2025년 추계학술발표대회 : 일반부문

KOSHA 규정 연계 비계 안전점검 자동화를 위한 Multimodal LLM 기반 다중 에이전트 시스템

A Multimodal LLM-based Multi-Agent System for Automated Scaffolding Safety Inspection Linked with KOSHA Regulation

○김 도 현* 김 진 욱** 이 상 훈***
Dohyun Kim Jinwook Kim Sanghoon Lee

Abstract

Construction is known as one of the most hazardous industries globally, reporting high rates of fatalities and injuries. To prevent accidents effectively at construction sites, proactive safety inspections are required to eliminate and correct the safety hazards. However, conventional safety inspections heavily depend on the experience and competency of human inspectors, along with labor-intensive manual documentation. This study proposes a multimodal large language model (MLLM)-based multi-agent system that automatically detects scaffolding-related safety hazards and generates inspection reports in compliance with KOSHA regulations. The evaluation results demonstrated that the proposed system can effectively support inspectors by automatically providing the critical safety issues with regulation-driven guidelines.

키워드 : 건설안전, 도메인 특화 인공지능, 멀티모달 대형언어모델, 비계, 다중 에이전트 시스템, 검색증강생성
Keywords : Construction Safety, Domain-specific AI, Multimodal Large Language Model, Scaffolding, Multi-Agent System, RAG

1. 서론

1.1 연구의 배경 및 필요성

건설 프로젝트에서 안전은 프로젝트 성공의 핵심 요인 중 하나이다(Wanberg, et al., 2013). 하지만, 건설업은 국제적으로 고위험 산업군 중 하나로 잘 알려져 있으며, 이에 따른 체계적인 안전관리가 필수적이다. 미국노동통계국(Bureau of Labor Statistics; BLS)에 따르면, 2023년 민간분야에서 건설업은 총 1,075명의 사망자를 기록하며 모든 산업 중 가장 높은 수치를 보였다(BLS, 2024). 또한, 대한민국 고용노동부에 의하면 2024년 전체 산업 사망자 598명 중 건설업 관련 사망자는 276명으로 약 46%에 해당하는 것으로 나타났다(고용노동부, 2025).

안전사고를 효과적으로 예방하기 위해서는 위험 요소들을 식별하고, 관련 규정에 따라 이를 제거하거나 개선함으로써 안전한 작업환경을 유지하는 것이 중요하다. 그러나 현재 대부분의 점검활동은 전문인력의 경험과 역량에 의존하고 있어 점검품질의 일관성을 확보하는 것에 한계가 있다. 더불어, 안전에 대한 사회적 요구가 확대됨에 따라 점검 과정과 결과에 대한 체계적인 문서화, 명확한 근거 제시, 법적 요구사항 준수를 통한 점검품질 확보가 강조되고 있다. 그러나 이러한 문서화 요구사항의 증가는 현장 전문인력의 업무 부담을 가중하며, 특히 복잡하고 다양한 규정을 근거로 한 문서 작성 작업은 시간 소모적이고 노동집약적이다. 또한, 문서화 과정에서 발생할 수 있는 누락이나 오류는 점검의 신뢰성을 훼손시킬 우려가 있으며, 과도한 문서 작성 업무로 인해 현장 안전점검에 집중해야 할 시간과 노력이 분산되어 안전관리의 본질적 목적 달성을 저해할 수 있다(Seo et al., 2015).

멀티모달 대형언어모델(Multimodal Large Language Model; MLLM)의 발전은 이러한 문제점의 해결 가능성을 제시한다. 기존의 컴퓨터비전 및 자연어 처리 기술은 건설현장의 복잡한 상황을 종합적으로 이해하고, 관련 지식과 연계하여 처리하는 데 한계가 있었다. 이에 반해 MLLM은 다양한 유형의 데이터를 동시에 처리하고 통합함으로써 복합적인 상황을 보다 정밀하게 인식하고 추론할 수 있다(Yin et al., 2024). 나아가 생성형 AI 기반 에이전트 개념의 등장으로 위임된 목표(delegated objective)를 달성하기 위해 자율적으로 행동하며 인간이 수행하던 작업 프로세스의 일부를 대체할 수 있게 되었다(Acharya et al., 2025).

이에, 본 연구에서는 건설현장에서 촬영된 이미지로부터 위험 요소를 검출하고 규정과 연계된 조치사항을 제시하기 위한 MLLM 기반 다중 에이전트 시스템을 제안한다.

