows flexible adaptation across diverse hydraulic engineering projects and other types of infrastructure systems. Users can achieve customized deployment and rapid system migration by simply adjusting workflows and data configurations to meet specific application requirements. For instance, by replacing the domain-specific knowledge graph, fine-tuning input data, and customizing irrigation-related workflows, the framework can be seamlessly extended to support intelligent irrigation management. Looking ahead, efficient deployment and broad applicability can be realized through targeted adjustments to model parameters, knowledge graph content, and business logic. This not only facilitates the widespread adoption of intelligent operation and maintenance technologies but also promotes their implementation across various infrastructure domains. Meanwhile, this study establishes a technical paradigm for constructing transferable and adaptive general-purpose infrastructure systems, which can serve as a valuable reference for advancing research in related domains, such as sustainable development and spatial justice (Ghalehteimouri et al. 2021).

Finally, as artificial intelligence systems become increasingly integrated into the operation and maintenance of hydraulic engineering projects, their potential ethical and social impacts are drawing growing attention. Models may inherit biases from historical data during training, which can result in misjudgments during real-world decision-making. This necessitates enhanced data preprocessing procedures, including rigorous data screening, ongoing bias detection and correction, and the involvement of multi-disciplinary experts in model evaluation to improve the transparency and interpretability of system decisions. Meanwhile, the automation of operation and maintenance tasks may lead to the reduction of certain basic jobs. To mitigate employment-related impacts, it is essential to simultaneously promote skill upgrading, job transitions, and mechanisms for human–machine collaboration, thereby fostering both improved efficiency and workforce development. On the data front, areas densely covered by sensors and image acquisition devices may raise privacy concerns. This requires the strict implementation of data anonymization, access control, and encrypted storage strategies. Where necessary, privatized system deployment can further ensure data security and regulatory compliance (Timofte et al. 2024). In future deployments and research practices, continued emphasis on the integration of explainable AI (Maußner et al. 2025), fairness-aware learning mechanisms, and privacy-preserving technologies will be crucial to ensuring the safety, accountability, and sustainable development of intelligent systems in the water conservancy sector.

## 6 Conclusion and Future Work

This study addresses key challenges in intelligent water conservancy engineering operations, including knowledge fragmentation, decision-making delays, and model hallucination, by innovatively constructing a multi-agent collaborative decision-making framework enhanced with multimodal knowledge. A dual knowledge enhancement mechanism is proposed, integrating domain-adaptive fine-tuning and dynamic knowledge base augmentation. Through LoRA-based efficient fine-tuning, the performance of multimodal large models in equipment identification and engineering maintenance defect detection has been significantly improved, with F1-scores increasing by 7.2% and 6.9%, respectively. Meanwhile, the introduction of dynamic knowledge graph construction effectively reduces the model hallucination rate.Compared to traditional RAG methods, GraphRAG employs a hierarchical retrieval mechanism based on the mapping-reduction paradigm, utilizing community detection and semantic relevance calculations to enable cross-domain reasoning. This leads to a significant improvement in F1-scores for complex query scenarios, validating the effectiveness of structured knowledge representation in enhancing domain-specific tasks.The proposed modular architecture, comprising perception, cognition, and decision, enables a closed-loop management system from data collection to maintenance verification.

Future research will be carried out in the following three directions:

1. **Edge intelligence:** Develop lightweight edge computing modules to achieve real-time collaboration among drones, inspection robots, and central systems.
2. **Knowledge evolution:** Build an incremental knowledge evolution mechanism to dynamically update the knowledge graph through online learning.
3. **Federated learning:** Explore cross-basin knowledge sharing under the federated learning paradigm to overcome data silo constraints and enhance model generalization.

The engineering application of this framework will promote the transformation of hydraulic engineering operation and maintenance from experience-driven to knowledge-driven. Furthermore, the methodology provides a transferable technical framework applicable to the intelligent operation and maintenance of other infrastructure sectors, such as transportation and energy.

