SAGE: A Knowledge-Grounded Multi-Agent Framework for Automated MEP Drawing Interpretation

Vivek Vishwas Vichare[1], Pratik Khandelwal[1], Aditya Debnath[2], Divya Singh Rathore[3], Gauri Lamb[4], and Paul O'Neill[4]

[1]Crafted Data Labs Private Limited (Encraft AI), India, [2]Indian Institute of Technology Kharagpur, India, [3]Indian Institute of Technology (BHU) Varanasi, India, [4]FHP Group, United Kingdom
vivek@encraft.ai, pratik@bingeclip.ai, aditya0202debnath@kgpian.iitkgp.ac.in, divyasingh.rathore.phy19@itbhu.ac.in, g.lamb@fhpgroup.com, P.O'Neill@fhpgroup.com

Abstract – Mechanical, Electrical, and Plumbing (MEP) engineering drawings serve as the contractual foundation for construction projects, yet verifying their consistency remains manual and error-prone. While Vision-Language Models show promise, they exhibit symbol fragility, spatial reasoning failures, and limited cross-document reasoning on technical drawings. We propose SAGE, a multi-agent framework employing: (1) YOLO11 for region segmentation (0.912 mAP50); (2) Cache-Augmented Generation for project-specific knowledge grounding; (3) a dual-mode adaptive extraction engine with coordinate grid overlay; and (4) hierarchical assembly combining extractions across project drawings. Evaluation across three construction projects (23 drawings, 1043 components) demonstrated 91.9% precision, 98.6% recall, and 97.4% spatial accuracy—substantially outperforming zero-shot VLM baselines (GPT 5.2, Claude Opus 4.6, Gemini 3 Pro) which achieved comparable precision but recall degrading sharply with drawing complexity (20–29% across projects vs. SAGE's consistent 98–100%). Ablation studies confirm the critical role of each architectural component.

Keywords –
MEP Drawings; Vision-Language Models; Multi-Agent Systems; Construction Automation; Drawing Interpretation

1 Introduction

The construction industry struggles to maintain consistency between CAD drawings and specifications, especially in retrofit projects where existing conditions overlap with new designs. Most design errors arise during handovers between architects, engineers, and contractors, causing costly rework, delays, and material conflicts. MEP systems account for over 50% of project investment [1], while rework consumes 4–10% of costs—52% linked to poor data and miscommunication [2] leading to $88.7 billion in global losses and $15.8 billion in annual interoperability costs [3]. Recent Vision-Language Models (VLMs) show promise for document understanding [4] but still face spatial reasoning and symbol interpretation challenges [5]. Fine-tuning improves accuracy—up to 52% F1 gain on GD&T extraction [6]—yet scales poorly with new component types.

In MEP projects, information flows sequentially from architects to MEP engineers to contractors, with each handover introducing potential for specification drift and interpretation errors. Design-intent drawings may reference equipment schedules, legend sheets, and specification documents that reside in separate files, creating cross-document dependencies that manual review struggles to verify consistently. Discrepancies such as mismatched equipment ratings between schedules and plan drawings, or spatial conflicts between HVAC ductwork and electrical conduit routing, often remain undetected until construction, when correction costs escalate significantly.

This paper presents SAGE, a multi-agent framework addressing these limitations through: (1) YOLO11 [7]-based region segmentation isolating drawing blocks, legends, notes, and title blocks; (2) Cache-Augmented Generation for knowledge grounding, enabling project-specific symbol interpretation; (3) a dual-mode adaptive extraction engine with coordinate grid overlay for spatial referencing; (4) hierarchical assembly combining extractions across all project drawings; and (5) industry validation on real construction projects. Unlike prior automated drawing interpretation pipelines that rely on static processing, SAGE introduces four novel mechanisms: (a) an adaptive grid density function that dynamically adjusts tiling granularity based on measured drawing complexity; (b) a Surgical Merge Agent implementing zone-based deduplication for overlapping tile regions; (c) the first application of Cache-Augmented Generation for domain-specific symbol grounding in

technical drawings; and (d) a multi-agent orchestration pattern with hierarchical semantic assembly, representing a principled four-stage decomposition specifically designed for the MEP drawing domain, where each agent targets an identified VLM failure mode through focused task specialisation. Importantly, as the underlying VLM capabilities continue to advance, the agentic architecture is designed to inherit these improvements—stronger base models directly translate into improved extraction quality within the orchestrated pipeline, without requiring architectural changes.

