ICICLE: An Interactive VLM-based System for Information Extraction from Ambiguous Engineering Diagram Legends

ICICLE: An Interactive VLM-based System for Information Extraction from Ambiguous Engineering Diagram Legends

Vasil Shteriyanov
v.shteriyanov1@tue.nl
Mathematics and Computer Science, Eindhoven University of Technology
Eindhoven, The Netherlands

Rimma Dzhusupova
rdzhusupova@mcdermott.com
Engineering, McDermott
The Hague, The Netherlands

Jan Bosch
jan.bosch@chalmers.se
Computer Science and Engineering, Chalmers University of Technology
Gothenburg, Sweden

Helena Holmström Olsson
helena.holmstrom.olsson@mau.se
Computer Science and Media Technology, Malmö University
Malmö, Sweden

Abstract

Engineering legend sheets are vital in Engineering, Procurement, and Construction (EPC) projects, visually defining complex assemblies. Automating information extraction from these legends can significantly reduce manual effort, yet their spatially ambiguous layouts pose barriers to digital transformation. To address this, we introduce ICICLE (In-Context Interactive Cue-guided Legend Extractor), a Generative AI system for parsing these documents. ICICLE's core innovation is the In-Context Multimodal Annotation Prompting (ICMAP) method. This technique instructs a Vision Language Model (VLM) by providing a single annotated visual example alongside a textual prompt. This enables the VLM to ground abstract concepts like "detailed assembly" in visual evidence, making the system adaptable without per-query annotation or costly model retraining. Validated on legend sheets from four industrial projects, ICMAP achieved superior extraction accuracy (96–100%) compared to established alternatives and demonstrated robustness across varied visual examples. An ablation study confirms that the tight coupling of visual annotations and textual instructions is critical for success. ICICLE serves as a successful case study for operationalizing Generative AI, offering a blueprint for solving complex information extraction tasks in challenging industrial domains.

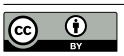
CCS Concepts

• **Applied computing** → **Engineering**; • **Computing methodologies** → **Computer vision**.

Keywords

Information extraction, Multimodal Prompt Engineering, Vision Language Models (VLMs), Engineering Diagrams, EPC

ACM Reference Format:
Vasil Shteriyanov, Rimma Dzhusupova, Jan Bosch, and Helena Holmström Olsson. 2026. ICICLE: An Interactive VLM-based System for Information Extraction from Ambiguous Engineering Diagram Legends. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3748522.3779736



SAC '26, Thessaloniki, Greece

ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3779736

1 Introduction

The Engineering, Procurement, and Construction (EPC) industry executes projects involving the construction of large-scale facilities. During the early project stages, engineers receive numerous engineering diagrams and legend sheets, often in PDF format to protect intellectual property [13]. Engineers are required to analyze these drawings and their corresponding legends to produce deliverables for later project stages. A critical task is understanding component assemblies, which are often depicted in a simplified representation on the main diagrams but explained in full detail on the project's legend sheets [14]. As a standard EPC project can involve hundreds of such drawings, the manual analysis of legends represents a significant expenditure of engineering hours and cost and a critical bottleneck in the digitalization of engineering workflows [12].

Automating the extraction of this structured information is therefore a high-impact goal for the digital transformation of the EPC industry. However, this task presents a distinct challenge for AI systems because engineering legends often contain semantically distinct but spatially ambiguous regions. As shown in Figure 1, a legend may depict a "detailed" and a "simplified" assembly side-by-side with no explicit visual separators like lines or boxes to distinguish them. Due to the absence of standardized formats, this ambiguous layout is common, making it difficult to automatically and correctly associate extracted information with its intended region.

The difficulty of handling such spatial ambiguity exposes specific limitations in current AI methodologies. Classical computer vision pipelines using Optical Character Recognition (OCR) can extract text but lack the contextual understanding to group it correctly without brittle, hand-crafted rules [3]. More advanced approaches using segmentation models like SAM [9] are designed to identify visually salient objects (e.g., individual shapes), not conceptual groupings like an "assembly", making them ill-suited for the task. The need for extensive, per-project fine-tuning for these methods limits their scalability and practicality in an industrial context [24]. This highlights the need for more advanced intelligent systems,

leading to our guiding research question: **"How can we engineer a practical and scalable intelligent system capable of automating information extraction from blueprint legends, including those characterized by ambiguous spatial layouts?"**

To answer this question, we designed, implemented, and evaluated ICICLE (In-Context Interactive Cue-guided Legend Extractor), a proof-of-concept intelligent system for automated engineering legend information extraction. This work isolates and expands upon the core Vision Language Model (VLM) component from our previously published hybrid pipeline for engineering document analysis [19]. While our prior study demonstrated this component's effectiveness within an end-to-end industrial workflow, its scope did not include a standalone analysis or a comparison against other VLM-based strategies.