**Author Contributions** Yangrui Yang: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization, Funding acquisition. Pengfei Wang: Writing – review & editing, Visualization, Formal analysis. Xuemei Liu: Resources, Investigation. Wenyu Luo: Resources, Investigation. Libo Yang: Writing – review & editing.

**Funding** This work was funded by the National Natural Science Foundation of China (No. 72271091), Henan Provincial Science and Technology Research Project (No. 220901008), and the Key Scientific Research Project of Higher Education Institutions of Henan Province (No. 25A520006; No. 25A520016).

**Data Availability** Data will be made available on request.

## Declarations

**Conflict of Interest** The authors declare that they have no Conflict of interest.

**Ethical Approval** Not applicable.

**Consent to Participate** Not applicable.

**Consent to Publish** Not applicable.

## References

Altermatt FR, Neyem A, Sumonte N et al (2025) Performance of single-agent and multi-agent language models in spanish language medical competency exams. BMC Med. Educ. 25(1):1–11. https://doi.org/10.1186/s12909-025-07250-3

Aydin ME, Keleş R (2021) A multi agent-based approach for energy efficient water resource management. Comput. Ind. Eng. 151:106679. https://doi.org/10.1016/j.cie.2020.106679
Bahrami N, Nikoo MR, Al-Rawas G et al (2023) Optimal treated wastewater allocation among stakeholders based on an agent-based approach. Water Resour. Manage 37(1):135–156
Baronian I, Borna R, Jafarpour Ghalehteimouri K et al (2024) Unveiling the thermal impact of land cover transformations in Khuzestan province through modis satellite remote sensing products. Paddy Water Environ, 22(4):503–520. https://doi.org/10.1007/s10333-024-00981-x
Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 1877–1901
Chang Y, Wang X, Wang J et al (2024) A survey on evaluation of large language models. ACM Trans Intell Syst Technol 15(3):1–45
Cheng Y, Chen J, Huang Q et al (2024) Prompt sapper: a llm-empowered production tool for building ai chains. ACM Trans Softw Eng Method 33(5):1–24. https://doi.org/10.1145/3638247
Faramarzzadeh M, Ehsani MR, Akbari M et al (2023) Application of machine learning and remote sensing for gap-filling daily precipitation data of a sparsely gauged basin in east Africa. Environment Process 10(1). https://doi.org/10.1007/s40710-023-00625-y
Ghalehteimouri KJ, Shamaei A, Ros FBC (2021) Effectiveness of spatial justice in sustainable development and classification of sustainability in Tehran province. Regional Stat 11(2)
GLM T, Zeng A, Xu B, et al. (2024) Chatglm: a family of large language models from glm-130b to glm-4 all tools. arXiv:2406.12793
Golestani Z, Borna R, Khaliji MA et al (2024) Impact of urban expansion on the formation of urban heat islands in Isfahan, Iran: a satellite base analysis (1990–2019). J Geovisual Spat Anal 8(2). https://doi.org/10.1007/s41651-024-00189-x
Guo T, Chen X, Wang Y, et al (2024) Large language model based multi-agents: A survey of progress and challenges. arXiv:2402.01680
Hao Z, Cai X, Ge Y et al (2024) Resolving data gaps in global surface water monthly records through a self-supervised deep learning strategy. J. Hydrol. 640:131673. https://doi.org/10.1016/j.jhydrol.2024.131673
Hong S, Zheng X, Chen J, et al (2023) Metagpt: Meta programming for multi-agent collaborative framework. arXiv:2308.00352
Hu EJ, Shen Y, Wallis P, et al (2022) Lora: low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR). p 3
Li G, Hammoud H, Itani H et al (2023) Camel: communicative agents for" mind" exploration of large language model society. Adv. Neural. Inf. Process. Syst. 36:51991–52008
Liu H, Li C, Wu Q et al (2023) Visual instruction tuning. Adv. Neural. Inf. Process. Syst. 36:34892–34916
Maußner C, Oberascher M, Autengruber A et al (2025) Explainable artificial intelligence for reliable water demand forecasting to increase trust in predictions. Water Res. 268:122779. https://doi.org/10.1016/j.watres.2024.122779
Meng M (2024) Construction of ecological water resources management leadership model based on multi-agent collaboration. Water Resour. Manage 38:1–14. https://doi.org/10.1007/s11269-024-03939-0
Motlaghzadeh K, Eyni A, Behboudian M et al (2023) A multi-agent decision-making framework for evaluating water and environmental resources management scenarios under climate change. Sci. Total Environ. 864:161060. https://doi.org/10.1016/j.scitotenv.2022.161060
Qian C, Liu W, Liu H, et al (2024) Chatdev: communicative agents for software development. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp 15174–15186. https://doi.org/10.18653/v1/2024.acl-long.810
Qu Y, Nathaniel J, Li S, et al (2024) Deep generative data assimilation in multimodal setting. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp 449–459. https://doi.org/10.1109/cvprw63382.2024.00050
Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PmLR, pp 8748–8763
Rasley J, Rajbhandari S, Ruwase O, et al (2020) Deepspeed: system optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp 3505–3506. https://doi.org/10.1145/3394486.3406703
Ren S, Cui Z, Song R, et al (2024) Emergence of social norms in generative agent societies: principles and architecture. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI), https://doi.org/10.24963/ijcai.2024/874
Russell S, Norvig P (2021) Artificial intelligence: a modern approach, 4th, global edition Pearson. Harlow, UK