2 Related Work

Engineering drawing digitization has evolved from rule-based systems to deep learning approaches [8]. For P&ID diagrams, deep learning pipelines have achieved 94.5% symbol recognition accuracy [9], while transformer-based approaches demonstrate 83.6% AP for simultaneous node and relationship detection [10]. Hybrid frameworks combining rotation-aware object detection with transformer-based vision-language parsers have demonstrated promising results for mechanical drawings [11].

VLMs face significant challenges with technical drawings, exhibiting 10-30% hallucination rates on complex scenes [12] and 50-60% spatial reasoning accuracy. Recent work demonstrates that fine-tuned open-source VLMs (e.g., Florence-2 with 0.23B parameters) can achieve 52% F1 improvement over zero-shot GPT-4o and Claude-3.5-Sonnet [6]. However, fine-tuning suffers scaling limitations, motivating our agentic approach leveraging zero-shot VLMs with structured knowledge grounding. Set-of-Mark (SoM) prompting enhances VLM visual grounding by overlaying markers on image regions [13], providing foundation for SAGE's grid overlay. Multi-agent architectures like AutoGen enable task decomposition with iterative refinement [14].

Despite these advances, no existing framework addresses the complete MEP drawing interpretation pipeline—from region isolation through cross-document knowledge grounding to spatially-aware component extraction. Rule-based systems lack the semantic flexibility to handle diverse drafting conventions across projects, while fine-tuned VLMs require costly per-project retraining. SAGE bridges this gap through a multi-agent architecture that preserves zero-shot VLM generality while grounding interpretation in project-specific context via CAG, eliminating the need for per-project fine-tuning. Notably, because SAGE treats the VLM as a replaceable inference engine within a structured agentic workflow, improvements in base model capabilities—such as enhanced spatial reasoning or reduced hallucination in future VLM generations—are expected to propagate directly into improved pipeline performance without architectural modification.

3 VLM Limitations on MEP Drawings

Systematic evaluation of zero-shot VLMs (GPT-4V, Claude 3.5, Gemini Pro Vision) on real-world MEP drawings identified four critical failure modes: Symbol Fragility—custom drafting symbols cause systematic misclassification without legend context; Spatial Reasoning Failure—models fail to establish spatial relationships, achieving only 50-60% accuracy [12]; OCR Degradation—performance on rotated, scaled, or densely packed annotations degrades significantly [15]; Limited Cross-Document Reasoning: VLMs struggle to incorporate information from referenced schedules and legend sheets. These findings motivated SAGE's multi-agent architecture, where specialized agents address each limitation through focused task decomposition. Section 5 presents a systematic quantitative baseline study using current-generation models (GPT-5.2, Claude Opus 4.6, Gemini 3 Pro) to validate that these limitations persist and intensify with drawing complexity despite substantial advances in base model capabilities.

4 Methodology

4.1 Framework Overview

SAGE decomposes MEP drawing interpretation into four sequential stages following multi-agent design principles [14]: S1 (Region Isolation) for semantic segmentation into functional regions; S2 (Knowledge Grounding) for pre-loading project-specific context; S3 (Adaptive Extraction) for dual-mode component extraction with coordinate grid overlay; and S4 (Assembly) for synthesis into structured semantic representation as shown in the Figure 1. We utilised Gemini 3 Pro (Vision API) for developing extractor, merging and assembly agents due to its superior visual reasoning capabilities. Each stage introduces a targeted mechanism addressing a specific VLM failure mode identified in Section 3: region isolation reduces visual search space, CAG eliminates symbol ambiguity, the adaptive grid forces exhaustive spatial traversal, and hierarchical assembly enables cross-document reasoning.