This paper provides a focused investigation by formalizing the technical contribution as a multimodal prompting strategy we call In-Context Multimodal Annotation Prompting (ICMAP). Rather than relying on a VLM's pre-trained knowledge to infer structure from non-standard layouts, ICMAP explicitly anchors abstract concepts (e.g., "detailed region") to specific visual patterns via a single, annotated example image integrated with referential textual instructions. This approach allows the VLM to resolve spatial ambiguity through in-context learning, enabling the ICICLE system to accurately parse new legend sheets without the need for modification or fine-tuning. We validate the ICMAP strategy on a real-world dataset from four distinct industrial projects, conducting the comprehensive evaluation that was absent in our prior work, which includes ablation studies and comparisons against strong VLM-specific baselines. ICICLE serves as an industrial case study of operationalizing Generative AI in intelligent systems. It provides generalizable insights into the architectural choices, prompting strategies, and workflow integration required to move from a raw AI capability to a practical intelligent system. Our work makes the following contributions:

- A systematic investigation of existing AI methods on industrial diagrams, empirically demonstrating why they are insufficient for this automation task.
- The formalization and evaluation of ICMAP, a multimodal prompting strategy requiring no model fine-tuning that enables a VLM to extract information from engineering legends with ambiguous layout. We demonstrate that integrating a single annotated visual example with referential textual cues serves as a highly data-efficient alternative to fine-tuning.
- The design and implementation of ICICLE, a proof-of-concept intelligent system with a user interface that employs ICMAP and serves as a blueprint for integrating VLMs into practical engineering workflows.
- An empirical validation of ICICLE, achieving 96-100% accuracy on a proprietary industrial dataset. Our ablation studies provide quantitative evidence that the interplay between visual annotations and textual cues is a critical requirement for accuracy in this domain.

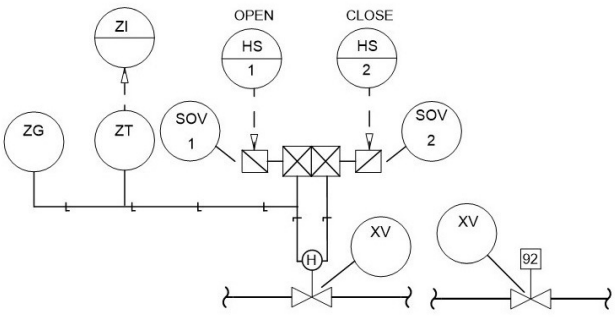


Figure 1: An example of a typical instrument assembly legend sheet. The core challenge is automatically distinguishing the detailed assembly from the simplified representation without any explicit visual separators between them.

2 Related Work

In this section, we review the existing research in blueprint information extraction, visual prompt engineering, as well as in-context and few-shot learning. Importantly, we also comment on the limitations of these methods in terms of automating information extraction from blueprint legend sheets.

2.1 Information extraction in engineering blueprints

Research on information extraction from engineering drawings primarily relies on Optical Character Recognition (OCR) tools and handcrafted heuristic rules [5, 8, 16, 20]. This typically involves using a text detector like EAST [30] followed by a recognizer like Tesseract [21]. The extracted text is utilized via proximity-based heuristics, such as Euclidean distance [3]. However, these rule-based methods are brittle as they often fail when faced with dense layouts or inconsistent spacing, which are common in real-world diagrams. Their fundamental limitation is a lack of contextual and semantic understanding, making them unsuitable for reliably parsing the conceptual regions found in engineering legends.

2.2 Visual Prompt Engineering

Several studies have explored visual prompt engineering techniques to guide VLMs. The authors of [17] modify the query images by manually inserting a red circle as a visual prompt to influence VLMs to analyze specific regions. However, this approach can introduce occlusions and is not practical, as it requires the manual annotation of every query image. To automate the creation of visual prompts, other methods leverage recognition models. For example, Contrastive Region Guidance [24] uses object detectors, while Fine-Grained Visual Prompting [27] employs the Segment Anything Model (SAM) [9] to generate segmentation masks. Similarly, the Attention Prompting on Image (API) method [28] uses CLIP [15] to locate objects based on a text query. Another example is the WorldAfford framework [2], which first uses an LLM with Chain-of-Thought prompting to reason about a task, and then explicitly employs SAM and CLIP as an intermediate step to localize relevant objects in a scene before final analysis. The fundamental limitation of these methods for our domain is their reliance on object-centric recognition. Models like SAM are trained to segment

visually distinct objects (e.g., a specific shape or entity). However, engineering legends require identifying concepts. A "detailed assembly" is not a single object, but a loose conceptual grouping of symbols and text with no clear visual boundary separating it from the adjacent "simplified" representation. As our results will show, object-centric methods struggle to identify these abstract groupings without costly, task-specific fine-tuning. This highlights the need for an end-to-end prompting strategy that guides the VLM directly without relying on a brittle, intermediate object recognition step.

More recently, Chain-of-Region (CoR) preprocesses diagrams by identifying visual regions using traditional computer vision tools and rule-based heuristics [11]. CoR then incorporates the extracted metadata into a VLM prompt for fine-grained analysis. However, this methodology is fundamentally unscalable for our domain because it relies entirely on brittle pre-processing. While CoR explicitly targets the "structured nature of scientific diagrams," our industrial legends are characterized by ambiguous spatial boundaries and high layout variance. This means CoR requires costly, per-project customization of its internal rules (e.g., for shape detection and region merging), a brittleness that our prompt-based ICMAP strategy is designed to circumvent.

