Enhancing OCR-based Engineering Diagram Analysis by Integrating Diverse External Legends with VLMs

Journal of Software: Evolution and Process

WILEY

Software: Evolution and Process

RESEARCH ARTICLE - TECHNOLOGY OPEN ACCESS

Enhancing OCR-based Engineering Diagram Analysis by Integrating Diverse External Legends with VLMs

Vasil Shteriyanov[1,2] | Rimma Dzhusupova[1,2] | Jan Bosch[2,3] | Helena Holmström Olsson[4]

[1]Engineering, McDermott, The Hague, the Netherlands | [2]Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Netherlands | [3]Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden | [4]Computer Science and Media Technology, Malmö University, Malmö, Sweden

Correspondence: Vasil Shteriyanov (vasil.shteriyanov@mcdermott.com)

Received: 28 April 2025 | **Revised:** 17 September 2025 | **Accepted:** 6 November 2025

Keywords: diagrams | information extraction | legends | multimodal prompt engineering | optical character recognition | vision language models

ABSTRACT

Manual analysis of diagrams and legend sheets in engineering projects is time consuming and needs automation. The lack of standardized legend formats complicates creating a general method for automated information extraction. Existing approaches require training and custom rules for each project. This study proposes a novel solution combining optical character recognition with vision language models and multimodal prompt engineering to automate information extraction from diverse legend sheets without training. It integrates legend information with information extracted from diagrams, unlike studies that only focus on diagrams. Our study shows that VLMs, guided by multimodal prompts, can accurately extract information from diverse legend sheets, enabling automatic information extraction in diagrams across engineering projects. We validate our method through a case study involving the extraction of instruments from piping and instrumentation diagrams (P&IDs) and their legends across three projects with varied formats and standards. The proposed method achieved 100% accuracy in legend classification and information extraction, and 99.68% precision and 95.91% recall in generating instrument listings. The results demonstrate the effectiveness of our approach, significantly enhancing the accuracy and efficiency of information extraction from diagrams. This method can be adapted to different legend formats and diagrams, providing a versatile solution for various industries.

1 | Introduction

Many industries, including engineering, construction, and manufacturing, rely on complex technical diagrams such as diagrams, schematics, and blueprints. These documents serve as essential references for designing, constructing, and maintaining industrial systems. However, these diagrams often employ simplified representations of equipment to enhance readability, omitting critical details about full equipment assemblies. As a result, engineers must reference external legend sheets that map simplified symbols to their detailed counterparts to fully interpret these diagrams [1]. This process is critical during the early phases of projects, such as feasibility studies, tendering, and design validation, where rapid and accurate analysis of diagrams is crucial. An example legend of a piping equipment assembly is shown in Figure 1. It highlights the discrepancy between the simplified representation found in diagrams and the detailed assembly described in the legend sheet.

This reliance on external legend sheets creates a significant bottleneck in diagram analysis. During the tender phase of engineering projects, engineers must manually cross-reference legend sheets with the diagrams to analyze the diagrams. Such manual processes have been reported to be time consuming and error prone [2]. The manual nature of this process introduces inefficiencies, delays, and the risk of misinterpretation, particularly in large-scale projects involving thousands of diagrams. This challenge is not limited to a single industry.

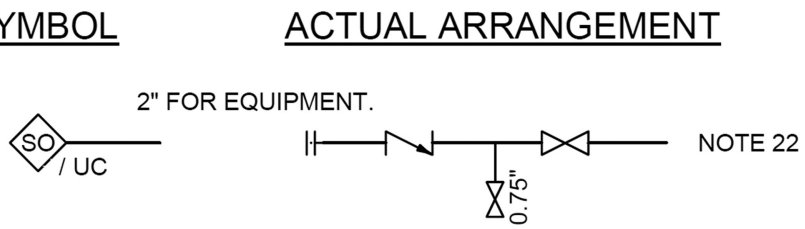


FIGURE 1 | Example legend of a piping equipment assembly showing the simplified representation in a diagram and its corresponding detailed assembly.

Similar workflows exist in various domains where technical diagrams and external reference documents must be cross-verified, such as electrical schematics in power systems, avionics blueprints in aerospace, and mechanical diagrams in manufacturing.

The problem is further compounded by the lack of standardization in legend sheet formats, which can vary significantly across projects and industries [3]. Unlike diagrams, which often follow industry standards (e.g., IEC or ANSI standards [4, 5]), legend sheets do not have universally accepted formats, making automation of information extraction particularly difficult. This variability requires any automated system to be adaptable to diverse legend structures without requiring extensive retraining or manual customization.

Existing research on information extraction from diagrams has predominantly focused on unimodal object recognition methods [2, 6–11, 36]. Attempting to integrate legend information using traditional AI approaches would require extensive training on multiple legend formats or custom heuristic rules for each project. Given the high variability in legend formatting, such an approach is neither scalable nor practical for real-world industrial applications.

Recently, vision language models (VLMs) have demonstrated significant potential in both extracting information from images and understanding the contextual structure within them. These models can be prompted using textual instructions and images, which enables them to be tailored to a wide range of application scenarios [12]. This capability could be particularly beneficial in industries dealing with diagrams, where legend sheets can vary significantly from project to project. Thus, VLMs could provide a flexible and adaptable solution in the case of legend sheet information extraction.

Another limitation of existing research in digitalizing diagrams focuses on extracting text and symbols exactly as they are depicted on these diagrams [2, 6–11, 13, 36]. However, as previously stated, diagrams often show only a subset of the actual assemblies and materials. Thus, to enable comprehensive digitalization, an effective method must go beyond simple extraction and incorporate legend-based contextual understanding, allowing for the automatic reconstruction of full equipment assemblies.

In our earlier work [14], we demonstrated that AI could extract assembly information by correlating simplified diagram representations with their detailed counterparts in legend sheets. However, our previous approach relied on a fixed legend format, requiring manual adjustments for different projects. This constraint limited its scalability and applicability in diverse engineering documents.

Based on the identified research gaps, this research is guided by the following research questions:

- *How can information be extracted from legend sheets following diverse formatting and standards in a training-free manner?*
- *How can legend information be incorporated into diagram information extraction?*

This manuscript extends our previous research [14] by developing a generalized method for integrating diverse legend information into engineering diagram analysis. Unlike our prior approach, the new method does not require the implementation of specific rules for customization to different projects. Instead, it leverages VLMs, multimodal prompt engineering, and in-context learning to dynamically adapt to varying legend formats. This novel approach addresses the fundamental challenge of legend sheet formatting variability. Furthermore, the method uses traditional optical character recognition (OCR) methods to extract information from the diagrams. This hybrid approach ensures high adaptability while maintaining precision in engineering applications. Importantly, our method is designed to be industry-agnostic, making it applicable to a wide range of domains where complex diagrams and external legends must be interpreted together.

The main contribution of the research is the following:

We introduce the first method for integrating diverse legend information into automated diagram information extraction, combining traditional OCR techniques and VLMs. This novel approach enables information extraction and integration of any legend format without requiring training. The method enables the extraction of components not explicitly depicted in the diagrams. This contrasts with previous studies that only extract information from diagrams.

We validate our method using a case study. We focus on extracting a listing of instruments from typical instrument assemblies in piping and instrumentation diagrams (P&IDs) and their legend sheets. The P&IDs show a simple representation, showing only a subset of the instruments in the assemblies, while the legends show both the simple representation and the detailed assembly. We evaluate our method on three different engineering projects following diverse legend formats and drawing standards. Importantly, the method achieves 100% information extraction

from the typical instrument assembly legends. Additionally, the instrument generation approach achieved an overall 95.91% recall and 99.68% precision.

Our method has the potential to benefit a wide range of industries reliant on technical diagrams for decision-making. It addresses the challenge of diagram information extraction in contexts where external reference documents, such as legend sheets, are required. The incorporation of external information in diagram information extraction is a concern wherever simplified representations of assemblies are used to improve the readability of diagrams. The proposed method contributes to the broader field of automated document understanding, offering a scalable solution for industries that require accurate and efficient interpretation of complex technical documents.

The remainder of this manuscript is structured as follows: Section 2 reviews current research on extracting information from industrial diagrams. Section 3 describes the research methodology. Section 4 introduces the proposed method to integrate diverse legend information with diagram data. Section 5 describes the validation case study. Section 6 presents the results. Section 7 discusses the results in terms of the prior work on the topic, the limitations of the study, and the directions for future work. Finally, Section 8 summarizes the contributions and potential impacts of our work.

2 | Related Work

Information extraction in industrial diagrams has traditionally relied on object recognition pipelines that combine deep learning models for symbol, text, and line detection with heuristics to associate the recognized objects [2, 6–10]. While effective at digitizing what is explicitly depicted on a diagram, these methods are fundamentally limited as they do not account for implicit information, such as detailed component assemblies that are only described in external legend sheets.

Recently, there has been a shift toward exploring multimodal models for information extraction from diagrams. Khan et al. [13] fine-tuned the Florence-2 VLM [15] to extract geometric dimensioning and tolerancing information from diagrams. The authors report achieving an F1 score of 61.51% in terms of extraction. Furthermore, Doris et al. [16] evaluate contemporary multimodal large language models (MLLMs), such as GPT-4o, on a benchmark that evaluates the MLLMs' ability to interpret engineering documents. The authors report that these models face challenges in recognizing technical components in CAD images and analyzing diagrams.

Nevertheless, a significant limitation of industrial diagram information extraction methods is that they do not incorporate information from the diagrams' legend sheets about simple and detailed representations of complex components. Sarkar et al. [11] proposed a method that references legend sheets to extract individual symbols from diagrams based on similarity matching. However, their method focuses solely on individual symbols and does not account for assemblies of elements or components represented through textual description. This limitation makes their method unsuitable for diagrams where components are represented in an abstract manner.

To advance research in this area, our previous study presented a method to extract complex components based on their simple diagram representation and their detailed legend representation for a single legend format [14]. Despite this progress, the diverse legend formatting standards employed by different engineering projects necessitate further research to generalize this method so that it can be easily customized to any legend format.

This review highlights several limitations in existing methods:

- Most diagram information extraction methods [2, 6–11, 13, 36] extract information exactly as it is depicted in the diagrams and do not account for implicit components represented only in reference documents, such as legend sheets.
- Although our previous study [14] proposed a method to capture implicit components targeting a single legend format, no existing work proposes an easily adaptable solution that addresses the formatting variability of the diagrams' legend sheets across projects.

To address the identified research gaps, our proposed method offers a novel solution to integrate diverse legend information into diagram extraction and enable the extraction of implicit components. It can easily be applied across different projects and standards. This advancement has the potential to significantly enhance the accuracy and efficiency of information extraction in diagrams, benefiting various industries that rely on these critical documents.