* 서울시립대학교 대학원 건축공학전공 석박사통합과정 수료
** 서울시립대학교 대학원 건축공학전공 석박사통합과정
*** 서울시립대학교 도시과학대학 건축학부 교수, 공학박사
(Corresponding author : Department of Architectural Engineering, University of Seoul, sanghoon.lee@uos.ac.kr)
이 연구는 2025년도 ㈜건축사사무소건원엔지니어링 연구비 지원에 의한 결과의 일부임. 과제번호:202507302002.

이를 통해, 안전점검 문서화 작업을 자동화하여 전문인력의 업무 효율성과 점검 결과의 품질을 향상한다.

1.2 연구의 범위 및 방법

본 연구에서는 공사객체 중 가장 안전사고가 많이 발생하는 가시설을 주요 대상으로 설정하였다(국토안전관리원, 2025). 특히 가시설 사고 중 높은 비율을 차지하는 비계에 대한 시설물 안전관리에 초점을 맞추어 연구 범위를 한정하였다. 이때, 비계 안전관리의 근거 기준으로 한국산업안전보건공단(Korea Occupational Safety and Health Agency; KOSHA)의 기술지원규정을 적용하였다.

본 연구 과정은 그림1과 같다. 우선, 문헌고찰을 통해 기존 연구의 한계와 본 연구의 필요성을 도출한다. 다음으로, MLLM의 미세조정(Fine-tuning)을 위하여 이미지와 해당 이미지에 대응하는 텍스트 데이터를 수집 및 전처리한다. 동시에, 안전 대응 지침의 근거가 되는 KOSHA 규정을 추출 및 재구조화한다. 이후, 안전 특화 AI 에이전트를 구축하기 위해 사전에 전처리한 이미지-텍스트 데이터셋을 활용하여 선정된 MLLM을 미세조정 한다. 이전 단계에서 구조화한 규정 데이터는 규정DB에 저장되며, 이를 활용한 검색증강생성(Retrieval Augmented Generation; RAG) 기반(Lewis, et al., 2020)의 검색 및 보고서 작성 AI 에이전트들을 구축한다. 마지막으로, 각각의 AI 에이전트를 통합한 시스템 프로토타입을 개발하고, 생성형 AI 및 RAG 평가지표에 기반한 성능 평가를 수행한 후 결과를 분석한다.

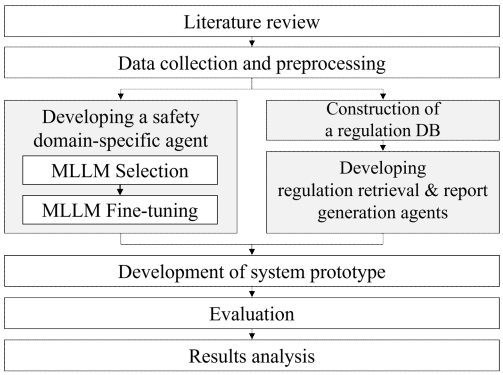


그림1. 연구 수행 절차

2. 선행연구 고찰

최근 건설안전 분야에서 MLLM 또는 LMM(Large Multimodal Model)을 적용한 연구들이 활발히 이루어지고 있다. 기존의 전통적인 컴퓨터비전 기반 접근법은 객체 탐지나 의미론적 분할과 같은 시각적 특징 검출에서는 우수한 성능을 보여 왔으나, 복잡한 현장 상황에 대한 맥락적 이해와 상황 해석 능력이 부족하여 적용 범위가 제한적이었다. 또한, 기존의 자연어 처리 기술은 텍스트 정보에만 의존하기 때문에, 시각적 단서와 환경적 요소가 결여된 상태에서 현장 상황을 온전히 해석하기에 한계가 있었다.

이러한 기술적 제한을 극복하고 건설현장에서 발생할 수 있는 다양한 안전사고에 대응하기 위해 MLLM 기반 접근법이 도입되고 있다. Uhm et al.(2025)은 LMM과 Graph RAG를 결합한 AcciVid 시스템을 통해 건설 안전사고 영상에 대한 KOSHA 규정 기반 분석을 실현하였다. Tsai et al.(2025)은 CLIP(Contrastive Language-Image Pre-training) 기반 모델을 활용하여 9가지 안전 위반유형에 대한 분류 및 위반 내용에 대한 장면 묘사 자동화를 구현하였다. 유사한 연구로, Tran et al.(2025)은 시각적 증거를 포함한 건설현장 안전 분석을 위하여 OSHA(Occupational Safety and Health Agency) 기준에 기반한 16가지 시나리오와 질문에 대한 VQA(Visual Question Answering)와 RES(Referring Expression Segmentation)를 결합한 접근법을 제안하였다. 이러한 연구들을 통해 MLLM 기반 접근법이 기존 안전관리의 효율성을 높이고, 기술 적용의 실효성을 확보하는 데 기여하고 있음을 확인할 수 있다.

