LLM-Driven Graph Chain of Thought Reasoning for Agentic Response to Indoor Fire Risk

Anonymous ACL submission

Abstract

Understanding fire risk and response planning in complex indoor environments requires reliable reasoning over incomplete perceptions and heterogeneous domain knowledge. Although large language models (LLMs) have demonstrated strong reasoning capabilities, enabling them to perform structured, interpretable, and adaptive reasoning over dynamic fire scenes remains a significant challenge. In this work, we present Insights on Graph (IOG), an LLM-driven multi-agent reasoning framework that performs adaptive graph-based chain-of-thought reasoning for early fire-risk detection and response recommendation. IOG constructs a fire-domain knowledge graph by integrating fire safety regulations and robotic emergency response protocols and orchestrates three collaborative agents to reason over the KG. Through the incremental construction of dynamic subgraphs aligned with scene observations, IOG enables traceable reasoning, context-aware decision-making, and adaptation to environmental changes. Extensive simulations and real-world experiments have demonstrated that IOG significantly improves fire-risk understanding and response planning compared to existing baselines, highlighting its effectiveness and robustness in complex, safety-critical environments. Our code is publicly available at https://anonymous.4open.science/r/IOG.

1 Introduction

Fire is a prevalent and destructive disaster that often causes severe human casualties and substantial property damage (Zhang, 2023). For instance, the 2025 residential fire in Tai Po, Hong Kong, resulted in at least 159 fatalities and economic losses exceeding HK$2.08 billion (Kong, 2025). Many fire incidents are closely related to negligence in daily management or inadequate regulatory oversight. However, existing fire risk prevention models mainly rely on manual monitoring and post-incident responses, struggling to provide early warnings of fire risks or prevent full-process dynamic supervision of safety management, thus leaving preventable hazards unaddressed. Large Language Models are highly capable of processing complex, dynamic environmental information and multi-domain knowledge (Bubeck et al., 2023; Luo et al., 2025), and demonstrate great potential for providing precise early warning and supervisory support for fire risk prevention.

Nevertheless, the reliable deployment of LLMs in high-risk, rapidly evolving fire emergencies still faces three key challenges. First, in knowledge-intensive tasks such as fire hazard risk prediction, LLMs struggle to integrate essential fire safety knowledge (Mallen et al., 2023). This includes material flammability, equipment connectivity, and risk escalation. Second, although LLMs can reason from historical and general knowledge (Wei et al., 2022; Yao et al., 2023), their fixed internal representations are ill-suited for rapidly evolving fire emergencies. New fire sources, equipment changes, and environmental shifts can quickly alter risks; however, LLMs lack real-time knowledge updating (Yu and Ji, 2024). LLM+KG interaction methods have been proposed to address this, but they often fail to quickly incorporate new entities or relationships, leading to delayed risk assessment and decisions (Jiang et al., 2023; Luo et al., 2023). Third, emergency decision-making requires practical requirements for process traceability and verifiable evidence (e.g., fire commanders must justify their actions to investigation teams) (Singh et al., 2024). However, existing LLM-based methods mostly employ end-to-end black-box inference, which neither documents the knowledge sources of risk judgments nor provides a quantitative link between expert knowledge in knowledge graphs and robotic action decision-making. This limitation undermines credibility verification and iterative optimization.

To address these issues, this study constructs a dedicated fire safety knowledge graph (KG) that integrates fire safety guidelines, historical incident data, expert experience, and robotic response knowledge. The KG establishes a unified semantic framework of entity attributes (e.g., flammability) and relations (e.g., spatial associations), bridging the knowledge gap in raw perceptual data. We then propose an LLM-driven multi-agent reasoning framework called Insights on Graph (IOG), which orchestrates Grounding, Thought, and Judgement agents through iterative coordination. Specifically, the Grounding Agent aligns real-time visual entities with the KG; the Thought Agent retrieves and constructs context-aware subgraphs, avoiding computational overhead; and the Judgement Agent binds decision outputs to specific subgraph elements, ensuring traceable reasoning. Through this adaptive graph-based chain-of-thought reasoning process, IOG continually evolves the subgraph to capture emerging risks while maintaining an explicit evidence chain. Consequently, the framework mitigates knowledge fragmentation, improves adaptability in dynamic environments, and eliminates black-box decision-making in fire scenarios.

We conducted high-fidelity fire simulations in Unity3D and real-world drills with wheeled robots. IOG consistently outperformed the baseline methods across all key metrics and demonstrated reliable operation, robust environmental adaptability, and sustained response capabilities in complex dynamic hazard scenarios.

The main contributions of this study are as follows:

- A fire-domain knowledge graph that integrates safety guidelines, historical incidents, expert insights, and robotic protocols into a unified entity-attribute-relation structure, addressing perception-knowledge gaps without complex statistical modeling. (See Section 3.2)
- A context-aware dynamic subgraph mechanism enabling real-time retrieval and online expansion under fire-specific environmental changes, overcoming decision latency and computational redundancy inherent in predefined LLM–KG interaction frameworks during abrupt scene transitions. (See Section 3.3)
- A cross-frame evolving derived Chain-of-Thought reasoning, modeling risk assessment and task planning as incremental paths over temporal subgraphs, yielding verifiable decision evidence chains for interpretability and post-incident analysis. (See Section 3.3.4)

2 Related Work

2.1 LLMs with Knowledge Graphs

Recent studies have increasingly explored the integration of LLMs with KGs to enhance the intelligence of emergency responses. Li demonstrated that multimodal and standardized KGs improve information accuracy and decision support in flood management (Li et al., 2024). E-KELL combines KG and LLM to enhance interpretability and consistency in emergency command (Chen et al., 2024a). Further advances include Althobaiti employing KG prompting for robotic safety verification, highlighting KG's potential in robotics (Althobaiti et al., 2024); Yao applied KG with retrieval-augmented generation for earthquake response (Yao et al., 2025), and Wang proposed KG-driven real-time mapping for fire evacuation (Wang et al., 2025a). However, existing studies have focused on information extraction, retrieval, single-agent decision-making, or limited safety checks in robotics. The use of structured knowledge for multi-robot autonomous coordination in real-time emergency contexts remains an open challenge.