3 | Research Methodology

The research was executed at McDermott, an international Engineering, Procurement, and Construction (EPC) company undertaking construction projects in the energy sector. The study employed dominant Action Research as its methodology [17], which involves mixing the Action Research method [18] with another research method. Given that the researchers were members of the case company, we used action research to study common challenges engineers experience at McDermott and to propose a method to solve these challenges. Furthermore, we employed a Case Study [19] to validate our proposed method. Despite potential biases arising from the researchers' direct involvement in the development process [20], the AR methodology offers the substantial advantage of granting unique access to industry-specific data that would not be accessible to external researchers.

The research process is guided by the CRISP-DM framework [21]. The activities involved in the research method are shown in Figure 2. Initially, we identified the problem based on the practices and experiences of engineers within McDermott. Our action research revealed that the full lifecycle of producing the full instrument listing by referring to the legend sheets for a single P&ID, including multiple revisions, requires an average of three engineering hours, confirming it as a significant time-intensive undertaking. The identified problem of incorporating legend information in the diagram information extraction is presented in Figure 3. As can be seen, engineers manually analyze diagrams and their legends to produce documents listing

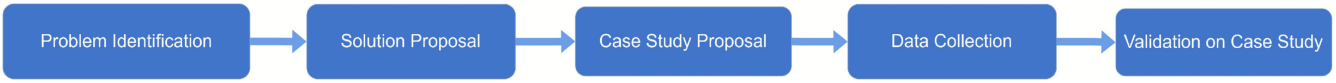


FIGURE 2 | Dominant action research activities [17].

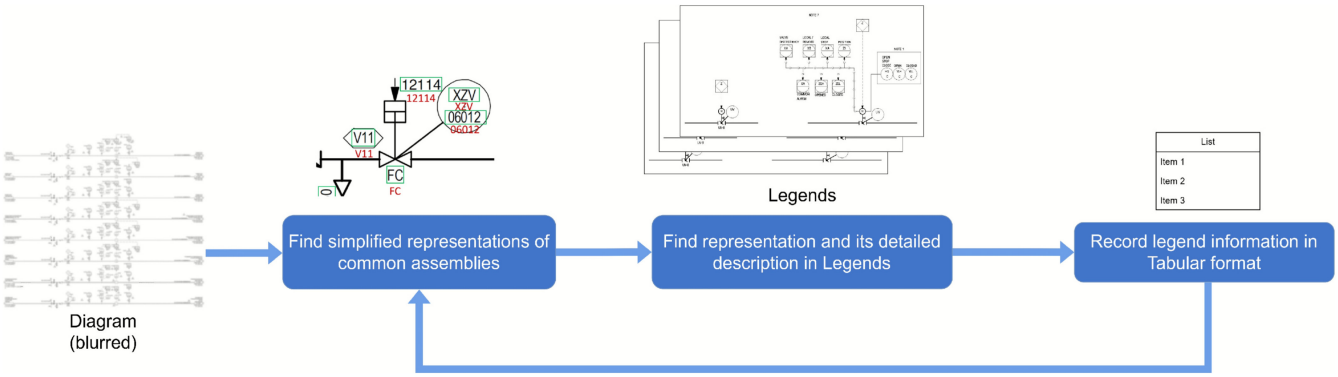


FIGURE 3 | Manual process of incorporating legend information in diagram analysis.

the equipment on the diagrams. This process can be time consuming and error prone whenever engineers need to analyze many diagrams. Next, we proposed a method for automatically extracting and integrating information from legends with diverse formats in the diagram information extraction process. In addition, we proposed a case study to validate our method based on an activity McDermott's engineers perform during the tender phase of EPC projects. The case involves the creation of instrument lists based on common instrument assemblies, which are represented by a simple representation on P&IDs, and their detailed representation is visualized in the P&IDs' legends. After the case study was proposed, we collected the case data. The training and evaluation data for the case study validation of the method was collected from past projects executed and owned by McDermott. The data consisted of P&IDs, as well as their legends, which are solely McDermott's intellectual property. None of the training data was used in the evaluation process. Finally, the proposed method was validated via the case study. A selected group of McDermott engineers were involved in the case study validation. Specifically, they evaluated the accuracy, precision, and recall of the method's extraction and integration of legend and diagram information. The proposed method, case study, data collection, and evaluation are explained in more detail in the following chapters.

The desired organizational change we aim to achieve with this research is to reduce the engineering hours required for the manual analysis of diagrams and legend sheets during the tender phase of EPC projects. The method could also reduce errors made by engineers. Compared with other methods, our method integrates legend information to extract components that are not explicitly depicted on the diagrams.

4 | Method for Enhanced Diagram Analysis via Training-Free Legend Extraction

This section presents our proposed general method for integrating legend information with diagram information extraction, given the variability of legend formats. The method is based on prior work on visual prompting and in-context learning, as well as information extraction from diagrams. Firstly, the processing of legends involves the use of a VLM and multimodal prompt engineering. Unlike traditional object recognition methods, VLMs have been shown to solve tasks without the need for fine-tuning via textual and visual prompting [12, 22, 23]. Additionally, this could also be achieved by providing the VLMs with in-context examples [24]. These abilities can be utilized to handle the variability of legend sheet formats without training specialized models and defining distance-based heuristic rules for each legend type. Furthermore, traditional computer vision methods, such as symbol recognition and text recognition, are employed to extract information from the diagrams. Moreover, existing diagram analysis methods have previously utilized heuristics to associate extracted information [2]. Thus, the extracted legend information can be integrated along with the information from the diagrams via heuristics to produce documents such as equipment lists. The overall approach is illustrated in Figure 4.

Initially, our automated solution for processing various legend sheets needs to determine the type of each legend. Thus, we explored the use of in-context learning with VLMs to identify the legend sheet type. This involves providing an image showing an example of each legend type, along with a textual instruction to determine the type of the query legend image based on the example. Depending on the identified legend type, an appropriate legend extraction method is selected for further execution.

Similarly to the legend identification method, the legend extraction method involves the use of a VLM along with a prompting method for information extraction. This method is inspired by prior work on visual in-context learning [24], where visual examples are provided to the VLMs to enable them to solve new tasks, as well as visual prompt engineering [23], where the queried image is modified to guide the attention of the model in the modified regions. In the context of this research, given a query legend, the method involves providing an example legend where

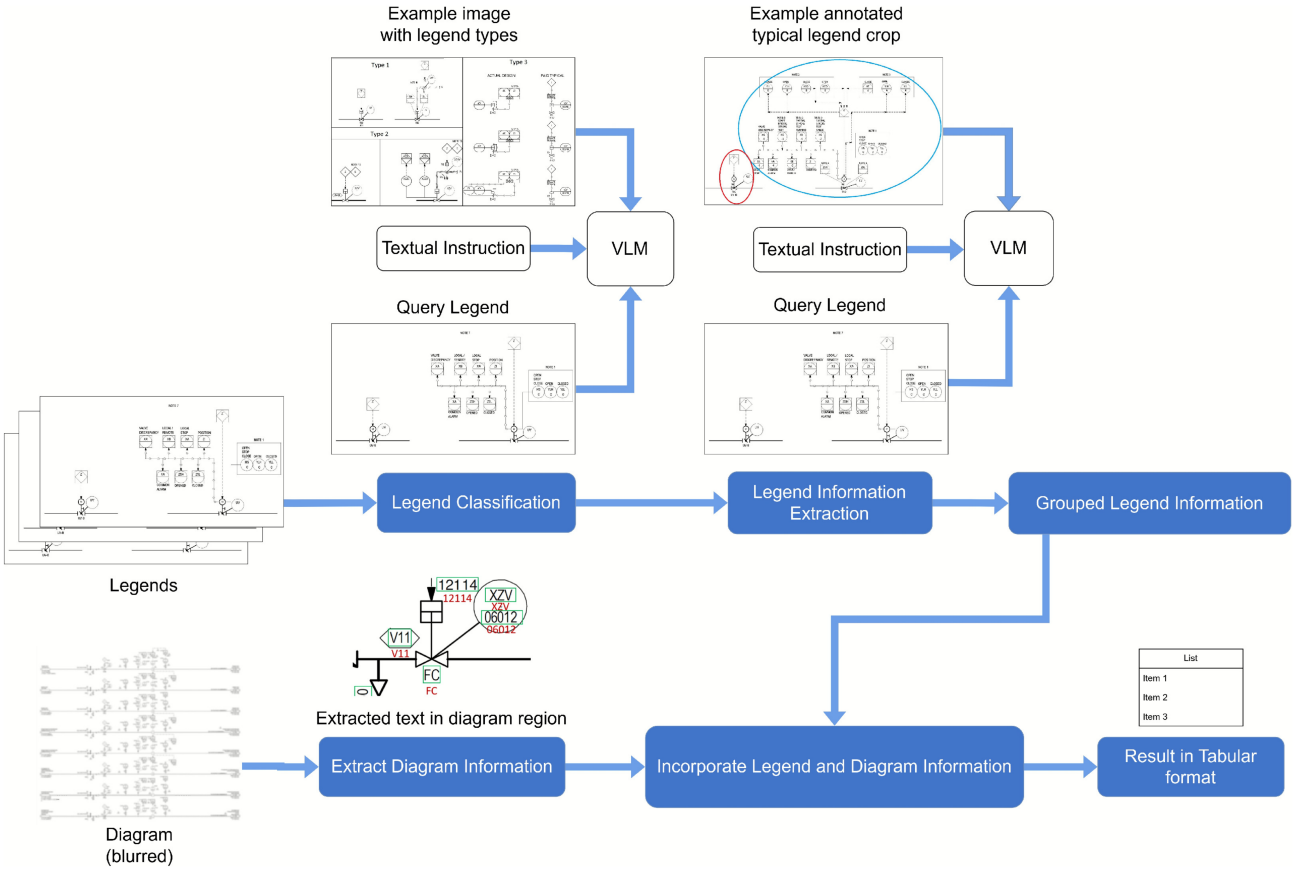


FIGURE 4 | Proposed method for seamless integration of diverse legend information in diagram information extraction.

regions of interest are annotated to guide the VLM to identify similar regions in the query legend. Furthermore, textual instructions are provided to guide the model on how to process each region. This approach allows for easy customization of various legend sheet types without the need for fine-tuning, unlike traditional computer vision methods.

A key advantage of this approach is its adaptability to new, previously unseen legend formats. The process for handling a new format is straightforward and does not require any model retraining. An engineer would simply need to

1. select a single, representative example of the new legend format;
2. manually annotate the example image by drawing a circle using a standard image editor to highlight the key information regions, if the layout is ambiguous;
3. adapt the textual portion of the prompt to describe the new layout and instruct the VLM on the desired extraction logic for that specific format.

This training-free customization process makes the system highly scalable and practical for real-world industrial environments where project-specific documentation is common, unlike traditional computer vision methods that would require extensive retraining.

Because diagrams are more standardized compared with legends, traditional computer vision extraction methods can be utilized to extract the text and symbols from the diagrams [2, 6, 14]. This extracted information can be used to expand the information in diagrams based on the information in the legend sheets. This would allow for the proper extraction of information, such as generating equipment lists.