하지만, 기존 연구들은 몇 가지 한계점을 보인다. 첫째, 대부분의 연구들이 특정 안전 위반유형이나 시나리오에 한정된 접근법을 채택함으로써, 현장에서 발생할 수 있는 다양하고 복합적인 안전 상황에 대한 포괄적 분석에 제한이 있다. 둘째, 기존 연구들은 주로 위험 요소의 식별과 분류에 초점을 맞추고 있으며, 인공지능 모델이 검출한 위험 요소와 연계한 체계적인 지침 제공이 부족하다. 셋째, 단일 모델 기반 접근법으로 인해 위험 인식, 규정 매칭, 조치사항 생성 등의 각기 다른 특성을 가진 작업들을 통합적으로 수행하는 데 어려움이 있다. 따라서 이러한 도전과제들을 극복하고 보다 실용적이고 포괄적인 건설 안전점검 시스템을 구축하기 위한 새로운 접근법이 필요하다.

3. MLLM 기반 다중 에이전트 시스템

3.1 시스템 개요

본 연구에서 제안하는 시스템은 이미지로부터 식별된 위험 요소를 KOSHA 규정과 연계하여, 구조화된 보고서를 생성 및 제공하는 것을 목표로 한다. 시스템은 그림2와 같이 세 가지 핵심 모듈로 구성된다.

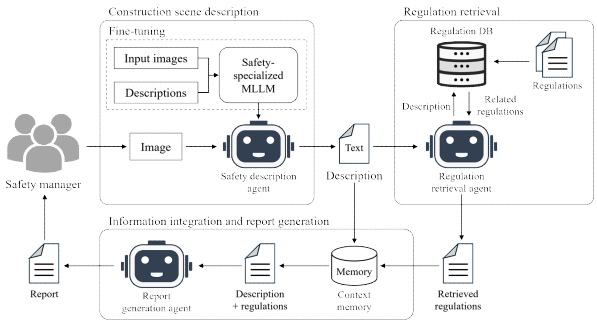


그림2. 시스템 프레임워크

첫째, construction scene description 모듈은 건설현장에서 취득된 이미지를 safety description 에이전트에 입력하여 위험 요소에 대한 캡션(caption)을 생성한다. 비계 관련 위험 요소 식별에 특화되도록 미세조정된 MLLM이 적용되며, 이미지 분석을 통한 상황 인식을 가능하게 한다. 둘째, regulation retrieval 모듈은 생성된 캡션을 바탕으로 규정DB로부터 관련 규정들을 검색하여 반환한다. 이때, 맥락 메모리(context memory)에는 서로 다른 시점에서 생

성된 캡션 및 규정 정보가 저장된다. 마지막으로, information integration and report generation 모듈에서는 report generation 에이전트가 맥락 메모리로부터 캡션과 관련 규정을 전달받아 조치사항을 생성하고, 최종적으로 보고서 형식의 결과물을 출력한다.

3.2 Construction Scene Description 모듈

해당 모듈에서는 safety description 에이전트를 구축하기 위해 사전학습된 MLLM을 미세조정하였다. 이를 위해 비계 관련 이미지-텍스트 데이터 400장을 수집하였으며, 각 데이터는 안전 전문가가 작성한 위반사항 설명을 포함한다. 본 연구는 LLaVA(Large Language and Vision Assistant)를 기반모델로 활용하였으며(Liu et al., 2023), 제한된 데이터셋과 컴퓨팅 자원 환경에서도 효과적인 미세조정을 수행하기 위하여 LoRA(Low-Rank Adaptation) 기법을 적용하였다(Hu et al., 2021). 이를 통해 사용자가 입력한 이미지에 대한 안전 위반사항을 서술할 수 있도록 하였으며, 생성된 캡션은 맥락 메모리에 저장된다.

3.3 Regulation Retrieval 모듈

본 모듈은 생성된 캡션에 포함된 위험 요소를 식별하고, 이와 관련된 규정을 체계적으로 제시하기 위해 RAG 기법을 적용하였다. 연구의 범위를 고려하여 KOSHA 이동식 비계(C-28-2018)와 시스템 비계(C-32-2020)를 대상 규정으로 선정하였다. 문서들은 문장 단위로 분할되며, 각 문장을 'id', 'content', 'metadata(code, category, source)' 스키마를 갖는 JSON 구조로 변환하였다. 이후, OpenAI의 임베딩 모델인 'text-embedding-3-small'을 이용하여 'content'를 벡터화하고 벡터DB에 저장함으로써 규정 DB를 구축하였다. 벡터DB로는 ChromaDB를 활용했다.