2.2 LLM-Based Decision Making

In recent years, LLMs have shown increasing promise in emergency response and robotic decision-making. Many efforts have combined LLMs with symbolic planners or crisis management platforms to translate natural language descriptions into action plans and adapt dynamically, thereby enhancing system adaptability and planning efficiency (Congès et al., 2024; Lee et al., 2024). Researchers have also applied LLMs to process multisource, real-time data (e.g., social media posts and emergency calls), enabling more efficient resource allocation and faster responses (Otal et al., 2024). Odubola and Wang further demonstrated the effectiveness of LLMs in disaster response, highlighting their strengths in data integration and multi-agent collaboration, which enhance system robustness. These studies underscore the significant potential of LLMs in emergency decision-making (Wang et al., 2025b; Odubola et al., 2025). However, with increasing complexity and dynamism in operational scenarios, current

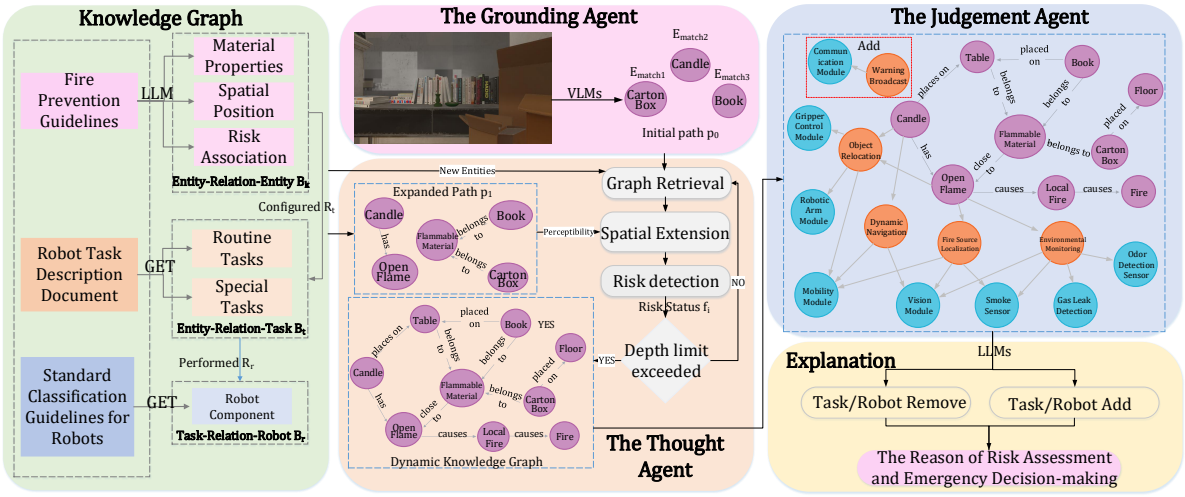


Figure 1: Overview of fire emergency response system, which consists of three main parts: Static Knowledge Graph Construction, Dynamic Knowledge Graph Generation and Robotic Response Decision Making.

approaches still lack sufficient decision flexibility and explainability capabilities. Moreover, the limited integration of perception and knowledge often leads to delayed decisions during rapid scenario evolution.

3 Method

3.1 Problem Definition

To enhance the objectivity and reliability of fire emergency responses, this study addresses the challenges of overreliance on on-site personnel's subjective experience and vigilance for risk detection, along with the lack of preventive measures due to insufficient enforcement of supervisory responsibilities. To address these challenges, we propose a multi-agent collaborative emergency response system based on LLMs. Specifically, given the current visual observation input I_t and fire knowledge graph $\mathcal{KG}$, the system aims to infer the existence of potential fire risks f_t and when risk prevention is required, generate emergency response tasks $T_a^{(t)}$ and robot assignments $R_a^{(t)}$ to ensure that each decision is traceable to explicit domain knowledge.

3.2 Construction of Static Knowledge Graph

Existing works have explored evacuation knowledge graphs (Da et al., 2023) and robotic situational awareness (Sung et al., 2025), yet they lack systematic integration of prevention knowledge with robotic execution knowledge. To address this, we follow the construction methodology of (Zhang et al., 2024) and construct a fire emergency response KG that encodes risks, task modules, and robotic components in a unified schema. The graph construction can be formalized as:

$$\mathcal{KG} = \bigcup_{k=1}^{n} \{(\mathcal{B}_k \xrightarrow{\mathcal{R}_t} \mathcal{B}_t), (\mathcal{B}_t \xrightarrow{\mathcal{R}_r} \mathcal{B}_r)\} \quad (1)$$

where $\mathcal{B}_k$, $\mathcal{B}_t$, and $\mathcal{B}_r$ denote risk-related entities, task modules, and robotic components, respectively, while $\mathcal{R}_t$ and $\mathcal{R}_r$ represent risk-to-task relations and task-to-robot relations.

The knowledge graph construction process can be expressed mathematically as follows:

1. Corpus Preparation. The corpus is composed of two sources:

(a) *Fire cases and prevention guidelines.* We collect indoor fire history case reports and prevention manuals from the fire departments and emergency management agencies of China and the United States. This corpus is denoted as $\mathcal{C}_{\text{fire}} = \{c_1, c_2, \ldots, c_{n_f}\}$.

(b) *Emergency robot standards.* We refer to international and national standards on task division and robot classification, including ISO/TC 299 (International Organization for Standardization, 2015), GA892.1-2010 (Standardization Administration of China, 2010), and ASTM E54.09 (ASTM International, 2022). This corpus is denoted as $\mathcal{C}_{\text{robot}} = \{c_1, c_2, \ldots, c_{n_r}\}$.

Thus, the complete corpus is:

$$\mathcal{C} = \mathcal{C}_{\text{fire}} \cup \mathcal{C}_{\text{robot}}. \quad (2)$$

2. Entity Extraction. Two types of entities are extracted using an LLM guided by structured

prompts, with additional prompt examples provided in Appendix A.