5 | Validation Case Study: Extracting Instrument Listings From P&IDs

We validate our proposed method for integrating legend information across diverse legend formats in diagram information extraction via a case study. The case involves extracting a listing of instruments from typical instrument assemblies in P&IDs given diverse legend sheets. This section provides background information on the specific problem, a description of the application of our method to the problem, as well as the evaluation strategy and data collection.

5.1 | Background

A type of diagram heavily used in the EPC industry is the P&ID. P&IDs visualize the equipment and the instruments required to control processes in EPC projects [25]. When engineers analyze P&IDs, they often need to create documents listing the equipment in the diagrams. An example of this is the "Instrument Index" document, which lists all utilized instruments in the P&IDs [26]. These listing documents are used in the later stages of engineering project execution.

In the safety-critical context of EPC projects, generating documents like the "Instrument Index" is a significant challenge. The accepted quality standard for final documentation is effectively 100% accuracy, a stringent requirement met through a mandatory human verification and sign-off stage for any workflow. This practice of human oversight is essential for both manual and automated processes and aligns with emerging regulations like the EU AI Act [27]. While the initial drafting is traditionally performed manually, this task is prone to human oversight and highly inefficient. For instance, based on McDermott project execution statistics, our previous study reports that manually processing a single P&ID through its revision lifecycle averages three engineering hours [14]. Although academic literature lacks formal metrics on manual error rates, the value of automation is evident. Its primary benefit lies in streamlining the initial, labor-intensive generation, transforming the engineer's role from data entry to the more manageable task of verification.

The P&IDs often show typical assemblies of instruments via a simplified schematic. An example of a simplified instrument assembly representation in a P&ID is shown in Figure 5. It indicates a typical assembly of instrument devices via the typical number "UV-04" but only depicts the instrument type "XZV" with tag number "60107." Furthermore, the typical number "UV-04" acts as a reference to the actual instrument assembly, which is indicated in the legends of the EPC project. The legend for this typical assembly is shown in Figure 6, which shows the detailed assembly with all the utilized instrument equipment, alongside the simple schematic representation visualized on the P&ID. Thus, to generate the listing of instrument devices, engineers need to analyze the P&IDs and refer to their legend sheets. The process of listing all utilized instruments in the P&IDs is time consuming and error prone.

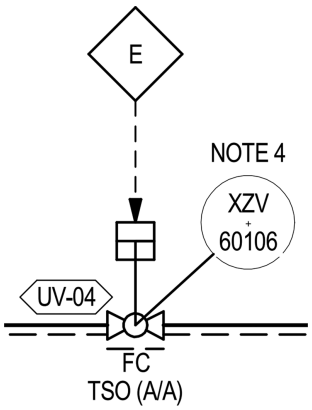


FIGURE 5 | Simplified instrument assembly representation in a piping and instrumentation diagram (P&ID).

A challenge to the automation of this process is that there is a lack of legend format standardization across EPC projects. For example, Figure 7 shows another legend sheet, which has different formatting and structure compared with the legend in Figure 6. An additional challenge is that some legend sheets do not indicate a clear separation between the simplified P&ID representation and its detailed representation. Figure 7 shows an example of this challenge, where some of the equipment of the detailed assembly is next to the instruments of the simplified diagram representation. Thus, it may not be trivial to create heuristic rules to extract and group the legend information separately for the simplified diagram representation and the detailed instrument assemblies.

5.2 | Automating Instrument Listings

The following section explains how the proposed method for incorporating diverse legend information in the diagram information extraction process, introduced in Section 4, was applied and customized to address the problem of generating a list of instruments in P&IDs despite the lack of format standardization in legends.

Specifically, we applied the method to 3 large EPC projects executed by McDermott, which we will refer to as A, B, and C. These projects were not chosen at random but were specifically selected because their legend formats represent a spectrum of key real-world formatting challenges:

- *Project A (the unstructured case)*: This format represents the most ambiguous challenge, where the simplified and detailed components lack any clear visual demarcation or consistent alignment.

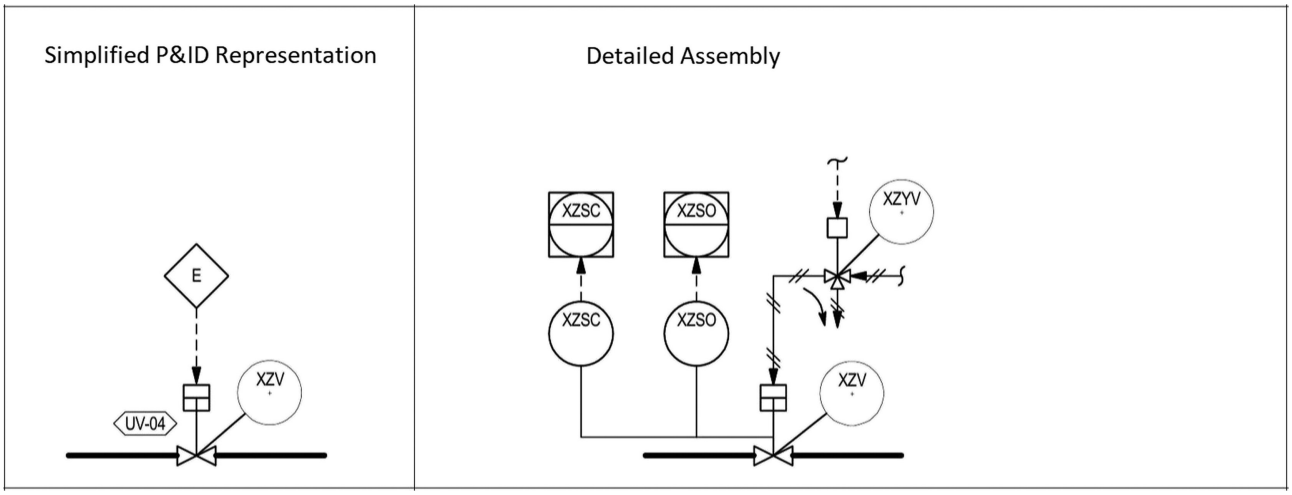


FIGURE 6 | Legend detailing the simplified representation of instrument assembly "UV-04" depicted in P&IDs and its corresponding detailed representation.

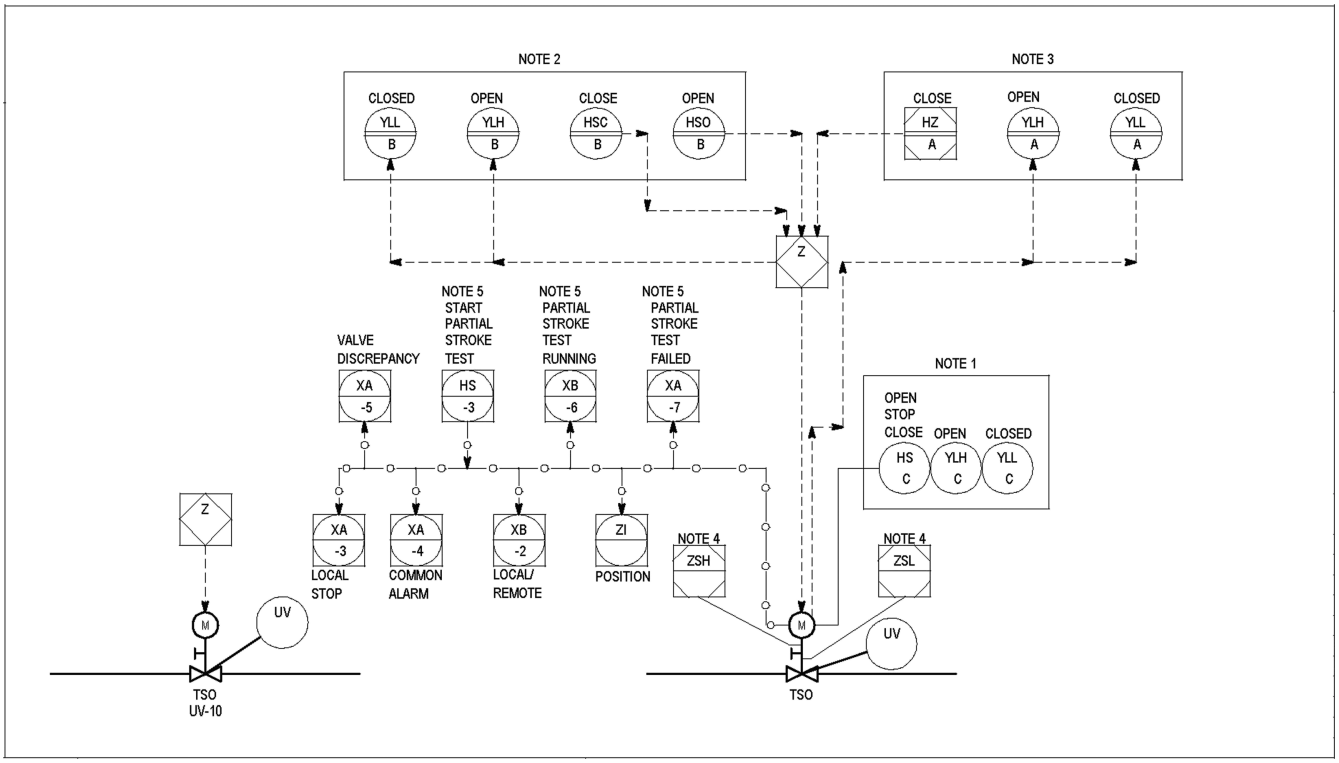


FIGURE 7 | Another example of a typical instrument assembly legend following a different formatting, which does not indicate a clear separation between the simplified P&ID representation and its detailed representation.

- *Project B (the semistructured case)*: This format is more complex, featuring multiple assemblies stacked vertically on a single page without explicit separators, relying only on spatial grouping.
- *Project C (the structured case)*: This format represents a well-structured layout with an explicit vertical line separating the simplified and detailed representations.

By testing our method against these distinct patterns, from clean and structured to dense and ambiguous, we aimed to rigorously evaluate its adaptability. The specific legend format for each project can be seen in Figure 8.

A key architectural choice for this case study was the hybrid use of a VLM for legend analysis and specialized deep learning models for P&ID text extraction. We opted for this approach because using a VLM for the P&ID extraction stage proved impractical. Processing a full, dense P&ID image with a VLM in a single pass resulted in incomplete text extraction, missing critical identifiers. Processing the diagram in smaller "tiles" also failed as VLMs cannot reliably output the precise bounding box coordinates necessary to accurately merge the tiled results. On the other hand, specialized OCR models provide both complete text detection and the accurate coordinate map required for our association step. Our hybrid design is therefore a pragmatic choice, leveraging the precision and speed of OCR for the standardized P&IDs while reserving the VLM's powerful contextual reasoning for the unstandardized and highly variable legend sheets.