4.2 Stage 1: Region Isolation

YOLO11, fine-tuned on 200 annotated MEP drawings, partitions each sheet into four semantic region classes: Drawing Blocks (primary schematic content), Legend Blocks (symbol definitions), Note Blocks (specifications), and Title Blocks (drawing metadata).

The model achieved 0.912 mAP50 and 0.696 mAP50-95 on held-out test data.

4.3 Stage 2: Knowledge Grounding

SAGE constructs project-specific context by aggregating legend symbol-to-specification mappings, note block content, and project documentation. This context is pre-loaded using Cache-Augmented Generation (CAG) [16], enabling interpretation of project-specific symbol conventions.

4.4 Stage 3: Adaptive Extraction Engine

The extraction implements dual-mode processing with adaptive visual grid overlay inspired by Set-of-Mark prompting. Grid density G (Complexity Decision) = f(ρs, ρt, R) is computed from symbol density (ρs), text density (ρt), and resolution (R), estimated via a preliminary VLM assessment call on each Drawing Block. Standard-density drawings are processed by a single VLM instance (Baseline Mode); high-density schematics are partitioned into overlapping sub-regions processed in parallel (Dynamic Tiling Mode). Both modes extract: (1) component identification, (2) metadata (specifications, model numbers, sizing), and (3) grid cell location.

In Dynamic Tiling mode, 10% tile overlap ensures boundary components are fully visible but introduces redundant detections. A "Surgical" Merge Agent accepts components outside the seam zone as "Safe Entities" while forwarding "Conflict Candidates" within overlap boundaries for IoU-based deduplication and attribute union—keeping processing linear to overlap count.

4.5 Stage 4: Semantic Model Assembly

The Aggregation Module combines results from all project drawings into unified JSON representation. Stage 4 performs hierarchical merger across drawing blocks via Global Coordinate Registration: local coordinates transform to global system using block origins from YOLO segmentation. Output encodes component instances, spatial relationships from grid proximity, and system membership classification.