(a) *Risk-related entities.* From $\mathcal{C}_{\text{fire}}$, we extract triples of the form (h, r, t) via an LLM guided by designed prompts. The extraction function is denoted as:

$$\mathcal{E}_{\text{fire}} = \{(h, r, t) \mid (h, r, t) = \text{LLM}(c),\ c \in \mathcal{C}_{\text{fire}}\}. \tag{3}$$

where h, r, and t denote the head entity, the relation, and the tail entity from the fire cases and prevention guidelines corpus.

(b) *Robot tasks and components.* From $\mathcal{C}_{\text{robot}}$, we extract pairs (T, R) via an LLM guided by designed prompts. The extraction function is denoted as:

$$\mathcal{E}_{\text{robot}} = \{(T, R) \mid (T, R) = \text{LLM}(c),\ c \in \mathcal{C}_{\text{robot}}\}. \tag{4}$$

where T denotes a firefighting task module and R denotes its corresponding robotic component.

3. Ontology Alignment. The extracted entities are organized into a three-layer ontology that reflects the hierarchical structure of emergency response:

$$\mathcal{O} = (\mathcal{B}_k, \mathcal{B}_t, \mathcal{B}_r), \tag{5}$$

We define a mapping function ϕ_{align} that normalizes heterogeneous entities into this three-layer schema:

$$\phi_{\text{align}} : \mathcal{E}_{\text{fire}} \cup \mathcal{E}_{\text{robot}} \quad \mapsto \quad \mathcal{O}. \tag{6}$$

4. Relation Induction. After ontology alignment, semantic relations across layers are induced to form a coherent knowledge graph. This step leverages both structured corpus knowledge and LLM-based reasoning for relation generation:

(a) *Risk-to-task relations ($\mathcal{R}_t$).* Given a risk-related entity $b_k \in \mathcal{B}_k$, we induce its corresponding task module $b_t \in \mathcal{B}_t$ by combining domain guidelines with LLM reasoning. Specifically, an LLM is prompted to select the most relevant task from $\mathcal{B}_t$ that directly mitigates or responds to the risk. For example, when the risk entity is *"electrical short circuit"*, the LLM selects the task module *"power cutoff"* from the task set. Formally:

$$\mathcal{R}_t = \{(b_k, b_t) \mid b_t = \text{LLM}(b_k, \mathcal{B}_t),\ b_k \in \mathcal{B}_k\}, \tag{7}$$

(b) *Task-to-robot relations ($\mathcal{R}_r$).* For each task module $b_t \in \mathcal{B}_t$, we determine the robotic component $b_r \in \mathcal{B}_r$ that can execute it. Candidate task–robot pairs are initially obtained from $\mathcal{E}_{\text{robot}}$, which are then refined through LLM reasoning to resolve inconsistencies across different robot standards. For instance, if a task is labeled as *"environmental detection"* in one standard and *"environment sensing"* in another, the LLM harmonizes these minor terminological variations by unifying them under a consistent designation. Formally:

$$\mathcal{R}_r = \{(b_t, b_r) \mid b_r = \text{LLM}(b_t, \mathcal{B}_r),\ b_t \in \mathcal{B}_t\}, \tag{8}$$

3.3 Dynamic Knowledge Graph Generation

Although recent studies have explored incorporating KGs into VLMs for multimodal image understanding (Li et al., 2023; Liu et al., 2025), the fragmented integration of perception and structured knowledge constrains decision-making, reducing flexibility and explainability and causing delays under rapidly evolving scenarios. To bridge the challenge, we design a hierarchical multi-agent mechanism inspired by adaptive graph exploration. The overall process involves three specialized agents working in a collaborative reasoning loop:

3.3.1 The Grounding Agent

The Grounding Agent serves as the entry point of the reasoning pipeline by anchoring visual observations into the knowledge space. For the current scene image frame $i \in \mathcal{I}$, it employs a VLM to extract key entities E_{img} that may constitute fire hazards and their spatial locations.

$$E_{\text{img}} = \text{VLM}(I_i) \tag{9}$$

As illustrated in the indoor fire scenario image in Figure 1, the image frame contains the entity set [Green Candlestick, Book, Cardboard Box]. To ensure semantic validity, a matching function ϕ_{match} aligns E_{img} with the closest entities in the KG, yielding a refined entity set E_{matched}, which maps image entities to the most similar entities E_{matched} in the KG. In this case, the matched entities retrieved are $[\text{Candle}, \text{Book}, \text{Carton Box}]$.

3.3.2 The Thought Agent

The Thought Agent expands the perceptual anchors into structured semantic pathways. Its operation is divided into three sequential stages:

1. Knowledge Graph Retrieval. Based on E_{matched}, it first selectively explores the KG by retrieving neighborhoods and associated relations, generating candidate entities E_{cand} and KG retrieval paths p_{kg}. For example, the initial entity

path is:

$$p_0 = \{\text{Candle}, \text{Book}, \text{CartonBox}\}$$

Following the association path for p_0 in the KG, the expanded path is:

$$p_{kg} = \{\text{Candle} \xrightarrow{\text{has}} \text{Open Flame},$$
$$\text{Book}, \text{Carton Box} \xrightarrow{\text{belongs to}} \text{Flammable Material}\}$$

The expanded path reveals ignition sources and flammable materials. The new candidate entities are:

$$E_{\text{cand}} = [\text{Open Flame}, \text{Flammable Material}]$$

By constraining retrieval to entity-relevant subgraphs, the Thought Agent mitigates unnecessary graph expansion, ensuring efficiency and semantic focus.

2. Spatial Extension via VLM. After the KG-based retrieval, the agent leverages visual grounding from a VLM to incorporate spatial relationships among entities.

$$P_{spatial} = \text{VLM}(E_{\text{matched}}, E_{\text{cand}}) \tag{10}$$

For instance, it may detect that the candle is located close to the carton box:

$$P_{spatial} = \{\text{Candle} \xrightarrow{\text{Close to}} \text{Carton Box}\}$$

These spatial constraints enrich the reasoning paths, producing candidate paths that reflect both semantic relations and the actual scene layout.