5.2.1 | Instrument Assembly Legend Classification

As mentioned in Section 4, we explored the use of VLMs and multimodal prompting to identify the type of legend. We provide an image showing an example of each typical instrument assembly legend type, along with textual instructions to determine the type of the query legend based on the examples. The prompt for typical instrument assembly legend classification is shown in Figure 9. The example image features a single typical instrument assembly for the legends of Projects A and C, as well as several typical assemblies for Project B, as the legends of Project B visualize multiple assemblies in a single legend. Furthermore, none of the typical assemblies in the example image were used as query images. Based on the response of the VLM, the pipeline determines the specific legend extraction prompt to apply to the VLM.

5.2.2 | Instrument Assembly Legend Extraction

We utilize a VLM along with multimodal prompt engineering to extract typical instrument assembly information from legends. As previously mentioned, the legends of Projects A and B are ambiguous and do not visually indicate a clear separation between the simplified and detailed assembly representations. To handle such cases, we employ a prompting method, where an annotated visual example is combined with detailed textual instructions.

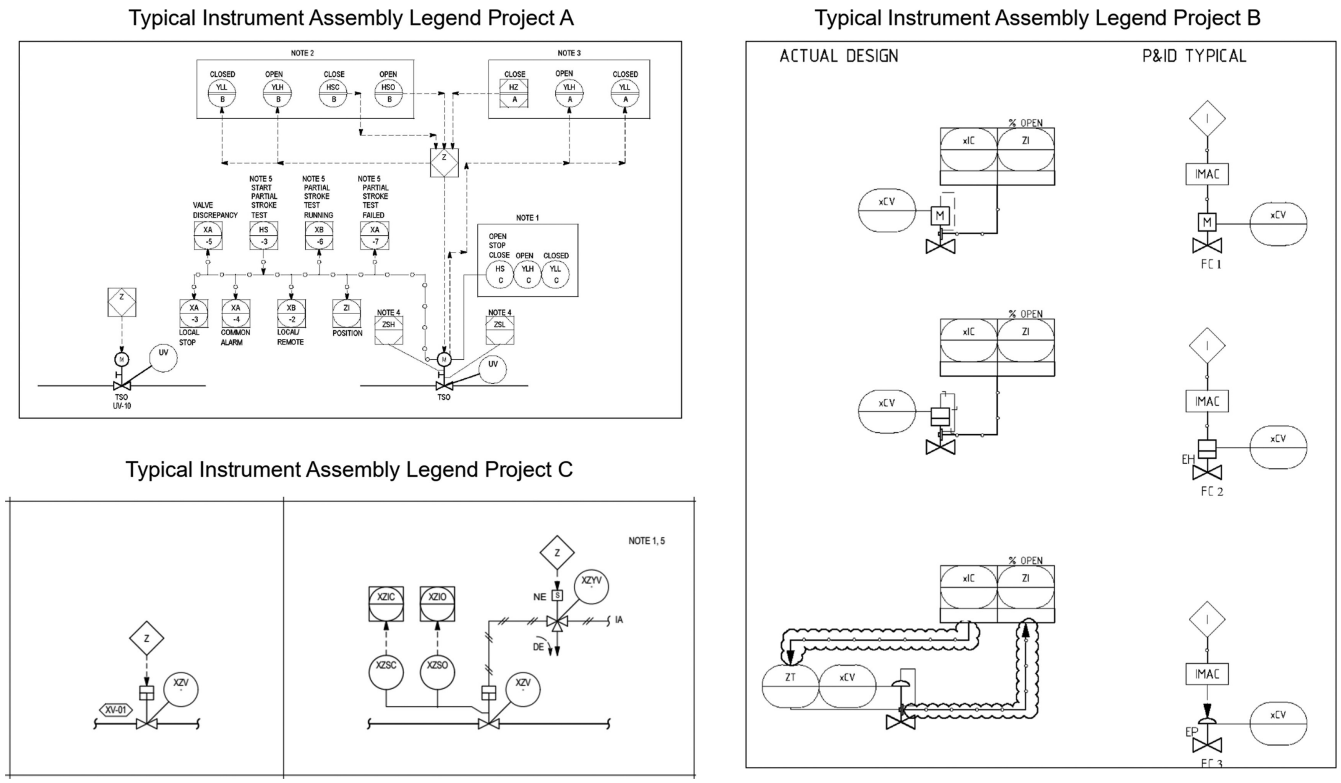


FIGURE 8 | Legend formats of Projects A, B, and C.

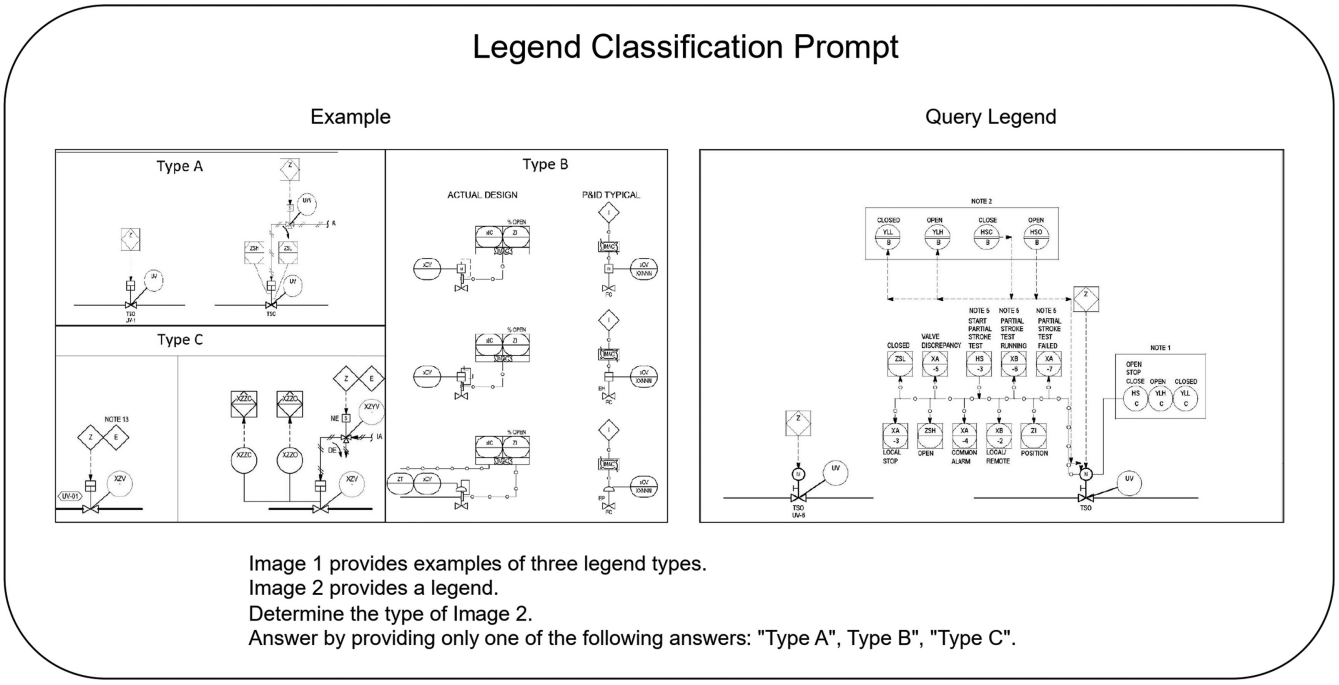


FIGURE 9 | Legend classification prompt.

The creation of these multimodal prompts was a systematic, multistep process. It began with selecting a representative example for each legend type that required visual guidance (Projects A and B). As the legends of Project B involved multiple typicals on separate rows, we cropped a single representative example. This example image was then manually annotated using a standard image editor to draw simple colored circles around the simplified and detailed assemblies. This annotated image serves as the visual component of the in-context learning example. Concurrently, the textual instructions were codeveloped and

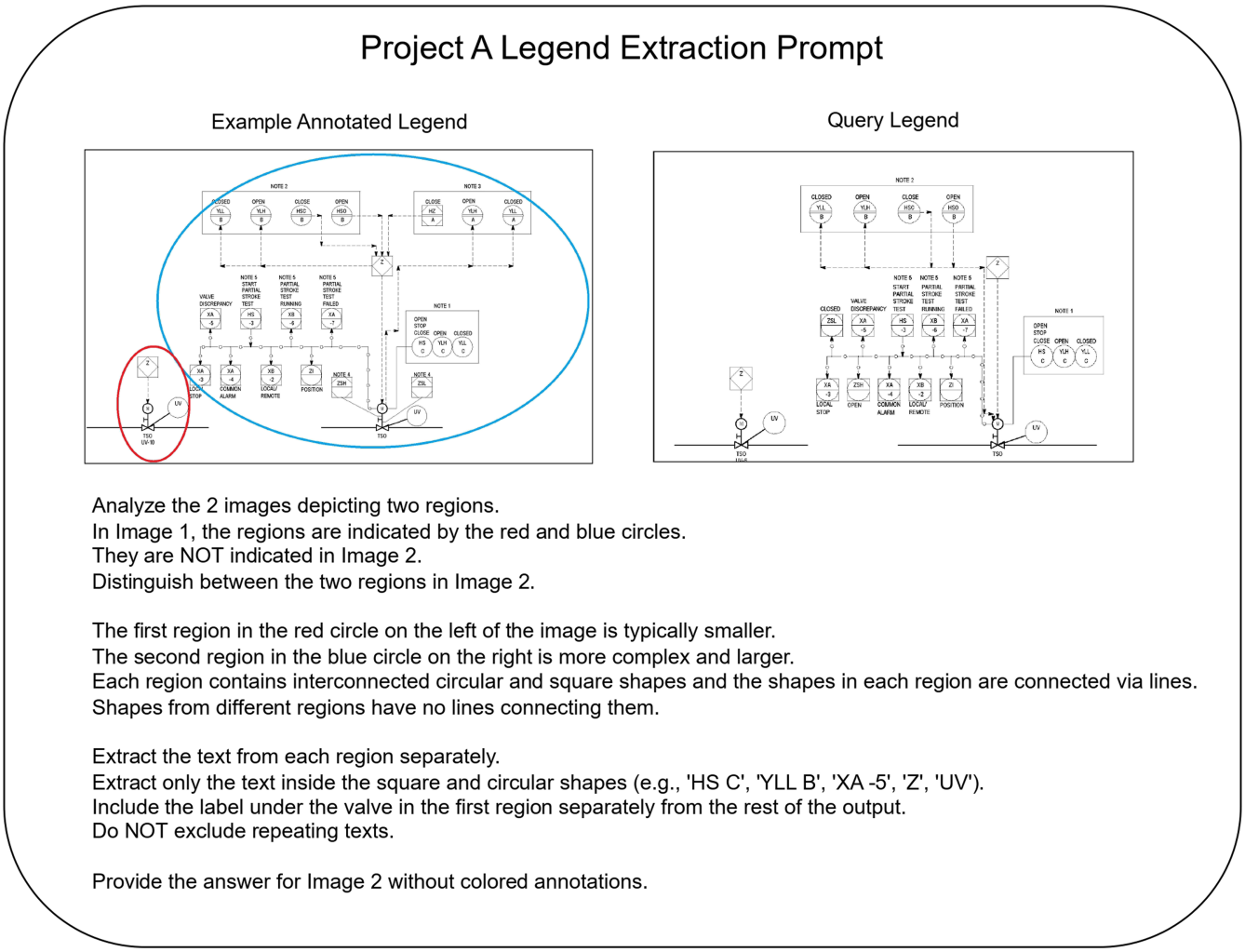


FIGURE 10 | Legend information extraction prompt for Project A.

refined through several iterations with domain experts. This iterative process was crucial for creating precise prompts that could reliably instruct the model on how to interpret the annotated regions, what text to extract, and what to ignore.