2.3 In-Context and Few-Shot Learning

In-context learning (ICL), which allows models to learn from a few examples provided directly in the prompt, has shown great promise in vision-language settings. Several models have been designed for specific in-context vision tasks. For instance, SegGPT [25] performs in-context segmentation, DINOv [10] uses visual prompts like scribbles for referring segmentation, and IMProv [26] leverages in-context examples for image inpainting. Despite their advancements, these methods are highly specialized for tasks like segmentation or inpainting. They do not directly address the challenge of structured information extraction from spatially ambiguous regions. Furthermore, they primarily focus on visual context and do not explore the structured interplay between annotated visual examples and referential textual instructions. While we have employed this structured prompting strategy to evaluate symbol recognition capabilities [18] and to enable end-to-end engineering drawing digitization [19], those studies prioritized domain-specific application outcomes over a systematic interrogation of the prompting mechanism itself. Consequently, they did not analyze the VLM component in isolation, perform ablation studies, or compare it against other VLM-based strategies. This work provides this missing analysis, formalizing the multimodal prompting strategy and validating it as a robust, data-efficient alternative to fine-tuning for parsing spatially ambiguous technical documents.

3 Methodology

In this section, we first describe the research context and the industrial problem. We then introduce the ICICLE proof-of-concept system and its core technical framework, the ICMAP prompting strategy. Finally, we detail the comparison methods, the industrial dataset, and the experimental setup used for our evaluation.

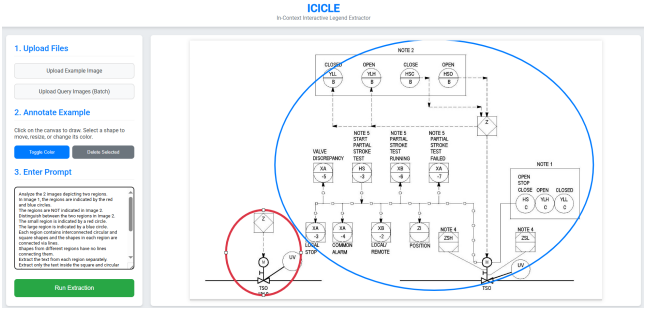


Figure 2: The user interface of the ICICLE Proof-of-Concept Software.

3.1 Research Context and Industrial Problem

The industrial problem is the manual, error-prone, and unscalable process of extracting structured information from instrument legend sheets. In large-scale EPC projects, these diagrams are critical engineering artifacts, and this manual process represents a significant bottleneck, leading to increased labor costs and potential project delays. The challenge is to develop an automated, accurate, and adaptable solution that handles the high diagram variability across different projects without requiring costly, per-project model retraining or the creation of brittle, rule-based parsers.

This study was conducted in McDermott, an international Engineering, Procurement, and Construction (EPC) company, employing an Action Research (AR) methodology [4]. This approach is ideal for applied problems, as it involves an iterative cycle where researchers and practitioners co-develop solutions to real-world challenges. Our AR process involved diagnosing the manual analysis of legend sheets as a key industrial bottleneck, developing ICICLE as a solution, and evaluating its impact using proprietary company data. While this deep immersion provides access to authentic problems, we acknowledge the potential for researcher bias [23], which we mitigate through rigorous quantitative evaluation against objective baselines.

3.2 The ICICLE System and the ICMAP Strategy

We developed ICICLE (In-Context Interactive Legend Extractor), a proof-of-concept intelligent system that provides engineers with a user-friendly interface using Python and the Flask web framework, shown in Figure 2. The system allows engineers to upload legend sheets, select an in-context example image, and perform annotations directly within the interface. Once the annotation is complete, the engineer can write the textual instruction. The system then automatically extracts structured information from all query legends and outputs the structured results as an Excel file for review.

The technical core employed by ICICLE is a multimodal prompting strategy we call In-Context Multimodal Annotation Prompting (ICMAP). ICMAP creates an explicit link between visual markers on an in-context example image and a textual instruction, which references these markers. This strategy enables a VLM to learn the visual characteristics of a conceptual region in a one-shot manner without fine-tuning. By referencing the annotated example, the VLM can ground abstract concepts (e.g., "detailed region") in

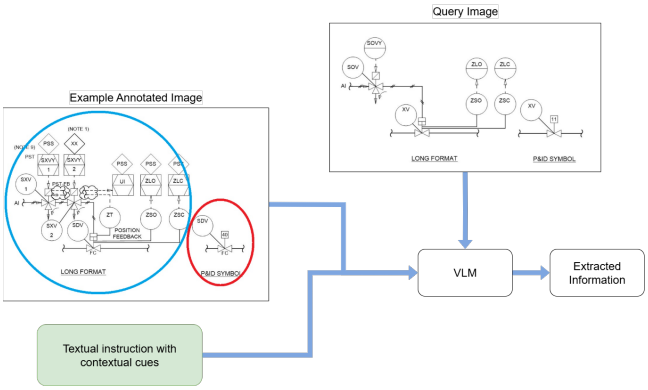


Figure 3: Visualization of the proposed In-Context Multimodal Annotation Prompting (ICMAP) strategy, which powers the ICICLE system.

specific visual patterns, allowing the ICICLE system to accurately parse new, unannotated legend sheets without modification or fine-tuning. The process, as illustrated in Figure 3, is composed of three key elements:

- **Query Image**: The unmodified legend sheet from which information needs to be extracted.
- **Annotated Example**: A separate example image where regions of interest (e.g., "detailed assembly", "simplified assembly") are highlighted with distinct visual markers, such as colored circles.
- **Referential Textual Prompt**: Instructions that direct the VLM's attention to the annotated example to learn the visual characteristics of each region. It then instructs the VLM to apply this learned knowledge to identify and extract information from the corresponding regions in the query image.