Finally, the semantic and spatial information are integrated to form new expanded reasoning paths:

$$p_1 = p_{\text{kg}} \cup p_{\text{spatial}} \tag{11}$$

3. Risk detection. For each reasoning path p_i, the agent uses an LLM to determine the presence of fire risk, assigning a binary label $f_i \in \{0, 1\}$. This process continues with newly retrieved candidate entities E_{cand}, which are provided to the step of Knowledge Graph Retrieval, until the depth limit $D_{\max}$ is reached or the explored path indicates a potential hazard with $f_i = 1$, allowing for early termination of the exploration.

Finally, for each frame $I_t \in \mathcal{I}$, the reasoning chain $\mathcal{C}$ integrates all reasoning paths $\mathcal{P} = \{p_0, p_1, \ldots, p_{D_{\max}}\}$ together with their relations, which are mapped into triplets (h, r, t) to construct the dynamic subgraph:

$$DG_t = \bigcup_{\mathcal{P} \in \mathcal{C}} \left\{ (h, r, t) \mid (h \xrightarrow{r} t) \in \mathcal{P} \right\}. \tag{12}$$

3.3.3 The Judgement Agent

The Judgement Agent then focuses on response configuration, serving as the decision-making layer after risk detection has been triggered.

When $f_i = 1$, the Judgement Agent advances from detection to response planning. For the current frame i, all entities along the reasoning path $\mathcal{P}$ are extracted and matched in the fire $\mathcal{KG}$ to retrieve candidate emergency tasks $T_i^{(G)}$ and corresponding robotic components $R_i^{(G)}$. Leveraging the latent associations within the reasoning chain $\mathcal{C}$ to guide LLM-based refinement of the response configuration. Specifically, the LLM identifies redundant tasks T_{rm} and components R_{rm} to be removed, as well as missing tasks T_{add} and components R_{add} to be added. The final task modules $T_a^{(i)}$ and the final robotic components $R_a^{(i)}$ for image frame i are defined as:

$$T_a^{(i)} = \left(T_i^{(G)} \setminus T_{\text{rm}}\right) \cup T_{\text{add}} \tag{13}$$

$$R_a^{(i)} = \left(R_i^{(G)} \setminus R_{\text{rm}}\right) \cup R_{\text{add}} \tag{14}$$

As illustrated in Figure 1, we add the Warning-Broadcast task to the original candidate task set, which is not directly associated with any currently identified fire entities. This task provides real-time alerts for potential hazards.

Prompt for Explainable Reasoning Generation

Explainable Reasoning Generation:
Inputs:
- Final selected tasks: {final_task_set}
- Final selected robots: {final_robot_set}
- Reasoning chain (text): {reasoning_chain_text}
- Supporting KG triplets (list): {supporting_triplets}
- Visual evidence: {visual_evidence}
Task: For each hazard/scene element identified in the reasoning chain, produce a step-by-step justification using the format below. For each line, include a short 'Rationale' (why this task/robot addresses the hazard), 'Supporting Evidence' (KG triplet ids, shortest path length).

Format (repeat for each hazard):
‣ [Hazard or Scene Element] (entity text; bbox_id)
The reasoning identifies: [brief explanation of risk]
Suitable Tasks: [Comma-separated tasks — each with brief functional role and an evidence pointer]
Suitable Robots: [Comma-separated robots — with roles/capabilities and evidence pointer]
Rationale: [1–2 sentence causal link from hazard -> task/robot]

Figure 2: Prompt for Explainable Reasoning.

3.4 Interpretable and Adaptive Robotic Response Decision Making

LLMs exhibit exceptional generalization and reasoning capabilities (Izacard and Grave, 2020). However, they often hallucinate in scenario-specific contextual reasoning, which can undermine the reliability of robotic response decision-making in fire environments (Tonmoy et al., 2024).

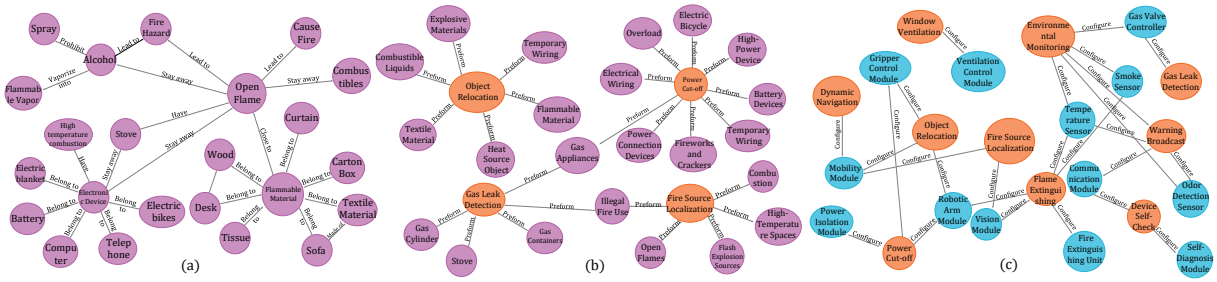


Figure 3: Illustration of the different hierarchical relationships in the static knowledge graph: (a) purple nodes denote fire scenario entities; (b) orange nodes denote task modules; (c) blue nodes denote robotic components.

To alleviate this issue, Chain-of-Thought (CoT) reasoning (Wei et al., 2022) encourages LLMs to decompose complex tasks into multiple intermediate steps, improving reasoning traceability and interpretability. Inspired by this idea, our approach formulates a fire scenario analysis as a stepwise reasoning chain that evolves with each incoming image. A dynamic scene graph explicitly models entities and their relationships, enabling continuous risk tracking and semantic enhancement for decision-making support.

Using this scene graph, the Judgement Agent generates structured justifications for each selected task–robot configuration. Once a fire risk is confirmed, hazards or scene elements from the reasoning chain are paired with candidate tasks and robots from the fire KG. A dedicated prompt guides the LLM to produce explanations in a standardized format, as shown in Figure 2. For each hazard, the output includes the Hazard or Scene Element, concise risk description, Suitable Tasks with roles and evidence, Suitable Robots with capabilities and evidence, and a short rationale linking the hazard to the task–robot choice. Grounding decisions in explicit hazards and KG evidence provides interpretable justifications usable by human operators for validation and robotic controllers for execution, ensuring adaptive and traceable responses.