본 모듈의 핵심인 regulation retrieval 에이전트의 기반 모델로 OpenAI gpt-4o가 적용되었다. 해당 에이전트는 입력 문장의 핵심어를 추출한 뒤 각 핵심어에 대하여 코사인 유사도 기반 최근접 검색(top-k)을 수행한다. 이후, 반환된 규정들을 통합 및 요약하고, 키워드별 관련된 규정들과 함께 JSON 구조로 출력하여 맥락 메모리로 전달한다.

3.4 Information Integration and Report Generation 모듈

본 모듈에서는 report generation 에이전트가 맥락 메모리에 저장된 캡션과 검색된 규정을 종합하여 최종 보고서를 생성한다. 해당 에이전트의 기반모델로 OpenAI gpt-4o를 적용했다. 보고서 생성은 사전에 정의된 구조화된 프롬프트를 기반으로 수행되며, 결과물에는 안전 규정 위반사항 및 관련 규정, 조치 및 개선방안 등이 포함된다.

4. 성능 평가 및 결과 분석

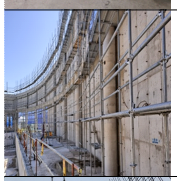
본 연구에서는 캡션 생성 및 RAG 성능을 구분하여 평가를 수행하였다. 우선, 생성된 캡션과 정답의 유사도를 비교하기 위해 캡션 생성 task에서 주요 평가지표로 사용되는 BERTScore(Zhang, et al., 2020), BLEU(Papineni, et al., 2002), METEOR(Denkowski & Lavie, 2014), ROUGE-L(Lin, 2004), CIDEr(Vedantan, et al., 2015)를 사용하였다. 비교 모델로는 OpenAI의 gpt-4o를 채택하였다. 평가 결과, 표1과 같이 미세조정된 LLaVA가 gpt-4o 대비 전 지표에서 우세했다. 이는 상대적으로 적은 데이터를 활용하여 미세조정된 모델이 범용 모델 대비 의미적, 구문적, 구조적 측면에서 성능이 향상되었음을 보여준다.

표1. 미세조정된 MLLM과 gpt-4o 모델 비교

Metric \ Model	Fine-tuned LLaVA	gpt-4o (OpenAI)
BERTScore	0.8017	0.7372
BLEU_1	0.3000	0.1890
BLEU_2	0.1837	0.1052
METEOR	0.1437	0.0785
ROUGE_L	0.3217	0.1850
CIDEr	0.4274	0.2536

RAG에 대한 성능 평가는 Es et al.(2025)이 제시한 RAGAs(Retrieval Augmented Generation Assessment) 프레임워크를 적용했다. 표2와 같이 5개의 이미지-생성 캡션을 바탕으로 규정DB와 연계하여 생성된 답변에 대한 faithfulness, factual correctness, context recall을 측정했다. Faithfulness는 생성된 답변 내 주장들이 검색된 문서에 의해 뒷받침되는 정도를 의미하며, factual correctness는 생성된 답변과 참조 답변 사이의 일치도를 평가한다. 마지막으로, context recall은 답변 생성을 위한 관련 정보가 검색된 문서에 얼마나 포함되어 있는지를 평가한다.

표2. 생성된 이미지 캡션 기반 RAG 성능 평가

Image & Caption \ Metric		Faithful.	Factual correct.	Context recall
	Mobile scaffold installation standard not complied (front guardrail, platform, etc.)	0.897	0.650	0.846
	System scaffold side safety net not installed, brace condition check required	0.625	0.550	1.000
	System scaffold lower horizontal member arbitrarily dismantled	0.409	0.520	1.000
	System scaffold fixation defect due to middle guardrail not installed on work platform	0.435	0.430	0.727
	The mobile scaffold safety guardrail and outrigger have not been installed, and the load capacity is not indicated	0.238	0.760	1.000

평가 결과, 상대적으로 높은 context recall 성능은 RAG 시스템이 캡션과 관련된 문서를 효과적으로 검색하고 있음을 보여준다. 반면, factual correctness와 faithfulness 점수의 변동성은 검색된 문서로부터 신뢰할 수 있는 답변을 생성하는 과정에서 캡션에 따른 편차가 존재함을 의미한다. 특히, 가장 큰 변동성을 보인 faithfulness는 검색된 문서와 생성된 답변 간의 정합성이 입력 캡션의 복잡성이나 길이에 따라 달라질 수 있음을 시사한다.