The PoC system was made available to a select group of engineers to validate feasibility and workflow integration. The PoC provides a tangible bridge between research and industrial practice by illustrating how prompting can be integrated into existing engineering workflows, providing a scalable and user-friendly approach to reducing manual effort.

3.3 Baseline Methods for Comparison

To demonstrate the engineering value of our approach, we evaluated several alternative methods:

- **OCR-Based Heuristic**: A rule-based baseline that uses an OCR engine to extract all text and their coordinates. It then applies a heuristic identifying the largest contiguous white space. The midpoint of this gap is then used as the vertical split point to group the tags. We evaluate a generalized heuristic, as relying on layout-specific parameter tuning hinders the scalability required for diverse industrial projects.
- **Text-Only Prompting**: A simple textual instruction asking the VLM to identify the two regions and extract their content. This baseline measures the model's out-of-the-box capability without visual guidance.

Table 1: The per-project distribution of the legend sheets.

	Project A	Project B	Project C	Project D	Total
Legends	10	9	9	10	38

- **Manual Annotation**: A zero-shot VLM technique [17] that involves manually drawing colored circles directly on every single query image. While effective, its primary drawback is its complete lack of scalability for any practical application.
- **Recognition-Model-Based Prompting**: We use external recognition models to automatically generate visual prompts. We investigate a Segmentation-Based approach where SAM [9, 27] produces hard binary masks based on object structure, and an Attention-Based approach (API) [15, 28] where CLIP generates soft attention heatmaps by locating regions relevant to a text query. Both SAM and CLIP operate in a zero-shot manner.

3.4 Experimental Setup

3.4.1 VLM Configuration. All experiments were conducted within a secure Microsoft Azure environment to comply with the strict confidentiality requirements of our industrial data. Consequently, all evaluations exclusively utilized the Azure OpenAI API with the GPT-4o model [7]. The image detail parameter is set to "high" and the temperature is set to 0 to promote deterministic outputs. However, previous studies indicate that LLMs and VLMs can still produce variable results even with the temperature set to 0 [6]. Therefore, all experiments involving the VLM are conducted three times, and the most common output is preserved. This practice ensures the reliability of our findings [1, 6].

3.4.2 Evaluation Data. The evaluation was conducted on a proprietary dataset, consisting of 38 instrument assembly legend sheets from four distinct industrial projects (referred to as A, B, C, and D) executed by McDermott. While the total volume is constrained by intellectual property restrictions, these documents represent the complete set of distinct layouts used in four major capital projects, ranging from offshore green energy infrastructure to onshore hydrocarbon complexes. The distribution of legends is shown in Table 1. The confidential nature of this real-world dataset ensures that the VLM has not been trained on it, providing a robust test of generalization. This dataset was specifically chosen for its diversity and challenges:

- **Layout Variance**: High variation in the arrangement and density of components across projects.
- **Spatial Ambiguity**: Minimal separation between the "detailed" and "simplified" assembly regions.

Examples of these challenges are shown in Figure 4 for projects A and D. As can be seen, the positions of the "detailed" and "simplified" assemblies vary between projects. Furthermore, there is limited spacing between the assembly regions. Moreover, the complexity in terms of the number of components varies.

3.4.3 Evaluation Metrics. The evaluation metric is Accuracy (Equation 1), defined to measure the correct grouping of texts inside the shapes (tags) to their corresponding assembly region ('detailed' or

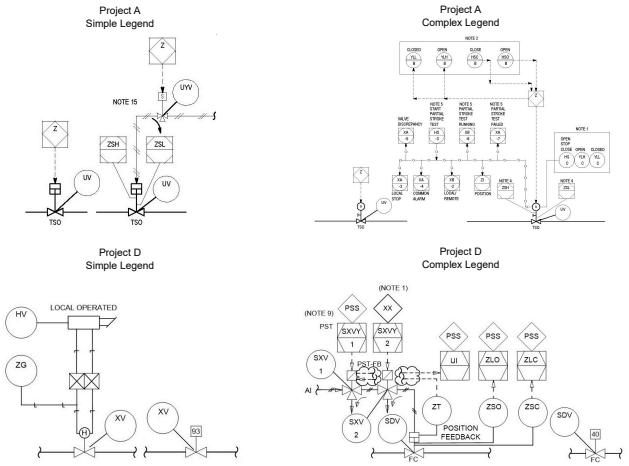


Figure 4: Examples of simple and complex legend images from two projects. In Project A, the right region corresponds to the detailed assembly, while in Project D, the left region corresponds to the detailed assembly.