4 Experimental Setup

We evaluate the proposed IOG framework in the context of fire emergency perception and robotic decision-making. Since no existing benchmark simultaneously supports fire scene understanding, emergency reasoning, and robot task planning, we construct a dedicated multimodal fire emergency dataset to facilitate systematic evaluation.

Based on this dataset, we compare IOG with representative emergency reasoning paradigms that differ in their use of structure and adaptability, including unstructured text reasoning and static knowledge graph reasoning. To ensure a fair comparison, all methods adopt the same multimodal perception model and decision making.

We further design domain-specific evaluation metrics to assess fire hazard recognition, reasoning quality, and task/robot decision effectiveness. All experiments are conducted under a unified hardware and simulation environment. Detailed descriptions of the dataset construction, baseline methods, evaluation protocol, and implementation details are provided in Appendix B.

5 Evaluation

5.1 Static Knowledge Graph Visualization

Figure 3 shows a partial example of the hierarchical structure of the fire emergency KG. The complex graph comprises approximately 650 node types and 80 relationship types, modelling approximately 1,100 pairs of interconnected entities and their semantic relations. It uniformly represents three core entities: risk factors, task execution modules, and robotic components. The graph is composed of multi-layer relationships, corresponding to semantic dependencies between environmental entities, logical matches between entities and tasks, and compatibility support between the task module and robotic components.

5.2 Overall Performance

Table 1 shows the experimental performance of the three methods on four core metrics in three scenarios: home and office environments with complex urban architectural structures under low lighting (OL) and intense lighting (OS) conditions. From the results, we can find that: 1) IOG significantly outperforms baselines across all metrics. 2) SKG performs poorly because indirect connections between directly associated nodes often introduce

Table 1: Performance comparison of STR, SKG, and IOG across different scenes. Evaluation metrics include Fire Status Accuracy (Status), Chain Completeness (Complet), and Task/Robot F1 (Task/Robot).

	Sence	Status	Complet	Task	Robot
IOG	home	92.00	70.01	87.13	89.72
	OL	88.36	74.32	84.16	85.38
	OS	85.71	79.70	88.01	87.93
	avg	**88.03**	**74.68**	**86.43**	**87.67**
SKG	home	76.00	39.57	66.97	70.61
	OL	59.09	33.66	49.70	54.67
	OS	57.14	30.08	46.60	48.44
	avg	64.08	34.44	54.42	57.91
STR	home	68.00	24.00	34.58	32.67
	OL	77.27	40.91	41.44	42.82
	OS	76.19	42.86	45.82	48.04
	avg	73.82	35.92	40.61	41.18

scene-irrelevant erroneous relationships, making it difficult to adapt to dynamic task requirements. 3) STR underperforms because it lacks the necessary domain knowledge for risk detection and problem resolution tasks. 4) Although IOG achieves the best overall performance, it still exhibits some fluctuations under extreme lighting conditions, indicating its reliance on perceived information quality and suggesting potential for further improvement.

5.3 Ablation Study

Does search depth matter for IOG? To investigate the influence of the search depth $D_{\max}$ on IOG's reasoning ability, we conduct experiments under settings with depths ranging from 2 to 4, as shown in Table 2. The results indicate that the best performance is achieved at depth 3 across all metrics. Because moderate neighborhood expansion helps IOG capture richer entity relations. However, increasing the depth to 4 leads to performance degradation due to the introduction of noise and redundant entities that hinder effective semantic association.

Table 2: Performances of IOG with different depths.

Metric/$D_{\max}$	2	3	4
F-Status	71.62	88.03	71.62
Complt	52.00	74.68	55.57
Task	66.54	86.43	66.56
Robot	67.83	87.67	68.03

How different model choices for IOG? In the main results, we adopt Qwen2-VL-7B-Instruct (Wang et al., 2024) and GPT-4o-mini (Ouyang et al., 2022) as the primary inference engines for IOG. In this section, we explore IOG with various model components, including multiple VLMs, including Qwen2-VL-7B-Instruct, InternVL2.5-8B (Chen et al., 2024b), and LLaMA3.2-11B-Vision-Instruct (Lee et al., 2025), as well as various LLMs, including GPT-4o-mini, Meta-LLaMA3.1-8B-Instruct (Vavekanand and Sam, 2024), DeepSeek-R1 (Guo et al., 2025), and InternVL2-8B (OpenGVLab, 2024). The performance of different model components in IOG's risk detection and rescue response tasks is presented in the heatmap, as shown in Figure 4. The result shows that the collaborative capability between VLMs and LLMs directly determines IOG's reasoning performance. A VLM with stronger text-image alignment and multi-entity recognition capabilities (i.e., Qwen2-VL-7B-Instruct), along with an LLM with more advanced semantic reasoning ability (i.e., GPT-4o-mini), can contribute to better reasoning performance in fire scenarios.

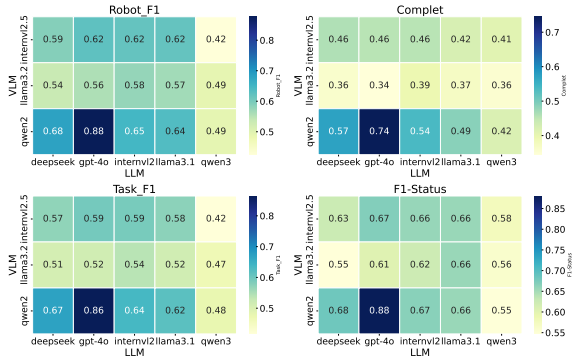
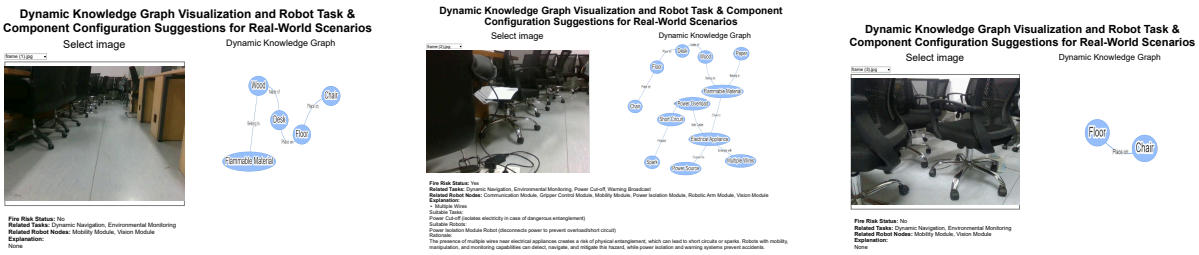


Figure 4: Performances of IOG with different model choices.