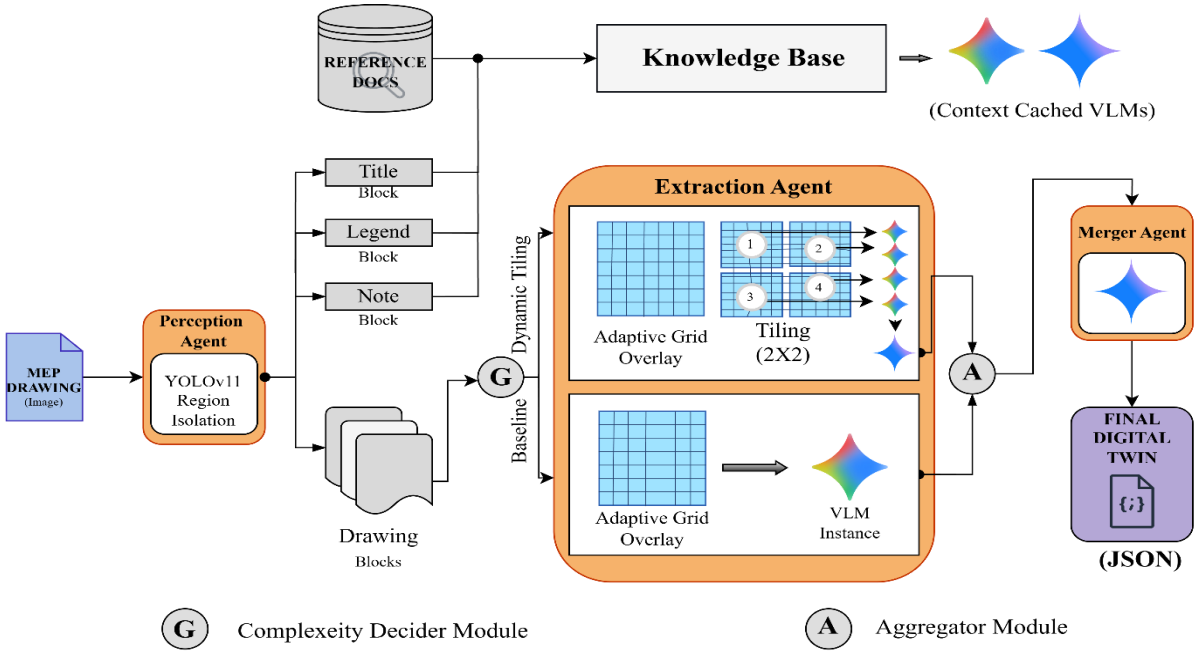

Figure 1. Multi-Agent Architecture (SAGE) for MEP drawing interpretation and Semantic Model Generation

5 Industry Case Study

5.1 Experimental Setup

We evaluated SAGE across three active construction projects provided by FHP Group (UK-based MEP consultancy). The dataset comprised 23 drawing sheets with 1,043 total extracted components spanning diverse MEP systems as detailed in Table 1. All drawings were exported as PDF from the original CAD/DWG files, representing the typical document format used in contractor handover and tendering workflows.

Table 1. Dataset Characteristics

Project	# of Drawings	Components	MEP System
1	5	66	Controls/BMS, Electrical, HVAC
2	6	339	HVAC, Electrical
3	12	638	Electrical, Public Health

5.2 Results

Results were obtained in JSON format as shown in Figure 2 and were accordingly compared with the ground truth validated by the expert design team at FHP Group.

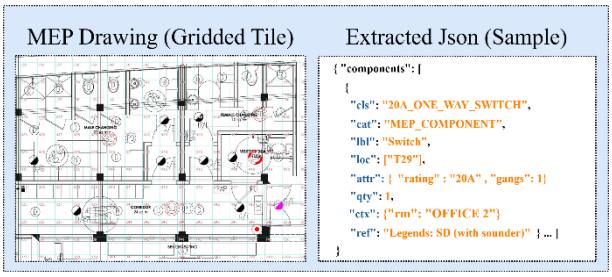

Figure 2. Sample extraction result showing a localized MEP drawing tile and its corresponding semantic JSON representation.

Precision measures the proportion of SAGE's extracted MEP components that match ground truth annotations, while recall captures the proportion of actual drawing components successfully detected. Spatial accuracy evaluates whether components are correctly localized to their grid cell positions.

Table 2. Evaluation Metrics

Project	Precision (%)	Recall (%)	F1-Score (%)	Spatial Accuracy (%)
1	100.0	98.48	99.24	100.0
2	90.16	99.71	94.69	97.05
3	92.05	97.96	94.91	97.44

To systematically quantify the failure modes identified qualitatively on earlier models, we conducted a baseline evaluation using current-generation VLMs. Notably, SAGE itself employs Gemini 3 Pro as its underlying VLM, making the Gemini 3 Pro baseline a direct controlled comparison—isolating the contribution of the agentic pipeline (region isolation, knowledge grounding, adaptive tiling, hierarchical assembly) from the base model's zero-shot capabilities. Spatial accuracy (Table 2) is excluded from the baseline comparison because it is architecturally inseparable from SAGE's coordinate grid overlay: without the grid, zero-shot VLMs produce only coarse qualitative localisations (e.g., "upper-left area") that are not commensurable with the grid-cell-level metric. The baseline comparison therefore focuses on detection quality (Precision, Recall, F1), where the models can be compared on equal terms. The results confirm that the same failure patterns persist even with substantially more capable models—and intensify with drawing complexity. Zero-shot VLMs achieve precision comparable to SAGE (88–91%) but weighted average recall of only 20–29%, with performance degrading sharply on complex drawings: on Project 1 (66 components), recall ranged from 51–92%, while on Project 3 (638 components, dense multi-system layouts), it collapsed to 8–23% across all three baselines. SAGE maintains near-complete recall (98–100%) regardless of drawing complexity. This demonstrates that raw VLM visual reasoning is adequate for sparse, well-separated layouts but fundamentally insufficient for the dense, information-rich drawings typical of real-world MEP projects; a limitation that persists across model generations. It is worth noting that as VLM architectures continue to mature with improved spatial reasoning, reduced hallucination, and stronger OCR on rotated annotations—these baseline capabilities are expected to improve. Because SAGE employs VLMs as inference components within a structured multi-agent workflow rather than as monolithic end-to-end systems, such advances in base model performance will propagate directly into improved pipeline-level metrics without requiring architectural changes to the framework.