'simplified'). To ensure this metric adequately evaluates recognition quality, we enforce a strict criterion: a tag is counted as correct only if the optical character recognition yields an exact string match with the ground truth.

$$\text{Accuracy} = \frac{\text{Correctly Extracted and Grouped Tags}}{\text{Total Number of Ground Truth Tags}} \quad (1)$$

To serve as the absolute baseline for comparison, a ground truth was established through manual annotation and verification by domain experts at McDermott. All automated methods are evaluated based on their ability to reproduce this expert consensus, which represents the target 100% accuracy.

3.4.4 Robustness to Visual Example Choice. The effectiveness of visual in-context learning greatly depends on the choice of demonstration images, as shown in research by [22, 29]. Therefore, we evaluated the accuracy of ICMAP with different example images with different arrangements and complexities. Examples of simple and complex legend images are shown in Figure 4.

3.4.5 Ablation Studies. To isolate the contribution of each key component of ICMAP, we evaluated its performance under two degraded conditions:

- **Without Contextual Cues**: The textual prompt was modified to remove explicit references to the visual markers, testing the impact of the visual annotations alone.
- **Without Visual Annotations**: The annotated example image was replaced with a standard few-shot example (a clean example image and its corresponding text answer), testing the impact of the visual markers themselves.

4 Results

This section presents the empirical results validating the ICICLE proof-of-concept system. We first compare the performance achieved by ICICLE against all baseline methods. We then present the findings from our robustness analysis, which is critical for any practical intelligent system. We conclude with the results of the ablation study, which deconstructs the key design principles of the ICMAP strategy that powers ICICLE.

4.1 Comparative Performance Analysis

We compare the performance of ICICLE with the different baseline methods. The performance of all methods was measured against the expert-verified ground truth as mentioned in Section 3.4.3. First, to validate the fundamental architectural decision to use a Vision-Language Model, we tested a classical computer vision approach. The OCR-Based Heuristic was evaluated on Project A and achieved only 24.1% accuracy. The heuristic rules were unable to cope with the minimal spatial separation between the assembly regions. Although specific constraints could be engineered to solve a specific format, such manual tuning is unsustainable across varying layouts. Consequently, heuristic approaches are not viable. Next, we conducted a comparison of different VLM prompting strategies across all four industrial projects, with the results shown in Figure 5. An example prompt and the output for each VLM-based method are also illustrated in Figure 6. The results clearly demonstrate that ICICLE's methodology is superior:

- **Performance of the ICICLE System**: Powered by ICMAP, our system achieves near-perfect results, reaching 100% accuracy in Projects A, B, and D. In Project C, it achieved 96% with a complex example and 100% with a simple one, demonstrating remarkable effectiveness and readiness for real-world industrial data.
- **Failure of General-Purpose VLM Baselines**: The results starkly highlight why a specialized strategy is necessary. The text-only prompt, representing the VLM's raw capability, performed poorly, confirming that a simple software wrapper around a VLM API is insufficient. Similarly, methods relying on external recognition models also failed. This confirms that these models cannot generalize to the abstract structures in technical diagrams, as they identify visually salient elements but fail to recognize the conceptual "assembly" of those elements. This failure provides the core engineering justification for our approach.
- **Comparison to Manual Annotation**: Interestingly, ICICLE also outperforms the manual annotation of the query image in several cases. We attribute this to the fact that manual annotation can sometimes occlude key details on the query image itself, a design flaw ICICLE avoids by annotating a separate example. Furthermore, ICICLE's one-shot automation is inherently more scalable.

4.2 Robustness to Example Choice

A key requirement for a practical intelligent system like ICICLE is its robustness to variations in user input. The results in Figure 5 also serve as our robustness analysis. As shown, the system's performance remains highly stable regardless of whether a simple or a complex legend is used as the annotated example. In three of the four projects, the accuracy is identical (100%). In Project C, the drop from 100% to 96% with the complex example is minor. This