This general process was then tailored to handle the specific complexities of each project. For the unstructured format of Project A, the textual instructions explicitly described the relative positions and visual characteristics of the simplified and detailed assemblies to help the model locate them. For the dense format of Project B, which contains multiple assemblies on a single page, the prompt instructed the VLM to identify all distinct "rows" containing a simplified-detailed pair and to process each pair sequentially. Finally, for the structured format of Project C, which has an explicit dividing line, a simpler text-only prompt was sufficient. The final prompts used for each project are shown in Figures 10–12.

To ensure a consistent and parsable response across all extraction tasks, a specific output format was required at the end of each prompt's instructions, though this is omitted from the figures for brevity. The model was instructed to structure its response as follows:

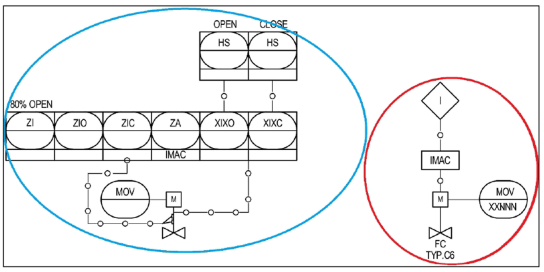
```
Use the following output format:
    **Label:**
    - ...

    **Region 1:**
    - ...

    **Region 2:**
    - ...
```

This structured output was essential for the reliable downstream processing of the extracted information.

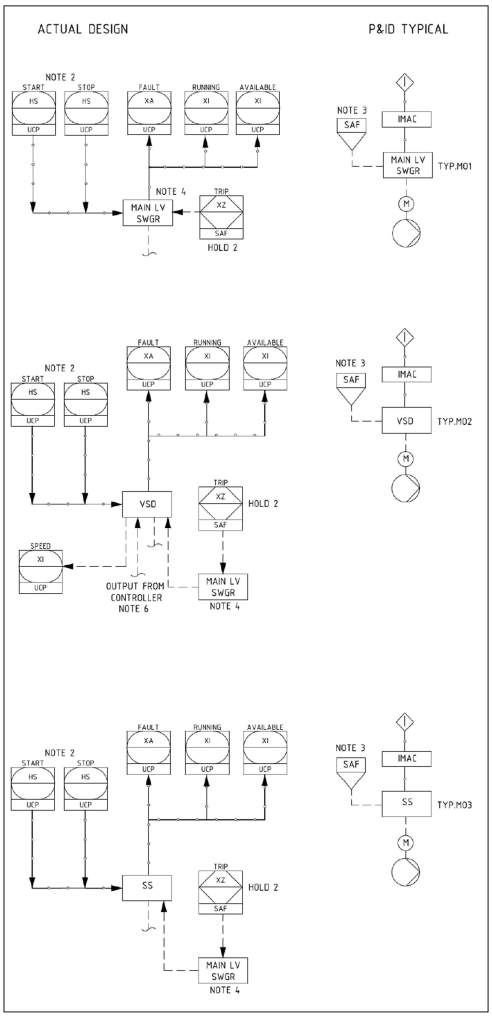
5.2.2.1 | Heuristic-Based Baseline for Comparison. To provide a direct comparison for our novel legend extraction method, we implemented a robust, rule-based baseline designed to replicate traditional document analysis techniques. This baseline employs a hybrid image segmentation strategy that first identifies strong structural cues, such as vertical lines, and then falls back to a more general

Project B Legend Extraction Prompt

Example



Query Legend



Analyze the 2 images.
Image 2 depicts multiple rows, with each row having two regions of assembled symbols.
In Image 1, there is one row with regions that are indicated by red and blue circles.
They are NOT indicated in the Image 2.

First, identify all rows with two regions in Image 2.
Then, distinguish between the two regions in Image 2.
The blue region on the left is more complex and larger.
The red region near the right is typically smaller.
Each region contains interconnected circle 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 in each row separately.
Additionally, ensure that the text extracted from Region 1 and Region 2 is separated in your response for each row.
Extract only the instrument types and their tags inside boxes and circular elements (e.g., 'MOV', 'SDV', 'HS OPEN').

Provide the answer for Image 2 without annotations.
Do NOT exclude repeating texts.
Do not exclude the HS, XZ texts that may be positioned slightly higher in Region 1.
Include the label close to the valve in the second region starting with "TYP" separately from the rest of the output.

FIGURE 11 | Legend information extraction prompt for Project B.

Project C Legend Extraction Prompt

Analyze the provided image.
Extract separately the text from the two regions separated by a line.
Extract only the text inside circles and square shapes in each region (e.g., 'HZ A', 'XZYV B').

Ensure that the text extracted from Region 1 and Region 2 is separated in your response.
Include the label in the diamond shape next to the valve in the first region separately from the rest of the output. Do NOT exclude repeating texts in Region 2. Exclude text such as NOTE, IA, S, NDE, DE, Z, and E.

FIGURE 12 | Legend information extraction prompt for Project C.

analysis of whitespace if these cues are absent. The process is as follows:

1. *Horizontal row segmentation*: The method begins by analyzing the entire legend image to identify distinct horizontal rows. This step is crucial for handling layouts like those in Project B, where multiple legend entries are stacked vertically on a single page. To achieve this, the algorithm creates a histogram that sums the pixel content for each horizontal line of the image. By finding contiguous regions in this histogram with very little to no pixel content, it can identify the large horizontal white spaces that separate the different entries. The image is then sliced at the midpoint of these detected gaps to create individual images for each legend entry.
2. *Vertical column segmentation*: For each extracted row, the baseline then attempts to find a vertical separator to distinguish the "simplified" from the "detailed" assembly. It employs a two-stage approach:
    1. *Line detection (primary method)*: It first searches for a prominent, solid vertical line near the center of the image. This is a strong, unambiguous feature found in many structured legends (like those in Project C). If a qualifying line is detected, its horizontal position is used as the split point.
    2. *Gap detection (fallback method)*: If no line is found, the method falls back to a similar whitespace analysis. It computes a histogram of pixel content for each vertical column of the row and identifies the largest contiguous white space. The midpoint of this largest gap is then used as the vertical split point. This fallback is designed to handle layouts that are columnar but lack explicit line separators.

This multistage heuristic is designed to be as robust as possible for structured and semistructured layouts. However, as demonstrated in our results, it is inherently brittle and fails on unstructured layouts (like those in Project A) where these geometric assumptions do not hold.

The performance of this baseline is sensitive to several key parameters, which were chosen empirically based on an analysis of our dataset. To be considered a valid separator, a horizontal gap was required to be at least 40 pixels high, and a vertical gap at least 50 pixels wide. These thresholds were selected to ensure the algorithm ignored small, incidental spaces between words or symbols, and only identified the larger gaps that define the main layout. For line detection, a line was required to span at least 70% of the image's height to be considered a global separator, a constraint designed to filter out shorter, incidental vertical lines that are part of the diagrams' symbols.

5.2.3 | P&ID Text Extraction

To detect text within P&IDs, the Progressive Scale Expansion Network (PSENet) [28] was employed. This model is noted for its efficiency in densely populated areas and its ability to handle text in various orientations, making it particularly suited for extracting information from P&IDs. Additionally, text recognition is carried out using the pretrained PP-OCR recognizer [29].

The PSENet detector's training process involves a tiling technique [30]. Each P&ID is divided into 16 overlapping sections, or "tiles," with a 200-pixel overlap to enhance the detection of small text elements. The training utilized the Adam optimizer with a learning rate of 0.001 across 40 epochs and a batch size of eight tiles.

During text extraction, the P&IDs are initially split into overlapping tiles. Then, the PSENet detector locates the bounding box coordinates of each text in the tiles. Following this, the bounding boxes are translated into the full P&ID, and overlapping detections due to the overlapping tiling are merged. The resulting text regions are then cropped and processed using the PP-OCR recognizer to extract the textual content.

5.2.4 | Instrument Listings Generation

The generation of the instrument listing is visualized in Figure 13 and is based on our earlier work [14]. The instrument listing generation begins by identifying all typical numbers that are found both in the legend sheet and the extracted P&ID text. Following this, the method determines the instruments linked to these typical numbers by locating the nearest texts using Euclidean distance. Only those texts that match the instruments from the legend's simplified representation are retained. Next, the method identifies the tag numbers for these instruments. This is accomplished by finding the closest text to one of the identified instrument types using Euclidean distance and then applying engineering knowledge rules to ensure the text complies with tag number formatting. Finally, the relevant instruments within the typical assembly are generated by assigning the identified tag number to the instruments from the detailed representation in the legend. This methodology adheres to the conventional practices employed by instrumentation engineers.

5.3 | Data Collection

The training dataset for the P&ID text detector, as well as the evaluation dataset for the methods for legend information extraction and the method for incorporating legend information into the diagram information extraction, comprised hundreds of industrial P&IDs from past large-scale EPC projects executed by McDermott. The diagrams and the legends are solely McDermott's proprietary data. Specifically, the training dataset for the P&ID text detector comprises 355 P&IDs from different EPC projects. Furthermore, the P&IDs comprising the evaluation dataset were collected from three large EPC projects executed by McDermott. Each evaluation project had unique legend and drawing standards. As previously mentioned, we will refer to these evaluation projects as A, B, and C. All P&IDs were initially in PDF format and were subsequently converted into image files. Furthermore, the data were heavily augmented to remove IP traces.

In total, the evaluation dataset consists of 50 P&IDs and 10 typical instrument assemblies in the legends of each project.