Does LLM-based refinement matter for emergency response reasoning? Table 3 explores the impact of incorporating LLMs on the emergency task and robot response reasoning capability of three methods. The results indicate that: 1) The LLM-based refinement provides STR with general knowledge for emergency robot task responses, significantly improving its overall performance. 2) LLMs' ability to identify and remove incorrect graph information further enhances SKG and IOG effectiveness. 3) The LLM-based refinement effectively enhances performance, even when the dynamic KG is built from models with limited vi-



(a) Time T_1 (b) Time T_2 (c) Time T_3

Figure 5: Knowledge graph evolution under complex disturbances (e.g., lighting changes, occlusions, and viewpoint shifts) at different time points.

sual and semantic understanding. 4) LLMs' generalizable reasoning effectively boosts traditional methods with insufficient or inaccurate emergency domain knowledge in task response.

Table 3: Performances of three decision-making reasoning with LLM-based refinement. T(-)/T(+) and R(-)/R(+) refer to Task and Robot F1 without/with LLMs.

	LLM	T(-)	T(+)	R(-)	R(+)
STR	deepseek	40.09	54.79	41.13	55.25
	gpt-4o	40.61	71.85	41.18	73.60
	internvl2	39.09	60.67	39.43	61.86
	llama3.1	44.34	61.18	46.09	62.17
SKG	deepseek	54.78	56.21	58.24	59.23
	gpt-4o	54.42	56.21	57.91	59.23
	internvl2	57.74	62.15	60.24	64.33
	llama3.1	54.77	56.59	59.39	59.74
IOG	deepseek	58.19	67.02	55.51	68.23
	gpt-4o	69.86	86.43	64.62	87.67
	internvl2	52.44	63.54	49.25	64.82
	llama3.1	56.33	61.76	45.05	63.63

5.4 Empirical Analysis

To evaluate IOG's real-world adaptability, we integrate it into an NXROBO Spark-T robot equipped with a depth camera and two active wheels for dynamic scene perception. A visual interface is developed to illustrate the evolution of the dynamic KG and the resulting configuration suggestions. Figure 5 shows three perception stages (T_1, T_2, and T_3) during robot navigation. Initially, IOG detects only furniture and retrieves their flammable attributes from the KG, leading to routine navigation and environmental monitoring. As the robot advances, multiple wires entangled with electrical appliances and nearby papers are detected, triggering dynamic KG expansion with relations such as wire–appliance entanglement, power connections, and proximity to flammable materials. The risk status is updated to TRUE, and IOG recommends actions including power cut-off and warning broadcast to mitigate potential hazards.

5.5 Explainability of IOG

IOG can present multi-level knowledge reasoning chains and provide structured explanations for response tasks, as shown in Figure 5(b). IOG outputs include three types of explicit chains: original knowledge chains, scene entity to task chains, and task to robotic component chains, providing a complete visualisation of the reasoning process from perception to response. Based on dynamic reasoning chains derived from scene images, IOG formally generates the basis for configuring execution task components and robotic components, namely "Suitable Tasks" and "Suitable Robots."

6 Conclusion

In this study, we investigated reliable reasoning in complex indoor fire environments with incomplete perceptions and heterogeneous knowledge. We constructed a fire-domain knowledge graph and proposed Insights on Graph, an LLM-driven multi-agent reasoning framework grounded in this graph. By reasoning over incrementally evolving scene-aware subgraphs, IOG enables adaptive and traceable graph-based chain-of-thought reasoning for early fire-risk detection and context-aware response recommendations. Simulations and real-world experiments demonstrated that IOG significantly outperformed existing baselines in fire-risk understanding and response planning. In future work, we will extend the framework to more complex scenarios, including larger-scale dynamic environments and real-time task prioritization.

Limitations

One limitation of the Insights on Graph framework is its reliance on LLMs, which are computationally demanding and may challenge real-time deployment on edge devices with limited resources. Although the Dynamic Subgraph mechanism helps alleviate this by focusing on relevant portions of the knowledge graph, the overall system still requires substantial processing power, particularly for real-time updates. Additionally, the system's performance depends on the accuracy of VLMs for perceptual grounding, which may introduce errors in dynamic, low-visibility environments. Future work will explore ways to balance model size, inference speed, and real-time adaptability for more efficient deployment.

References

Abdulrahman Althobaiti, Angel Ayala, JingYing Gao, Ali Almutairi, Mohammad Deghat, Imran Razzak, and Francisco Cruz. 2024. How can llms and knowledge graphs contribute to robot safety? a few-shot learning approach. *arXiv preprint arXiv:2412.11387*.

ASTM International. 2022. ASTM E54.09: Standard Test Methods for Response Robots. West Conshohocken, PA: ASTM International.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Minze Chen, Zhenxiang Tao, Weitong Tang, Tingxin Qin, Rui Yang, and Chunli Zhu. 2024a. Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction*, 113:104804.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Aurélie Congès, Audrey Fertier, Nicolas Salatgé, Sébastien Rebière, and Frederick Benaben. 2024. R-io suite: integration of llm-based ai into a knowledge management and model-driven based platform dedicated to crisis management. *Software and Systems Modeling*, pages 1–26.