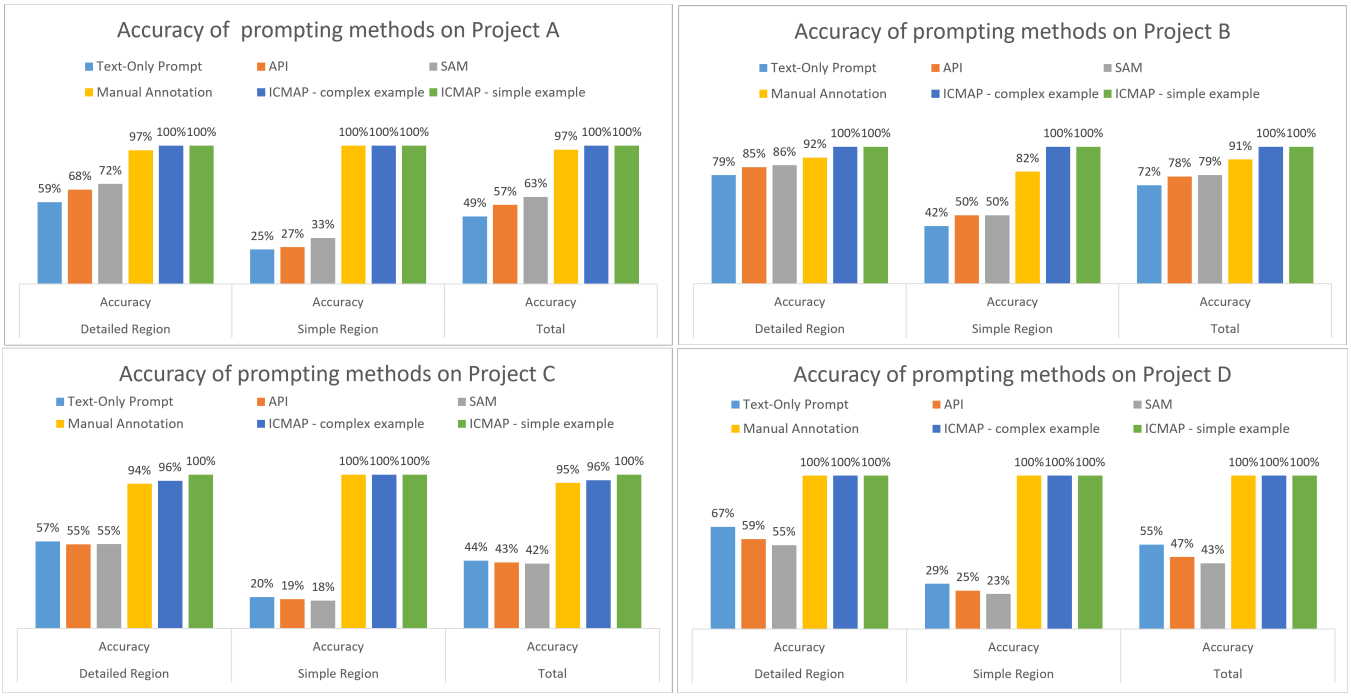
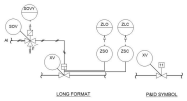
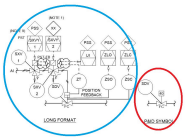
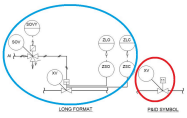


Figure 5: Results of the prompting methods on the evaluation projects.

Method	Query Image	Prompt	Extracted Instruments (Errors in red)
ICMAP		 Analyze the 2 images depicting two regions. In Image 1, the regions are indicated by the red and blue circles. The regions are NOT indicated in Image 2. Distinguish between the two regions in Image 2. The small region is indicated by a red circle. The large region is indicated by a blue circle. Each region contains interconnected circular and square shapes and the shapes in each region are connected via lines. Shapes from different regions have no lines connecting them. Extract the text from each region separately. Extract only the text inside the square and circular shapes. Do NOT exclude repeating texts. Provide the answer for Image 2 without colored annotations. Use the following output format: **Small (Red) Region:** - ... **Large (Blue) Region:** - ...	**Region 1:** - SOVY - SOV - XV - ZLO - ZLC - ZSO - ZSC **Region 2:** - XV
Manual Annotation		Analyze the image depicting two regions depicted in red and blue circles. Extract the text from each region separately. Extract only the text inside the square and circular shapes. Do NOT exclude repeating texts. Use the following output format: **Region 1:** - ... **Region 2:** - ...	

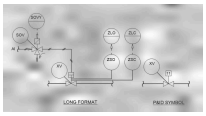
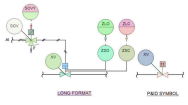
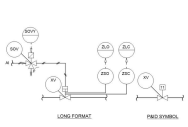
Method	Query Image	Prompt	Extracted Instruments (Errors in red)
API	 (Mask generated by CLIP via prompt: "two assemblies of instrument devices")	Analyze the image depicting two regions. Find the regions. The first region on the left of the image is more complex and larger. The second region on the right is typically smaller. Each region contains interconnected circular and square shapes and the shapes in each region are connected via lines. Shapes from different region have no lines connecting them. Extract the text from each region separately. Extract only the text inside the square and circular shapes. Do NOT exclude repeating texts. Use the following output format: **Region 1:** - ... **Region 2:** - ...	**Region 1:** - SOVY - SOV - XV **Region 2:** - ZLO - ZLC - ZSO - ZSC - XV
SAM	 (Mask generated by SAM in a zero-shot manner)		
Text-Only Prompt			

Figure 6: An example of a prompt and result for each investigated method.