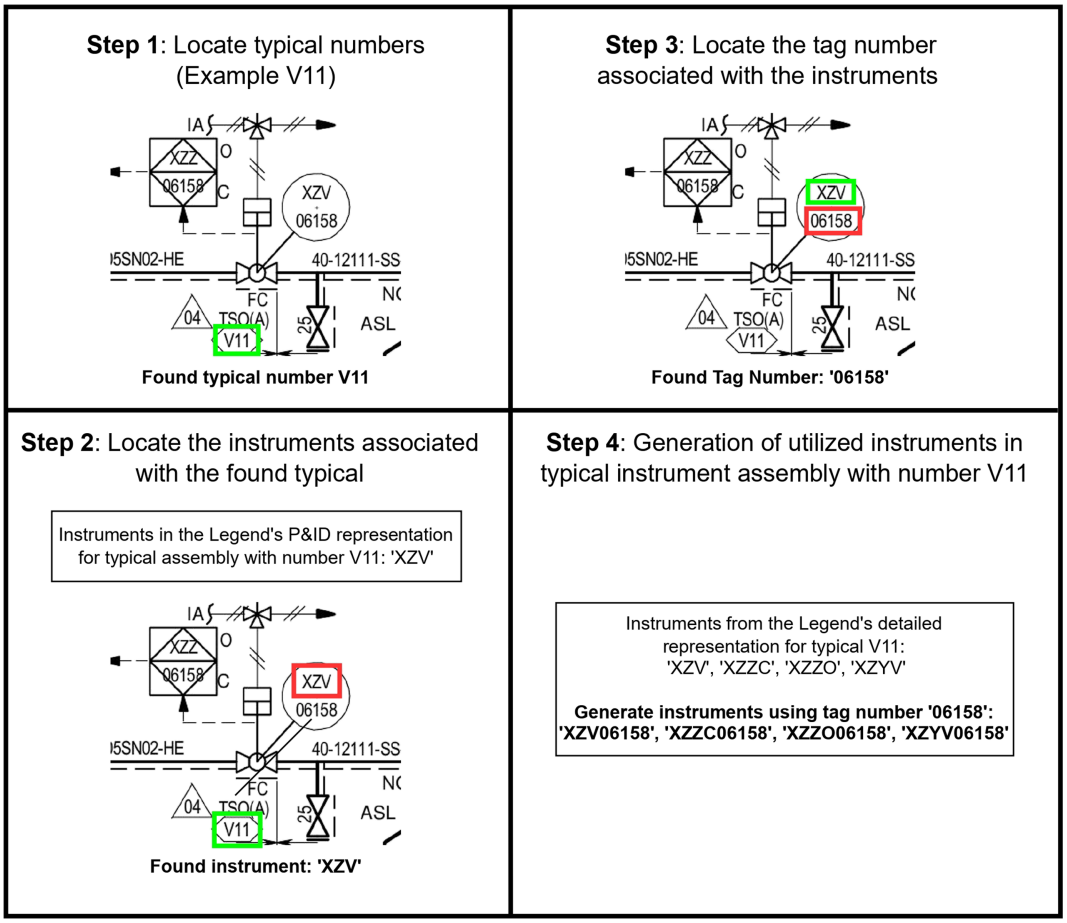


FIGURE 13 | Instrument listing generation method, which integrates the extracted information from the typical instrument assembly legends with the information from the P&IDs.

TABLE 1 | The distribution of the total number of typical instrument assemblies in the legends and the P&IDs per evaluation project.

	Project A	Project B	Project C
Typical instrument assemblies	10	10	10
P&IDs	20	10	20

None of the evaluation P&IDs were used in the training data of the PSENet detector. The distribution of the total number of typical instrument assemblies in the legends and the P&IDs per evaluation project are shown in Table 1. Project B contains fewer P&IDs as the scale of the P&IDs is larger, and each diagram contains more information compared with the diagrams of the other two projects. As a result, the project had fewer P&IDs.

5.4 | Evaluation

The following sections detail the methodologies and evaluation criteria for the methods for the legend classification and information extraction, text extraction from P&ID diagrams, and instrument listing generation. Additionally, we document the setup and evaluation of the VLM and querying process used for classifying and extracting information from legends.

5.4.1 | VLM and Querying Setup

To classify and extract information from the legends, we utilized the GPT-4o model (version "2024-11-20") via McDermott's Azure OpenAI API (version "2024-02-01"). The prompts were executed using the LlamaIndex library, which facilitates the construction of the multimodal API requests.

Several key configurations were implemented to ensure reliable and high-quality responses. We configured the requests to process images at high detail, providing the model with the best possible visual information for accurate interpretation. To handle the structured data output, the response length was capped at a limit of 1024 tokens for even the most complex legends. Additionally, to ensure robustness against transient network issues, the API calls were configured with a 180-second timeout and up to five automated retries.

The model's temperature was set to 0 to promote deterministic and consistent outputs. Moreover, previous research reports that the LLMs and VLMs can produce variable results despite setting the temperature to 0 [31, 32]. Thus, all experiments involving the VLM were executed 20 times.

5.4.2 | Evaluation of Legend Classification Method

We evaluate the legend classification method by calculating the classification accuracy of the VLM, that is, the number of correctly classified legends out of all legends:

$$\text{Accuracy} = \frac{\text{Number of correctly classified legends}}{\text{Total number of legends}} \times 100$$

As previously stated, each input legend is queried three times, and the results are averaged.

5.4.3 | Evaluation of Legend Extraction Method

We evaluate the legend information extraction method by calculating the number of correctly grouped instruments in the typical representation and the detailed representation out of all instruments:

$$\text{Grouping Accuracy} = \frac{\text{Number of correctly grouped instruments}}{\text{Total number of instruments}} \times 100$$

Furthermore, we calculate the accuracy of the text recognition of the VLM, that is, the number of correctly recognized instruments out of all instruments on the legends:

$$\text{Recognition Accuracy} = \frac{\text{Number of correctly recognized instruments}}{\text{Total number of instruments}} \times 100$$

Similarly to the legend identification method, each input legend is queried three times, and the results are averaged.

5.4.4 | Evaluation of P&ID Text Extraction Method

The text extraction method is evaluated based on the precision and recall of all extracted typical numbers, that is, the percentage of correctly extracted typical numbers out of all extracted typical numbers and the percentage of correctly extracted typical numbers out of all ground truth typical numbers:

$$\text{Precision} = \frac{\text{Number of correctly extracted typical numbers}}{\text{Total number of extracted typical numbers}} \times 100$$

$$\text{Recall} = \frac{\text{Number of correctly extracted typical numbers}}{\text{Total number of ground truth typical numbers}} \times 100$$

The same evaluation is done for the instruments associated with the typical numbers and their corresponding tag numbers.

5.4.5 | Evaluation of Instrument Listing Generation Method

The final end-to-end performance of our method is evaluated by calculating the precision and recall of the generated instrument instances. For this evaluation, a single "generated instrument" is defined as the unique combination of an instrument type from the legend's detailed assembly and the corresponding tag number extracted from the P&ID. For example, if the P&ID shows the typical number V11 associated with tag number 06158, and the legend's detailed assembly for V11 contains the types XZV and XZZC, the method is expected to generate two distinct instrument instances: XZV06158 and XZZC06158.

A correctly generated instrument is therefore defined as a generated instance where both the instrument type and the assigned tag number perfectly match the ground truth. The total count of these correct instances is then used to calculate precision and recall.

$$\text{Precision} = \frac{\text{Number of correctly generated instruments}}{\text{Total number of generated instruments}} \times 100$$

$$\text{Recall} = \frac{\text{Number of correctly generated instruments}}{\text{Total number of ground truth instruments}} \times 100$$

In these formulas, the "Total number of generated instruments" is the complete set of unique instrument instances produced by our method across all P&IDs. The "Total number of ground truth instruments" is the complete set of all instances that should have been generated according to the ground truth data. This per-instance evaluation provides a granular and accurate measure of the entire integration pipeline's performance.

6 | Results

This section presents the results of the case study for extracting instrument listings from P&IDs. Specifically, it reports the results of the methods for legend classification and extraction, the P&ID text extraction, and the instrument listing generation. Each subsection showcases the accuracy and reliability of the methods used for different projects.

6.1 | Instrument Assembly Legend Classification Results

The results of the legend identification method for each project are presented in Table 2. As can be seen, the method achieved 100% classification accuracy for each typical in the legend sheets in the three projects. These results showcase that VLMs can successfully be applied for automating the legend information extraction process, given the variability of legend formats.

6.2 | Instrument Assembly Legend Extraction Results

The results of the legend information extraction method for each project are presented in Table 3. As can be seen, the method

TABLE 2 | The accuracy results of the instrument assembly legend classification method.

	Project A	Project B	Project C
Classification accuracy	100%	100%	100%

TABLE 3 | The grouping and text recognition accuracy results of the instrument assembly legend extraction method.

	Project A	Project B	Project C
Grouping accuracy	100%	100%	100%
Recognition accuracy	100%	100%	100%

TABLE 4 | Grouping accuracy of heuristic-based baseline legend information extraction method.

Project A	Project B	Project C
24.1%	96.5%	100.0%

achieved 100% grouping and recognition accuracy for each typical in the legend sheets in the three projects.

We also explored alternative prompting strategies. A purely textual prompt without a visual example was insufficient, as the VLM struggled to accurately group the instruments. Conversely, we tested a full few-shot approach by providing the annotated example image along with its correct output. This common technique proved detrimental by causing the model to hallucinate and copy content from the example answer into its response for the new query. This finding revealed a critical trade-off in multimodal prompting for this task. Our final strategy of combining a visual annotation to guide attention with a separate textual instruction proved to be the most robust, as it effectively guided the model without overly biasing its output.

To evaluate the effectiveness of our proposed method, we compared its performance against the heuristic-based baseline described in Section 5.2.2.1. The primary metric for this comparison is Grouping Accuracy, which measures the percentage of correctly grouped instruments into "simplified" and "detailed" categories for each legend entry. The results are summarized in Table 4.

The performance of the heuristic-based baseline clearly demonstrates the limitations of rule-based approaches when faced with format variability. On the perfectly structured legends of Project C, which contain clean vertical line separators, the baseline achieved 100.0% accuracy. However, its performance degraded to 96.5% on the semistructured format of Project B, where it struggled with horizontal regions with limited vertical whitespace separation. Most notably, the baseline's performance collapsed to 24.1% on the unstructured format of Project A, as the absence of vertical whitespace separation rendered the rule-based segmentation ineffective.

6.3 | P&ID Text Extraction Results

The extraction models achieved 100% extraction precision and recall on all instruments and their tags in all projects. Furthermore, the typical number identification results are shown in Table 5. The text extraction methods were able to recognize 95.75% of typical numbers across the three projects.

TABLE 5 | Typical number extraction results.

	Project A	Project B	Project C	Total
Correct	51	218	69	338
Missed	4	11	0	15
Wrong	0	0	0	0
Recall	92.73%	95.20%	100.00%	95.75%
Precision	100.00%	100.00%	100.00%	100.00%

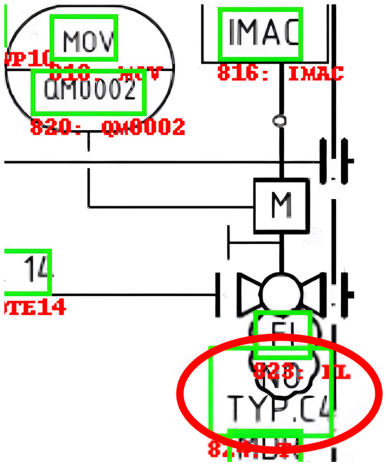


FIGURE 14 | An example of a missed typical number due to a detection bounding box, which captures additional text.