Mingkang Da, Teng Zhong, and Jiaqi Huang. 2023. Knowledge graph construction to facilitate indoor fire emergency evacuation. *ISPRS International Journal of Geo-Information*, 12(10):403.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shubham Gupta, Nandini Saini, Suman Kundu, Chiranjoy Chattopadhyay, and Debasis Das. 2024. Synergizing vision and language in remote sensing: A multimodal approach for enhanced disaster classification in emergency response systems. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 3278–3281. IEEE.

International Organization for Standardization. 2015. ISO/TC 299 – Robotics. LaTeX entry for the ISO technical committee on robotics. https://www.iso.org/committee/5915511.html.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

China Daily Hong Kong. 2025. Tai po fire: What insurance compensation can be expected? https://www.chinadailyhk.com/hk/article/624619. Preliminary estimates indicate payouts tied to the disaster could reach HK$2.08 billion. Accessed 2025-12-30.

Jacqueline Lee, Michelle Cantu, Joel Korb, Eva Meth, John D Griffith, Joanna Korman, Anna Yuen, Peter Schwartz, and Abigail S Gertner. 2024. Planning ai assistant for emergency decision-making (planaid): Framing planning problems and assessing plans with large language models. In *AAAI 2025 Workshop LM4Plan*.

Jewon Lee, Ki-Ung Song, Seungmin Yang, Donguk Lim, Jaeyeon Kim, Wooksu Shin, Bo-Kyeong Kim, Yong Jae Lee, and Tae-Ho Kim. 2025. Efficient llama-3.2-vision by trimming cross-attended visual features. *arXiv preprint arXiv:2504.00557*.

Mengkun Li, Chen Yuan, Kejin Li, Minzhong Gao, Yuan Zhang, and Huiying Lv. 2024. Knowledge management model for urban flood emergency response based on multimodal knowledge graphs. *Water (20734441)*, 16(12).

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. 2023. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466.

Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Botian Shi, and Ding Wang. 2025. Aligning vision to

language: Text-free multimodal knowledge graph construction for enhanced llms reasoning. *arXiv preprint arXiv:2503.12972*.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, and 1 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

O Odubola, TS Adeyemi, OO Olajuwon, NP Iduwet, AI Aniekan, and T Odubola. 2025. Ai in social good: Llm powered interventions in crisis management and disaster response. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 3(1):3353–3360.

Team OpenGVLab. 2024. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. *URL https://internvl. github. io/blog/2024-07-02-InternVL-2.0*.

Hakan T Otal, Eric Stern, and M Abdullah Canbaz. 2024. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jaibir Singh, Suman Rani, and Garaga Srilakshmi. 2024. Towards explainable ai: interpretable models for complex decision-making. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, volume 1, pages 1–5. IEEE.

Standardization Administration of China. 2010. GA 892.1-2010: General Technical Requirements for Firefighting Robots. Beijing, China: Ministry of Public Security.

Won Sung, Yiming Li, Jiahui Zeng, and Charles C. Kemp. 2025. Triffid: Autonomous robotic aid for increasing first responders efficiency. *arXiv preprint arXiv:2405.10021*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Raja Vavekanand and Kira Sam. 2024. Llama 3.1: An in-depth analysis of the next-generation large language model. *Preprint, July*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tianxin Wang, Ping Du, Pei Dang, Tao Liu, and Pengpeng Li. 2025a. A real-time mapping method for knowledge graph-driven large language models: a focus on indoor fire evacuations. *International Journal of Digital Earth*, 18(1):2468407.

Yi Wang, Chengliang Wang, Xueqing Zhang, and Li Zeng. 2025b. Towards intelligent emergency management: A scenario–learning–decision framework enabled by large language models. *Mathematics*, 13(21):3463.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Liwei Yao, Fu Ren, Du Kaixuan, and Qingyun Du. 2025. From knowledge graph construction to retrieval-augmented generation: a framework for comprehensive earthquake emergency support. *Geo-spatial Information Science*, page 1–21.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. 2023. Language is all a graph needs. *arXiv preprint arXiv:2308.07134*.

Pengfei Yu and Heng Ji. 2024. Information association for language model updating by mitigating lm-logical discrepancy. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 117–129.

Chenting Zhang. 2023. Review of structural fire hazards, challenges, and prevention strategies. *Fire*, 6(4):137.

Jialei Zhang, Shubin Cai, Zhiwei Jiang, Jian Xiao, and Zhong Ming. 2024. Firerobbrain: Planning for a firefighting robot using knowledge graph and large language model. In *2024 10th IEEE International Conference on Intelligent Data and Security (IDS)*, pages 37–41. IEEE.

A Prompts Details

A.1 Prompt Design for Static Knowledge Graph Construction

To enable structured knowledge extraction from fire prevention and emergency robotics–related textual data, we design a set of prompt templates that explicitly specify target entity categories, task semantics, and component attributes. As illustrated in Fig. 6, these prompts guide large language models to identify and organize risk-related entities, robot tasks, and functional components into a structured static knowledge graph, which serves as a foundational representation for subsequent reasoning and planning.

Prompt for Fire Risk-Related Triplets Extraction

The following is a fire prevention-related text. Extract up to 5 relevant triples (subject, relation, object). Focus on logical and causal understanding rather than literal extraction, and optimize triples using reasoning.
1.The text may contain typos, irrelevant info, or formatting issues. Clean and correct the text to make it clear and concise.
2.The subject must be a gerund or noun, the relation must be a verb, and the object must be a noun. Avoid named people or places.
3.Infer implied causality or intent. For example, 'card ignited' may imply 'cause fire'; 'fire drill in kindergarten' may imply 'prevent fire'.
4.Only use relations from this list:
['cause', 'produce', 'belong to', 'lead to'... 'prevent']
5.Use reasoning to extract the most relevant fire prevention triples. If a triple lacks context, infer and complete it logically.
6.Output format:
1)(subject, relation, object)
2)(subject, relation, object)
...
If no relevant triples, return: "No fire-related triples"
Original Text:{Corpus}

Prompt for Robot Tasks and Components Extraction

Please extract functions or tasks related to emergency firefighting robots or fire emergency robots from the following text.
1. Output only the Task or Robot function list one per line. Do not include any additional text or explanation.
2. If there are similar or redundant tasks keep only one.
3. Output format:
1)(Task, Robot)
2)(Task, Robot)
...
Original Text:{Corpus}

Figure 6: Structured prompt used to guide the extraction of risk-related entities, robot tasks, and components using LLMs.