Table 3. Zero-Shot VLM Baseline Comparison (Average across 3 Projects)

Model	Precision (%)	Recall (%)	F1-Score (%)
GPT 5.2	88.2	19.5	31.0
Claude Opus 4.6	89.3	29.3	43.3
Gemini 3 Pro	90.8	23.0	32.3
SAGE	91.9	98.6	95.1

As with the baseline comparison, the ablation evaluates detection quality rather than spatial accuracy, which is architecturally inseparable from the grid overlay.

Table 4. Ablation Study (Weighted Average across projects)

Configuration	Precision (%)	Recall (%)	F1-Score (%)
SAGE w/o CAG	50.5	49.7	48.0
SAGE w/o Region Isolation	91.5	79.5	86.2
SAGE w/o Grid Overlay	90.3	75.6	82.3
SAGE w/o Dynamic Tiling	89.9	45.2	59.3
SAGE	91.9	98.6	95.1

Project 1 (66 components, low-density) showed negligible ablation effects, as simpler layouts fall within zero-shot VLM capability; the impact is concentrated on Projects 2 and 3. Removing CAG caused the most severe degradation (F1: 95.1%→48.0%), as the VLM hallucinates without legend-to-specification grounding for visually homogeneous MEP symbols. Removing dynamic tiling caused the largest recall drop (recall: 98.6%→45.2%, F1→59.3%), as single-pass extraction misses dense component clusters. Removing the grid overlay maintained precision (90.3%) but penalised recall (75.6%), confirming its role as a deterministic spatial traversal mechanism. Removing YOLO produced moderate degradation (F1: 86.2%) from increased background noise exposure. The full pipeline achieves the optimal balance: CAG grounds semantics, the grid ensures spatial completeness, YOLO isolates signal from noise, and tiling ensures exhaustive extraction.

6 Discussions

SAGE demonstrates that knowledge-grounded multi-agent architectures effectively overcome zero-shot VLM limitations in technical drawings. The coordinate grid overlay achieved 97.4% spatial accuracy, while the Surgical Merge Agent eliminated duplicate detections in low-to-medium density cases. High-density seam regions remain challenging. The baseline comparison (Table 3) confirms that zero-shot VLMs maintain high precision (88–91%) but only 20–29% weighted recall. This high-precision/low-recall pattern reflects three limitations: VLMs lack mechanisms for exhaustive spatial scanning of dense technical layouts; project-specific drafting symbols fall outside pre-training distributions, causing systematic omission; and attention mechanisms struggle with counting repetitive elements in dense arrays. SAGE targets each limitation through CAG (symbol grounding), the adaptive grid (exhaustive traversal), and YOLO (signal-to-noise isolation). The framework accurately extracted and validated drawing details against specifications, identifying discrepancies in parameters such as pressure and power ratings—highlighting its value for automated compliance verification.