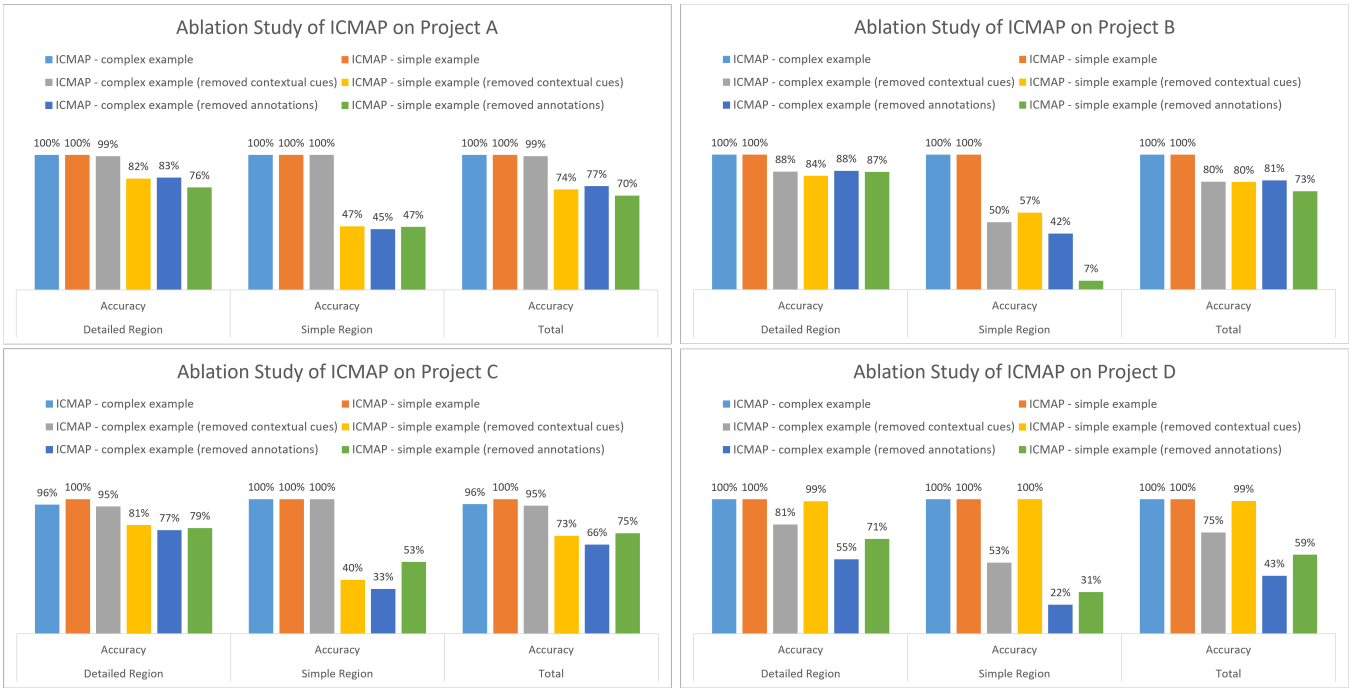


Figure 7: Ablation study of ICMAP on the evaluation projects.

result provides strong evidence that the ICMAP strategy is not reliant on a specific, perfectly chosen example to function effectively, highlighting its robustness for practical deployment.

4.3 Ablation Study: Deconstructing ICMAP

We performed an ablation study to validate that the interplay between visual annotations and referential text is critical for ICMAP's success. The results are presented in Figure 7.

- **Removing Contextual Cues**: When the textual references to the annotations were removed, performance dropped significantly. Furthermore, the model's accuracy became dependent on the choice of the example image. This demonstrates that the visual markers alone are insufficient; the VLM requires the textual instruction to reliably interpret the semantic meaning of the colored regions.
- **Removing Visual Annotations**: Conversely, when the visual annotations were removed from the example image and replaced with a text example answer, performance also degraded. This approach was prone to "hallucinating" content from the example answer into the final output.

These results provide evidence that the coupling of visual markers and referential text enables the reliability of ICICLE. Removing either component leads to a substantial loss in performance.

5 Discussion

The results validate that our ICICLE proof-of-concept is a highly effective intelligent system for its industrial domain. The near-perfect extraction accuracy is significant. The failure of baseline methods confirms that a specialized approach is necessary. The ICICLE system serves as a blueprint to guide VLMs to achieve practical automation.

A crucial takeaway is the context of our VLM choice. Due to corporate data policies, all experiments used the powerful GPT-4o model. Within this controlled setting, the failure of baselines demonstrates that the VLM's raw capability alone is insufficient. The significant performance delta highlights that the design of the prompting mechanism is the primary driver of success, a key lesson that thoughtful system design is critical.

5.1 Driving Digital Transformation through Efficiency Gains

The primary value of ICICLE lies in driving digital transformation through tangible cost and efficiency gains:

- **Reduction of Manual Effort:** Automating the analysis of legend sheets eliminates substantial labor costs, directly freeing skilled engineers for higher-value tasks.
- **Elimination of Model Fine-Tuning:** By relying on in-context learning, ICMAP avoids the expensive data collection and retraining cycles required by traditional AI solutions.
- **Streamlined Integration:** The system improves adoption by outputting data in standard formats (e.g., Excel), allowing for seamless integration into existing engineering processes.

5.2 Design Principles for Intelligent Systems

The development of ICICLE yielded several best practices:

- **Avoid Costly Specialization:** Methods relying on external recognition models fail to identify domain-specific conceptual regions without extensive tuning. Since model training

for every unique legend type is unscalable, our approach that bypasses this fragile pre-processing step offers a more robust and cost-effective architecture.

- **Adopt Model-Agnostic Patterns:** As a prompting pattern, ICMAP allows the underlying VLM to be swapped or upgraded without re-architecting the solution. This flexibility is critical for long-term maintenance.
- **Leverage Multimodal Synergy:** Success relies on the referential coupling of visual annotations and textual instructions. Intelligent systems should guide the VLM by linking modalities rather than simply providing unstructured context.
- **Prioritize High-Performance VLMs:** While ICMAP is model-agnostic, high-capability models are a prerequisite for industrial-grade accuracy.