A.2 Prompt Design for Dynamic Knowledge Graph Generation

To support temporally adaptive knowledge graph construction from video-based scenarios, we design a collection of structured prompts tailored for vision language models. As shown in Fig. 7, the prompt design decomposes dynamic scene understanding into three complementary processes: visual entity extraction, verification of visually grounded relationships, and detection of potential hazards or abnormal events. By enforcing explicit reasoning steps and visual evidence constraints, these prompts enable reliable and interpretable dynamic knowledge graph generation in complex and evolving environments.

Prompt for Image Entity Extraction

Please observe the image and list five to seven main objects that are highly relevant for assessing fire hazards.
- Only include objects that are commonly associated with fire risk or fire safety (e.g., heat sources, flammable items, fire suppression tools).
- Do not include information about color, size, position, or quantity—only object names.
- List object names only, without additional descriptions.
- Output format: Object1, Object2, Object3

Prompt for Visual Confirmability of Relationships

Please carefully analyze the image and determine whether the following relationships between entities can be visually confirmed: "{graph}"
You must judge based only on the image content. If the relationship exists, answer "Yes"; otherwise, answer "No".
- Only reply "Yes" or "No", do not add explanations or extra text.

Prompt for Fire Risk Detection

You are to assess the overall fire hazard risk based on the following reasoning chain.
Only choose one result from ["Fire Hazard Present", "Normal", "Uncertain"] and return only that result.
Reasoning Chain: "{graph}"
Judgment Guidelines:
1. "Fire Hazard Present" if:
  - The chain includes a fire source or potential fire source (e.g., open flame, fire appliances, overloaded electricity, leakage, aging wires, vaporized alcohol);
  - AND flammables are also present;
  - AND there is a spatial or causal connection like "close to", "touching", "can ignite".
  - Having only flammables without a fire source does not imply risk.
2. "Normal" if:
  - The reasoning chain describes a common, safe scene;
  - Flammable objects may exist but are not linked to fire sources or dangerous interactions.
3. "Uncertain" if:
  - The chain contains only object attributes or spatial placement info;
  - Lacks a fire source or key trigger condition.
Important Notes:
- Just having flammables does not mean a hazard exists; a clear fire source and ignition path must be present;
- Example: "Cardboard - is - flammable" + "Books - are - flammable" ≠ fire hazard;
- A clear fire source and close or causal relation are required to assess fire risk.
Please strictly follow these rules and return only one result from ["Fire Hazard Present", "Normal", "Uncertain"]. Do not add any additional text.

Figure 7: Unified prompt design for dynamic knowledge graph generation.

B Experimental Details

B.1 Baselines

Since the limited availability of fire-specific emergency response research, we benchmark our proposed IOG framework against two representative reasoning paradigms in the general emergency domain: unstructured text reasoning (STR) and static knowledge graph reasoning (SKG). These baselines were selected to ensure a comprehensive comparison across different levels of structure and adaptability in reasoning approaches.

STR (Gupta et al., 2024): This method relies solely on scene descriptions generated by VLMs, without using any explicit emergency knowledge structures.

SKG (Ye et al., 2023): This method uses an emergency KG to perform decision-making through fixed entity-task mappings.

IOG: Our method combines static emergency knowledge and real-time scene understanding to construct an adaptive KG for interpretable decision-making.

For all baselines, we use Qwen2-VL-7B-Instruct for multimodal perception and scene understanding, and GPT-4o-mini for risk detection and task decision-making.

B.2 Dataset

Due to the absence of publicly available datasets that effectively integrate fire emergency scene information with robot task response configurations, we construct a dedicated multimodal fire emergency dataset to support the evaluation of the IOG framework. The dataset is designed to combine textual knowledge, visual scene data, and robot-task reasoning annotations, thereby bridging the gap between emergency scenario representation and robotic response capability.

Text Data: Include 4,000 official documents, each comprising approximately 800 words. Covering fire safety guidelines and prevention education, fire incident case studies, and industry classification standards for fire emergency response robots.

Simulated data: Include 1,000 high-resolution frames (1920×1080), collected by a virtual robot operating within Unity3D-generated environments, and data augmentation techniques are applied to simulate real-world environmental noise, as shown in Figure 8(a). These images represent typical indoor scenarios, including both home and office settings.

Real-world data: Include ten video segments, each five minutes long, captured by a wheeled robot operating in typical indoor environments.The frame extraction process samples at 8 frames per second, resulting in approximately 300 representative images.

Reasoning Evaluation: Constructed from the simulated images. Each sample includes an image and four labels: fire hazard status (true/false), task execution module, chain reasoning of scene, and robot capability component. They are generated from LLMs-produced analyses of real incidents and are manually verified using expert judgment.

(a) (b)

Figure 8: Simulation environments and robotic platform.

B.3 Evaluation Metrics

In the absence of a standardized evaluation system for our dataset, we designed a comprehensive evaluation framework tailored to the fire emergency domain. The framework consists of three dimensions that jointly capture the critical aspects of reasoning quality and practical usability. **Fire Status Accuracy** evaluates the correctness of fire hazard predictions, ensuring that the model can reliably identify hazardous conditions. **Chain Reasoning Completeness** measures the ability of the reasoning process to reconstruct semantic relations from predicted reasoning chains, reflecting the coherence and depth of the model's inferential capacity. **Task/Robot F1** assesses the effectiveness of task planning and robot decision-making in emergency response scenarios, emphasizing the practical operability of the generated outputs.

B.4 Experimental Environment

All experiments are conducted on NVIDIA GeForce RTX 4090 GPU (24GB VRAM) with Python 3.11 and PyTorch 2.0. The simulation is built in Unity3D 2021.3 LTS with high-fidelity lighting and physics, and the mobile robot platform is NXROBO Spark-T equipped with a 1280×960@7fps depth camera, as shown in Figure 8(b).