As can be seen, some typical numbers were not detected in the P&IDs. These typical numbers were likely missed as the text was positioned close to other text elements. The results indicate their reliability for P&ID information extraction. An example of a missed typical number is shown in Figure 14, where the detection captured additional information besides the typical number, preventing its proper recognition.

6.4 | Instrument Listing Generation Results

The instrument assembly generation results are shown in Table 6. These results demonstrate our method's ability to generate the instruments in typical assemblies accurately across projects. The missed instruments reported in Table 6 are due to the missed typical numbers reported in Table 5. Furthermore, the wrongly expanded instruments in Table 6 are due to an incorrectly associated instrument tag number, due to the Euclidean distance association heuristic. An example of this issue is shown in Figure 15, where Instrument "MOV QM0001" is closer to the typical number "TYP.C4" compared with the instrument "MOV QM0015." Nevertheless, most of the instruments were generated correctly.

7 | Discussion

This study introduces a novel method for automating information extraction from technical diagrams by integrating

TABLE 6 | Instrument listing generation results.

	Project A	Project B	Project C	Total
Correct	443	1614	430	2487
Missed	18	88	0	106
Wrong	0	8	0	8
Recall	96.10%	94.83%	100.00%	95.91%
Precision	100.00%	99.51%	100.00%	99.68%

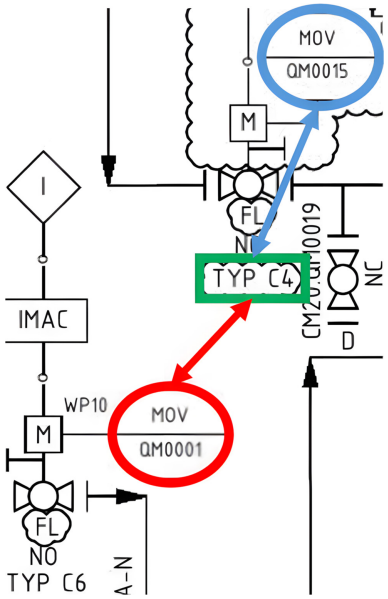


FIGURE 15 | An example of a missed typical number due to a detection bounding box, which captures additional text.

information from external, variably formatted legend sheets. The following discussion analyzes our findings in the context of prior work, evaluates the practical implications of our results, and outlines the limitations and future directions of this research.

7.1 | Comparison With Alternative Methods

A significant limitation of most previous studies on diagram information extraction [2, 6–10, 13, 36] is their focus on extracting only what is explicitly depicted, overlooking implicit components detailed in reference documents like legends. The few methods that do reference legends, such as that of Sarkar et al. [11], are limited to simple symbol matching and cannot handle complex assemblies. In comparison with these studies, we propose a novel method that integrates information from both the diagrams and their corresponding legend sheets. We validated our method via a case study focusing on integrating legend and diagram information to produce a listing of all instruments listed on the diagrams. This case study involved three large EPC projects with varying legend formats and diagram standards, demonstrating the flexibility and robustness of our approach.

To provide a direct quantitative comparison for the novel task of legend extraction, we evaluated our VLM-based approach against a rule-based baseline using hybrid line and gap detection heuristics. As detailed in our results (Table 4), the heuristic baseline's performance was highly dependent on the legend format. It achieved 100% accuracy on the perfectly structured layouts of Project C but saw its performance degrade on the semistructured (96.5%) and unstructured (24.1%) formats of Projects B and A, respectively. In contrast, our VLM-based method achieved 100% accuracy across all formats. This result empirically validates that our approach is not just effective but is fundamentally more robust and adaptable than traditional methods, which are too brittle to handle the document variability common in real-world industrial settings.

For the P&ID text extraction component of our pipeline, we chose the PSENet detector. This decision was informed by our prior comparative research [10, 36], which demonstrated its superior performance on dense engineering diagrams compared with other detectors, such as EAST [33] and CRAFT [34]. By leveraging a proven, state-of-the-art model for this established subtask, we focused our novel contributions on the primary challenge of legend integration.

7.2 | Practical Implications

The performance of our end-to-end pipeline, particularly the overall 95.91% recall and 99.68% precision, must be interpreted within its real-world industrial context. In this workflow, an automated tool generating an initial “Instrument Index” draft with such high accuracy is considered very good. This level of performance provides engineers with a highly reliable and nearly complete document, transforming their task from one of laborious creation from scratch to one of efficient validation and correction.

The errors produced by the method also have practical implications. The false negatives (missed instruments) are potentially critical, as their omission could lead to material shortfalls and subsequent project delays if not rectified. Conversely, the cost of false positives is relatively low, typically resulting in a negligible amount of surplus material. The primary goal of our automated approach is to minimize such omissions during the initial drafting stage. While these errors must be managed, the cost of the verification step itself is significantly lower than the cost of a fully manual generation process from scratch, representing a substantial net gain in efficiency.

However, a manual review remains an essential step in the process to achieve the 100% accuracy required for final, safety-critical deliverables. The established, multistage engineering review process is specifically designed as a safeguard to identify and correct any errors from the generation phase. For an experienced engineer, this review is conceptually straightforward but can be meticulous and time consuming. The key benefit of our method is that it fundamentally simplifies the work of engineers. Instead of the complex cognitive load of creating a list from scratch, the engineer's role shifts to the more manageable task of validating a nearly complete list, a significantly faster and less error-prone activity.

7.3 | Generalizability and Scalability of the VLM-Based Approach

This study provides strong evidence that VLMs, guided by multimodal prompt engineering and in-context learning, can enable information extraction from diverse technical documents in a training-free manner. The 100% accuracy in legend classification and extraction across varied formats indicates that VLMs, when properly prompted, can interpret complex visual and textual data, making them a powerful tool for automating document understanding.

While our method involves manual annotation for the in-context examples, this requirement is minimal. It only necessitates the annotation of a single example per legend type, a stark contrast to traditional deep learning methods that require thousands of annotated samples for training. This efficiency makes our approach practical and scalable, allowing for quick adaptation to new projects with different legend formats without undermining its feasibility.

The proposed method is highly generalizable. The use of VLMs and multimodal prompting allows for the handling of diverse legend formats without the need for extensive retraining or the creation of custom heuristic rules for each new project. This adaptability makes our method applicable to a wide range of industries beyond EPC, such as aerospace (for avionics blueprints), manufacturing (for mechanical diagrams), and healthcare (for medical device schematics), where similar workflows of cross-referencing simplified diagrams with detailed reference documents exist. It represents a versatile solution for any domain requiring accurate and efficient interpretation of complex technical documents.

7.4 | Deployment and Performance Considerations

For practical application in an industrial setting, our hybrid method is deployed as a secure web application. This section outlines the computational requirements and performance considerations for both the VLM and local OCR components.

A key aspect of our deployment is the use of the McDermott Azure OpenAI API for all VLM-based tasks. This enterprise-grade service ensures that all data remains within a secure, private corporate environment, addressing critical data privacy and confidentiality concerns.

The OCR-based P&ID processing is computationally intensive and is optimized to run on GPU-enabled infrastructure. To achieve high throughput, the processing environment requires a GPU with 16 GB of VRAM and 12 GB of system RAM. We also leverage a tiling preprocessing technique, which not only improves detection accuracy but also significantly enhances scalability. This allows the system to process up to 200 P&IDs in a single batch operation. With these resources, the system achieves a processing throughput of 100 P&IDs in 25 min for the text extraction stage.

7.5 | Limitations

Our study presents significant advancements in the field of information extraction from diagrams incorporating legend information. Nevertheless, several limitations should be acknowledged.

One significant limitation of this study is the challenging reproducibility due to the confidentiality of the data used. The training and evaluation datasets comprised industrial P&IDs and legend sheets from past projects executed by McDermott, which are proprietary and cannot be publicly shared. This restricts other researchers' ability to replicate our experiments and validate the results independently. While we have provided a detailed methodology, the lack of access to the specific datasets used in this study may hinder the reproducibility and verification of our findings. Thus, it is recommended that other industries and researchers replicate our methods using their specific datasets. This would enable them to determine the applicability of our method in other use cases.

Furthermore, our evaluation was conducted on three distinct legend formats. These formats were specifically selected to represent a spectrum of real-world challenges—from highly structured (Project C) to dense and semistructured (Project B) to fully unstructured (Project A). However, they do not encompass all possible variations. Other complex layouts, such as those with a vertical orientation, undoubtedly exist. Although our method's training-free adaptability is designed to handle such variability by simply creating a new one-shot prompt, its performance on formats beyond those tested has not been empirically validated.

Another limitation is the need to evaluate alternative VLM models. While the study employed the GPT-4o model, other VLMs were not extensively investigated. It would be beneficial to investigate the effectiveness of the method using alternative models. This could provide deeper insights into the robustness and applicability of our approach.

We did not systematically investigate the effect of visual noise and congestion, which could impact the performance of our hybrid pipeline. The impact of these factors, however, is likely to differ between the two main components of our method. Legend sheets are typically clean documents but can still suffer from low-resolution scans or degradation, which could affect the VLM's text recognition. Furthermore, while uncommon, handwritten annotations on a legend could potentially interfere with our multimodal prompting strategy, which uses colored shapes to guide the model's attention. In contrast, P&ID diagrams are far more susceptible to visual degradation. They are often dense, subject to revision clouds and markups, and can contain overlapping annotations. Based on our prior work, which showed that such occlusions can negatively affect symbol recognition [9], it is reasonable to expect that heavy congestion or noise on the P&IDs could degrade the performance of the PSENet text detector. This could lead to missed typical numbers or instrument tags, directly impacting the recall of the final instrument list. A detailed analysis of the pipeline's robustness to varying levels of noise and congestion on P&IDs remains a key direction for future work.

A limitation of the legend extraction method is that its accuracy might depend significantly on the utilized prompt [35]. While our prompting methods have demonstrated success, variations in prompt construction could lead to different levels of accuracy in the extracted information.

Lastly, although our method was validated by a selected group of McDermott engineers, we have not deployed the method across the entire organization. Thus, at this time, we cannot gauge whether the method will be adopted by the organization and enable the reduction of manual engineering hours.

7.6 | Future Work

As discussed, the validation of the proposed method has high results in terms of legend classification and extraction, diagram information extraction, as well as integrating the legend and diagram information. Nevertheless, additional research is still needed.

As we only explored the GPT-4o VLM, it is needed to investigate the performance of alternative VLMs with different parameter sizes. This includes assessing models with fewer parameters to see if they can achieve similar accuracy, as well as models with more parameters to determine if they offer improved performance. This research could provide insights into the scalability and efficiency of various models.