6 Conclusion

This paper presented ICICLE, an intelligent Generative AI system designed to automate information extraction from ambiguous engineering diagram legends. Central to our approach is In-Context Multimodal Annotation Prompting (ICMAP), which enables VLMs to parse complex documents using a single annotated example, thereby eliminating retraining costs. Evaluation on industrial datasets demonstrated 96–100% accuracy, significantly outperforming OCR heuristics and standard VLM prompting. Our ablation studies further confirmed that coupling visual annotations with textual instructions is critical for success. Ultimately, ICICLE serves as a blueprint for operationalizing Generative AI in engineering, proving that intelligent systems can be accurate, scalable, and adaptable. Future work will pursue two directions: evaluating the ICMAP pattern on diverse VLMs to confirm its adaptability, and adapting ICICLE for other engineering tasks.

Acknowledgments

This work is supported by McDermott, Software Center (Gothenburg, Sweden), and Eindhoven University of Technology.

References

[1] Alexander Bendeck and John Stasko. 2024. An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[2] Changmao Chen, Yuren Cong, and Zhen Kan. 2024. Worldafford:Affordance Grounding Based on Natural Language Instructions. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*. 822–828. doi:10.1109/ICTAI62512.2024.00120

[3] Rimma Dzhusupova, Mina Ya-alimadad, Vasil Shteriyanov, Jan Bosch, and Helena Holmström Olsson. 2024. Practical Software Development: Leveraging AI for Precise Cost Estimation in Lump-Sum EPC Projects. In *2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. 1023–1033. doi:10.1109/SANER60148.2024.00110

[4] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. Selecting empirical methods for software engineering research. *Guide to advanced empirical software engineering* (2008), 285–311.

[5] Mathieu Francois, Véronique Eglin, and Maxime Biou. 2022. Text detection and post-OCR correction in engineering documents. In *International Workshop on Document Analysis Systems*. Springer, 726–740.

[6] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14375–14385.

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[8] Laura Jamieson, Carlos Francisco Moreno-Garcia, and Eyad Elyan. 2020. Deep learning for text detection and recognition in complex engineering diagrams. In *2020 International joint conference on neural networks (IJCNN)*. IEEE, 1–7.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.

[10] Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, et al. 2024. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12861–12871.

[11] Xue Li, Yiyou Sun, Wei Cheng, Yinglun Zhu, and Haifeng Chen. 2025. Chain-of-region: Visual language models need details for diagram analysis. In *The Thirteenth International Conference on Learning Representations*.

[12] Shouvik Mani, Michael A Haddad, Dan Constantini, Willy Douhard, Qiwei Li, and Louis Poirier. 2020. Automatic digitization of engineering diagrams using deep learning and graph search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 176–177.

[13] Shubham Paliwal, Arushi Jain, Monika Sharma, and Lovekesh Vig. 2021. Digitize-PID: Automatic digitization of piping and instrumentation diagrams. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25*. Springer, 168–180.

[14] Process Industry Practices. 2023. *PIP PIC001 - Piping and Instrumentation Diagram Documentation Criteria - Technical Correction: 6/2023*. Technical Report. Process Industry Practices (PIP).

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[16] Amir Saba, Rim Hantach, and Mohammed Benslimane. 2023. Text Detection and Recognition from Piping and Instrumentation Diagrams. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 711–716.

[17] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11987–11997.

[18] Vasil Shteriyanov, Rimma Dzhusupova, Jan Bosch, and Helena Holmström Olsson. 2025. BlueprintSymVL: A discriminative benchmark for VLM symbol recognition in engineering blueprints. *Results in Engineering* 28 (2025), 108171. doi:10.1016/j.rineng.2025.108171

[19] Vasil Shteriyanov, Rimma Dzhusupova, Jan Bosch, and Helena Holmström Olsson. 2025. Enhancing OCR-based Engineering Diagram Analysis by Integrating Diverse External Legends with VLMs. *Journal of Software: Evolution and Process* 37, 12 (2025), e70072. doi:10.1002/smr.70072

[20] Vasil Shteriyanov, Rimma Dzhusupova, Jan Bosch, and Helena Holmström Olsson. 2026. From text to meaning: Semantic interpretation of non-standardized metadata in piping and instrumentation diagrams. *Computers & Chemical Engineering* 204 (2026), 109436. doi:10.1016/j.compchemeng.2025.109436

[21] R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Vol. 2. 629–633. doi:10.1109/ICDAR.2007.4376991

[22] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. 2025. Exploring Effective Factors for Improving Visual In-Context Learning. *IEEE Transactions on Image Processing* (2025).

[23] Geoff Walsham. 1995. Interpretive case studies in IS research: nature and method. *European Journal of information systems* 4, 2 (1995), 74–81.

[24] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. 2024. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*. Springer, 198–215.

[25] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. 2023. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1130–1140.

[26] Jiarui Xu, Yossi Gandelsman, Amir Bar, Jianwei Yang, Jianfeng Gao, Trevor Darrell, and Xiaolong Wang. 2023. Improv: Inpainting-based multimodal prompting for computer vision tasks. *arXiv preprint arXiv:2312.01771* (2023).

[27] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023. Fine-grained visual prompting. *Advances in Neural Information Processing Systems* 36 (2023), 24993–25006.

[28] Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*. Springer, 251–268.

[29] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems* 36 (2023), 17773–17794.

[30] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5551–5560.