최종적으로, 그림3에서 보여주는 보고서는 현장 이미지를 기반으로 캡션 생성 및 관련 규정 검색을 수행하고, 해당 정보들을 통합하여 생성된 결과물이다. 이는 제안된 시스템이 안전점검 사항을 체계적으로 정리하고 문서화할 수 있음을 보여준다.

그림3. 보고서 생성 예시

5. 결론

본 연구에서는 KOSHA 규정 연계 비계 안전점검 자동화를 위한 MLLM 기반 다중 에이전트 시스템을 제안했다. 제안된 시스템은 이미지로부터 안전 관련 캡션을 생성하고, 이를 바탕으로 연관된 규정을 검색 및 활용하여 보고서를 자동으로 생성한다. 성능 평가 결과, 미세조정된 MLLM이 범용 모델 대비 전반적으로 우수한 성능을 보였다. RAG 시스템은 높은 context recall 성능으로 효과적인 규정 검색이 가능함을 검증하였으며, 생성된 정보들을 바탕으로 보고서를 생성할 수 있음을 확인했다. 이를 통해, 기존의 인력 의존적 안전점검과 문서화의 한계를 극복하고, 규정 기반의 체계적인 점검 수행 가능성을 확인했다.

그러나, faithfulness와 factual correctness의 변동성은 향후 프롬프트 최적화와 캡션 복잡도를 고려한 적응적 답변 생성을 위한 추가적인 연구가 필요함을 보여준다. 또한, 현재 비계에 한정된 연구 범위를 다양한 공종 및 공사 객체로 확대하여 범용성을 높일 필요가 있다. 이에 따른 데이터 불균형 및 품질 문제를 해결하기 위해서는 합성데이터 생성 및 지식 증류 기법 등의 추가적인 연구가 필요하다. 본 연구에서는 시스템의 기술적 타당성 검증에 초점을 맞추었으나, 향후 모바일 기반 안전점검 애플리케이션 개발을 통해 실사용자를 대상으로 현장 적용성과 실효성을 검증할 계획이다. 이러한 향후 연구를 통한 한계점 개선 및 현장 검증을 바탕으로 안전사고 예방과 안전관리 효율성 향상에 기여할 수 있을 것으로 기대된다.

참고문헌

1. Wanberg, J. et al. (2013). Relationship between construction safety and quality performance. J. Constr. Eng. Manage. 139, 04013003.
2. U.S. Bureau of Labor Statistics. (2024). National Census of Fatal Occupational Injuries in 2023
3. 고용노동부, 중대재해통계, 2025
4. Seo, J. et al. (2015). Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inform. 29 239–251.
5. Yin, S. et al. (2024). A survey on multimodal large language models, Natl. Sci. Rev. 11, nwae403.
6. Acharya, D.B. et al. (2025). Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey, IEEE Access 13 18912–18936.
7. 국토안전관리원, 2024년 건설사고정보 리포트, 2025
8. Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS 2020.
9. Liu, H. et al. (2024). Visual instruction tuning. NeurIPS 2023.
10. Hu, E. et al. (2021). LORA: Low-Rank Adaptation of Large Language Models, arXiv preprint arXiv:2106.09685v2
11. Uhm, M. et al. (2025). Automated analysis of construction safety accident videos using a large multimodal model and graph retrieval-augmented generation, Autom. Constr. 177, 106363.
12. Tsai, W.-L. et al. (2025). Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention, Autom. Constr. 169, 105863.
13. Tran, D.Q. et al. (2025). Visual Question Answering-based Referring Expression Segmentation for construction safety analysis, Autom. Constr. 174, 106127.
14. Zhang, T. et al. (2020). BERTScore: Evaluating text generation with BERT. ICLR 2020.
15. Papineni, K. et al. (2002). Bleu: A method for automatic evaluation of machine translation. In Proc. 40th Annual Meeting of the ACL (pp. 311–318).
16. Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In Proc. 9th Workshop on Statistical Machine Translation (pp. 376–380).
17. Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74–81).
18. Vedantam, R. et al. (2015). CIDEr: Consensus-based image description evaluation. IEEE CVPR, pp. 4566–4575.
19. Es, S. et al. (2025). Ragas: Automated Evaluation of Retrieval Augmented Generation, arXiv preprint arXiv:2309.15217v2.