Future work should systematically investigate the impact of noise and congestion, particularly on the P&ID processing stage. This would involve creating a benchmark dataset of diagrams with varying levels of degradation (e.g., low-resolution scans, handwritten markups, and high component density) to quantify the performance limits of the text detector. Investigating techniques to make multimodal prompts more resilient to unexpected annotations on legend sheets would also be a valuable contribution.

Another area for future research could be the study of alternative prompting methods for information extraction from blueprints and legends with diverse formatting. This involves experimenting with different prompt methods and analyzing their impact on extraction accuracy. Future work should include a comprehensive evaluation of various prompt engineering techniques to determine their effectiveness across different legend formats.

Another promising direction is to improve the instrument association strategy. Our current Euclidean distance heuristic can fail in visually congested areas where multiple tag numbers are near a single instrument. Future work could develop a hybrid, two-stage process. First, the fast heuristic would identify all potential associations. Then, only for ambiguous cases where multiple candidates have similar distances, the system would trigger a VLM. This would involve cropping a small "region of interest" containing the instrument and its candidate tags and feeding it to the VLM to make the final determination. This targeted use of the VLM would resolve the most challenging errors while maintaining overall efficiency by avoiding unnecessary API calls.

8 | Conclusion

This research aimed to automate the extraction of information from diagrams that require referencing external legend sheets for correct diagram analysis. The challenge of integrating legend information into an automated pipeline arises from the different formats of legends in various industries.

Prior research in diagram information extraction has focused predominantly on extracting information exactly as it is depicted on the diagrams, without considering components that are not explicitly presented. The diagrams' legend sheets are essential for understanding the diagrams as they present the simplified representations of components often found in diagrams, as well as their detailed representation. Thus, if this legend information is not incorporated, this can lead to the incomplete and inaccurate digitalization of diagrams. Furthermore, the legends can vary in formatting across engineering projects and industries. As existing methods typically rely on unimodal information extraction techniques, they require extensive training and the development of custom heuristic rules, thereby limiting their scalability and adaptability.

This study introduces a novel method that integrates information from both diagrams and their corresponding legend sheets, which enables the inclusion of elements that are not explicitly depicted in the diagrams. By leveraging a novel integration of traditional OCR tools along with VLMs, multimodal prompt engineering, and in-context learning, our approach can extract information from legend sheets with diverse formats. The method enables a more comprehensive and accurate extraction of information, incorporating the information in the legend sheets, which is crucial for the accurate interpretation of diagrams.

The method was validated through a case study involving extracting a list of instruments from three large EPC projects, each with unique legend formats and diagram standards. The results demonstrated the flexibility and robustness of our approach, achieving high precision and recall rates in information extraction and integration. Specifically, the instrument listing extraction approach attained an overall recall of 95.91% and precision of 99.68%, highlighting the method's reliability in extracting and utilizing information from both the diagrams and legends.

Our study also provides evidence that VLMs, guided by multimodal prompts, can classify and extract information from diverse legend sheets with 100% accuracy in both legend classification and information extraction. This indicates that VLMs are capable of interpreting complex visual and textual data in legend sheets, making them a powerful tool for automating the information extraction process.

One of the significant advantages of our method is its minimal annotation requirement. It only necessitates the annotation of a single example per legend type, significantly reducing the annotation effort compared with traditional methods. This efficiency, combined with the adaptability of VLMs and multimodal prompting, makes our method practical for quick adaptation to new projects with different legend formats.

Furthermore, the proposed method is highly generalizable and can be easily customized for various use cases, industrial diagrams, and legend formats. This generalizability ensures that the method can be applied across a wide range of industries and projects, providing a versatile solution for information extraction from industrial diagrams and legends. The adaptability of the method can significantly reduce the time and potential errors associated with the manual analysis of industrial diagrams, enhancing the accuracy and efficiency of information extraction.

Acknowledgments

This work is supported by McDermott Inc. and Software Center (Gothenburg, Sweden) and conducted in collaboration with Eindhoven University of Technology, the Netherlands. During the preparation of this work, the authors used the GPT-4o model in order to improve the readability of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This work is supported by McDermott Inc. and Software Center (Gothenburg, Sweden) and conducted in collaboration with Eindhoven University of Technology, the Netherlands.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

This research was conducted using proprietary data with legal and commercial restrictions. Thus, due to the nature of the research, the supporting data are not available.

References

1. Process Industry Practices, *PIP PIC001—Piping and Instrumentation Diagram Documentation Criteria—Technical Correction: 6/2023*," Tech. rep. (Process Industry Practices (PIP), 2023).

2. R. Dzhusupova, M. Ya-alimadad, V. Shteriyanov, J. Bosch, and H. Holmstrm Olsson, "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)* (IEEE, 2024), 1023–1033.

3. N. GCR, *Cost Analysis of Inadequate Interoperability in the US Capital Facilities Industry* (National Institute of Standards and Technology (NIST), 2004), 223–253.

4. International Electrotechnical Commission, "IEC Understanding Standards," accessed February 26, 2025, https://www.iec.ch/understanding-standards.

5. American National Standards Institute, "ANSI Introduction," accessed February 26, 2025, https://www.ansi.org/about/introduction.

6. B. C. Kim, H. Kim, Y. Moon, G. Lee, and D. Mun, "End-to-End Digitization of Image Format Piping and Instrumentation Diagrams at an Industrially Applicable Level," *Journal of Computational Design and Engineering* 9, no. 4 (2022): 1298–1326.

7. R. Rahul, S. Paliwal, M. Sharma, and L. Vig, "Automatic Information Extraction From Piping and Instrumentation Diagrams," preprint, arXiv, January 28, 2019, https://doi.org/10.48550/arXiv.1901.11383.

8. S. Mani, M. A. Haddad, D. Constantini, W. Douhard, Q. Li, and L. Poirier, "Automatic Digitization of Engineering Diagrams Using Deep Learning and Graph Search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, 2020), 673–679, https://doi.org/10.1109/CVPRW50498.2020.00096.

9. V. Shteriyanov, R. Dzhusupova, J. Bosch, and H. H. Olsson, "Unraveling the Impact of Density and Noise on Symbol Recognition in Engineering Drawings," in *2024 IEEE 12th International Conference on Intelligent Systems (IS)* (IEEE, 2024), 1–7.

10. V. Shteriyanov, R. Dzhusupova, J. Bosch, and H. Holmstrm Olsson, "Robust Detection of Line Numbers in Piping and Instrumentation Diagrams (P&IDs)," in *2024 International Conference on Machine Learning and Applications (ICMLA)* (IEEE, 2024), 888–893, https://doi.org/10.1109/ICMLA61862.2024.00129.

11. S. Sarkar, P. Pandey, and S. Kar, "Automatic Detection and Classification of Symbols in Engineering Drawings," preprint, arXiv, April 28, 2022, https://doi.org/10.48550/arXiv.2204.13277.

12. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning* (PMLR, 2021), 8748–8763.

13. M. T. Khan, L. Chen, Y. H. Ng, W. Feng, N. Y. J. Tan, and S. K. Moon, "Fine-Tuning Vision-Language Model for Automated Engineering Drawing Information Extraction," preprint, arXiv, November 6, 2024, https://doi.org/10.48550/arXiv.2411.03707.

14. V. Shteriyanov, R. Dzhusupova, J. Bosch, and H. H. Olsson, "Automating the Expansion of Instrument Typicals in Piping and Instrumentation Diagrams (P&ids)," *Proceedings of the AAAI Conference on Artificial Intelligence* 39, no. 28 (2025): 28885–28891, https://ojs.aaai.org/index.php/AAAI/article/view/35155.

15. B. Xiao, H. Wu, W. Xu, et al., "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 4818–4829.

16. A. C. Doris, D. Grandi, R. Tomich, et al., "DesignQA: A Multimodal Benchmark for Evaluating Large Language Models Understanding of Engineering Documentation," *Journal of Computing and Information Science in Engineering* 25, no. 2 (2025): 021009.

17. M. Chiasson, M. Germonprez, and L. Mathiassen, "Pluralist Action Research: A Review of the Information Systems Literature," *Information Systems Journal* 19, no. 1 (2009): 31–54.

18. S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting Empirical Methods for Software Engineering Research," in *Guide to Advanced Empirical Software Engineering* (Springer, 2008), 285–311.

19. J. Gerring, *Case Study Research: Principles and Practices* (Cambridge University Press, 2006).

20. G. Walsham, "Interpretive Case Studies in Is Research: Nature and Method," *European Journal of Information Systems* 4, no. 2 (1995): 74–81.

21. C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing* 5, no. 4 (2000): 13–22.

22. S. Shahriar, B. D. Lund, N. R. Mannuru, et al., "Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency," *Applied Sciences* 14, no. 17 (2024): 7782.

23. A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What Does CLIP Know About a Red Circle? Visual Prompt Engineering for VLMs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2023), 11987–11997.

24. Y. Zhang, K. Zhou, and Z. Liu, "What Makes Good Examples for Visual In-Context Learning?," *Advances in Neural Information Processing Systems* 36 (2023): 17773–17794.

25. M. F. Theisen, K. N. Flores, L. S. Balhorn, and A. M. Schweidtmann, "Digitization of Chemical Process Flow Diagrams Using Deep Convolutional Neural Networks," *Digital Chemical Engineering* 6 (2023): 100072.

26. Process Industry Practices, *PIP PCEDO001—Guidelines for Control Systems Documentation*," Tech. rep. (Process Industry Practices (PIP), 2015).

27. European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). L 168* (Official Journal of the European Union, 2024), 1–254, https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

28. W. Wang, E. Xie, X. Li, et al., "Shape Robust Text Detection With Progressive Scale Expansion Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), 9336–9345.

29. C. Li, W. Liu, R. Guo, et al., "PP-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System," preprint, arXiv, June 14, 2022, https://doi.org/10.48550/arXiv.2206.03001.

30. F. Ozge Unel, B. O. Ozkalayci, and C. Cigla, "The Power of Tiling for Small Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2019).

31. A. Bendeck and J. Stasko, "An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks," *IEEE Transactions on Visualization and Computer Graphics* 31, no. 1 (2025): 1105–1115.

32. T. Guan, F. Liu, X. Wu, et al., "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* (IEEE, 2024), 14375–14385.

33. X. Zhou, C. Yao, H. Wen, et al., "East: An Efficient and Accurate Scene Text Detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), 5551–5560.

34. Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), 9365–9374.

35. Y. Zhang, K. Zhou, and Z. Liu, "What Makes Good Examples for Visual In-Context Learning?," *Advances in Neural Information Processing Systems* 36.

36. Shteriyanov, V, R. Dzhusupova, J. Bosch, &H. H.. Olsson, "From Text to Meaning: Semantic Interpretation of Non-Standardized Metadata in Piping and Instrumentation Diagrams," *Computers & Chemical Engineering* 204 (2026): 109436, https://doi.org/10.1016/j.compchemeng.2025.109436.