Sibanda M, Mutanga O, Chimonyo VG et al (2021) Application of drone technologies in surface water resources monitoring and assessment: a systematic review of progress, challenges, and opportunities in the global south. Drones 5(3):84

Timofte EM, Ligia Balan A, Iftime T (2024) Ai driven adaptive security mesh: cloud container protection for dynamic threat landscapes. 2024 International Conference on Development and Application Systems (DAS) pp 71–77. https://doi.org/10.1109/das61944.2024.10541148

Uhm M, Kim J, Ahn S et al (2025) Effectiveness of retrieval augmented generation-based large language models for generating construction safety information. Autom Constr 170:105926. https://doi.org/10.1016/j.autcon.2024.105926

Wang L, Ma C, Feng X et al (2024) A survey on large language model based autonomous agents. Front Comp Sci 18(6):186345. https://doi.org/10.1007/s11704-024-40231-1

Wang T, Wang P, Yang F et al (2025) Multi large language model collaboration framework for few-shot link prediction in evolutionary fault diagnosis event graphs. J Process Control 145:103342. https://doi.org/10.1016/j.jprocont.2024.103342

Wang Z, Mao S, Wu W, et al (2023) Unleashing cognitive synergy in large language models: a task-solving agent through multi-persona selfcollaboration. arXiv:2307.05300

Ye D, Wang Z, Song Y, et al (2025) Intelligent cybersecurity defense strategies based on large language model agents. 2025 2nd international conference on Algorithms, Software Engineering and Network Security (ASENS) pp 225–228. https://doi.org/10.1109/asens64990.2025.11011130

Zhang M, Press O, Merrill W, et al (2023) How language model hallucinations can snowball. arXiv:2305.13534

Zhao C, Wei H, Feyisa GL et al (2025) Evaluating spectral indices for water extraction: limitations and contextual usage recommendations. Int J Appl Earth Obs Geoinf 139:104510. https://doi.org/10.1016/j.jag.2025.104510



## Authors and Affiliations

**Yangrui Yang[1] · Pengfei Wang[1] · Xuemei Liu[1,3,4] · Wenyu Luo[2] · Libo Yang[1]**

✉ Yangrui Yang
yangyangrui@ncwu.edu.cn

[1] School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China

[2] School of Electronic Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China

[3] School of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China

[4] Collaborative Innovation Centre for Efficient Utilization of Water Resources, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China