Current limitations include incomplete topological inference, reduced counting accuracy in dense regions, and a 3× computational overhead from dynamic tiling. Average processing time per drawing block is ~20 seconds when tiling is not included and ~120 seconds in Dynamic Tiling Mode. The grid density function $G = f(\rho s, \rho t, R)$ determines mode selection based on symbol density, text density, and resolution; across our dataset, approximately 40% of drawing blocks exceeded the complexity threshold triggering Dynamic Tiling, yielding an effective ~3× computational overhead at the project level. This cost is concentrated on high-density drawings where the accuracy-cost trade-off is most favourable. Early experiments using recursive legend-symbol prompting and integrating non-VLM methods (e.g., YOLO, R-CNN) show promise for improving counting.

Future work will extend region detection to new drawing types, explore graph-based spatial reasoning, adapt tile sizes dynamically, and develop hybrid pipelines combining VLM semantics with deterministic detection for greater robustness.

7 Conclusions

This paper presented SAGE, a multi-agent framework enabling automated semantic model generation from MEP drawings, providing AI-powered quality assurance before contractor handover. Evaluation across three real-world construction projects demonstrated 91.9% precision and 98.6% recall (F1: 95.1%), with 97.4% spatial accuracy. Zero-shot VLM baselines (GPT-5.2, Claude Opus 4.6, Gemini 3 Pro) achieved comparable precision (88–91%) but only 20–29% weighted average recall, confirming that knowledge grounding and structured spatial traversal, not raw visual reasoning alone, are necessary for reliable MEP drawing interpretation. Ablation studies confirmed the complementary roles of CAG (semantic grounding), adaptive grid overlay (spatial completeness), YOLO-based region isolation (signal-to-noise filtering), and dynamic tiling (exhaustive extraction in dense regions). As VLM capabilities continue to advance, SAGE's architecture is designed to inherit these improvements: stronger base models translate directly into enhanced pipeline performance without structural modification, positioning the framework for sustained relevance in construction automation workflows.

Acknowledgments

The authors thank FHP Group UK for providing project drawings and validation expertise.

References

[1] B. Wang et al. *Omni-Scan2BIM: A ready-to-use Scan2BIM approach based on vision foundation models for MEP scenes*, 162:105308, 2024.

[2] FMI Corporation. *Harnessing the Data Advantage in Engineering and Construction*, 2024.

[3] M. P. Gallaher et al. *Cost Analysis of Inadequate Interoperability*, NIST GCR 04-867, 2004.

[4] OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2023.

[5] C. Picard et al. *From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design*, Artificial Intelligence Review, 2024.

[6] M. T. Khan et al. *Fine-Tuning VLM for Engineering Drawing Extraction*. arXiv:2411.03707, 2024.

[7] G. Jocher and J. Qiu, Ultralytics YOLO11," Version 11.0.0, Ultralytics, 2024. [Online]

[8] C. F. Moreno-García et al. *Digitisation of complex engineering drawings*, Neural Computing and Applications, 31:1695-1712, 2019.

[9] G. Su et al. *P&ID recognition based on deep learning*, Biomimetic Intelligence and Robotics, 4(1):100142, 2024.

[10] R. Rahul et al. *P&ID Digitization using Transformers*, arXiv:2411.13929, 2024.

[11] M. T. Khan et al. *From Drawings to Decisions*, Robotics and Computer-Integrated Manufacturing, 2025.

[12] Y. Huang et al. *A Survey on Hallucination in Large Vision-Language Models*, arXiv:2402.00253, 2024.

[13] J. Yang et al. *Set-of-Mark Prompting: Improving Visual Grounding in VLMs*, arXiv:2310.11441, 2023.

[14] Q. Wu et al. *AutoGen: Enabling Next-Generation Multi-Agent Collaboration*, Microsoft Research, 2023.

[15] T. Li et al. *Visual Merit or Linguistic Crutch? A Close Look at DeepSeek-OCR*, arXiv:2601.03714, 2026.

[16] Chan et al. *Don't Do RAG: When Cache-Augmented Generation Is All You Need for Knowledge Tasks.*, *arXiv:2412.15605*, Dec 2024.