2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

OPEN-SOURCE VLMS FOR ENGINEERING DRAWING METADATA: QWEN-VL TITLE-BLOCK DETECTION AND INFORMATION EXTRACTION

Mengyan Peng, Angran Cui, Steffen Marx, and Chongjie Kang
Dresden University of Technology, Institute of Concrete Structures
Corresponding author: mengyan.peng@tu-dresden.de

Abstract

Title-blocks in engineering drawings contain essential metadata, but automated interpretation remains difficult because of historical scans and heterogeneous layouts. This paper presents a vision-language model (VLM)-based workflow for title-block detection and structured information extraction using the open-source Qwen2.5-VL backbone. The workflow combines promptable visual grounding on full drawings with schema-constrained extraction from cropped title-block images. The study focuses on three representative fields: scale, drawing name, and route. Results show promising performance for both tasks and indicate that open-source VLMs can provide a simplified alternative to conventional multi-stage pipelines for architecture, engineering, and construction (AEC) document analysis.

Introduction

Engineering drawings remain a central information carrier throughout architecture, engineering, and construction (AEC) workflows. Despite the increasing adoption of digital design tools, large volumes of legacy drawings, often available only as scanned images, are still routinely accessed for maintenance, renovation, compliance checking, and digital asset management. Within these drawings, the title-block plays a critical role, as it encodes essential metadata such as the drawing title, scale, drawing number, creation date, author, and responsible organizations, making reliable extraction important for indexing, version control, and building information modeling (BIM) integration.

However, automated extraction of title-block information remains challenging in practice. Historical scans often exhibit low image quality, heterogeneous layouts, and inconsistent typography. Engineering drawings are visually complex documents in which textual content is embedded in spatially structured layouts and mixed with graphical elements, making reliable parsing difficult (Khan et al., 2025; Wang et al., 2024). Moreover, many older drawings contain handwritten or semi-structured annotations, further complicating recognition. While modern optical character recognition (OCR) systems perform well on clean, machine-printed text, their performance can decline for handwritten engineering-drawing content and under challenging imaging conditions (Peng et al., 2023; Peng et al., 2026).

To address these challenges, existing approaches to engineering drawing information extraction have often relied on rule-based heuristics or multi-stage learning pipelines. Rule-based methods often require manual customization and may become less robust when layouts deviate from expected templates (Gölzhäuser et al., 2023; Faltin et al., 2024). More recent data-driven approaches often adopt detector-based pipelines, sometimes augmented with large language models (LLMs) or vision-language models (VLMs) for post-processing and reasoning (Lombardi et al. 2025). While such hybrid solutions can improve extraction accuracy, they often involve complex multi-model workflows, may depend on proprietary components, and can require additional adaptation when applied to new document styles, particularly in privacy-sensitive AEC environments.

Recent advances in VLMs suggest that such models can align visual content with textual queries and generate structured outputs from complex images (Li et al., 2025; Shinde et al., 2025). However, their use for engineering drawings remains limited, particularly for workflows that employ the same open-source VLM backbone family for both title-block localization and structured information extraction. This gap motivates the VLM-based workflow proposed in this paper.

In this paper, we investigate a VLM-based workflow for title-block detection and structured information extraction using the open-source Qwen2.5-VL backbone. The two tasks are fine-tuned and evaluated separately: detection is formulated as promptable visual grounding on full engineering drawings, while extraction is performed on cropped title-block images. By using the same backbone family in both stages, the workflow avoids a heterogeneous multimodal toolchain and reduces pipeline complexity. Experimental results on bridge engineering drawings show promising performance and motivate further study of open-source VLMs for AEC document understanding.

Related Work

Engineering Drawing Information Extraction

Engineering drawing information extraction has attracted attention for decades, particularly due to the need for digitization and archival management (Scheibel et al., 2021; Kashevnik et al., 2023). Early work focused on

title-block localization and metadata extraction, a key step for indexing and retrieval from technical drawing archives (e.g., Najman et al., 2001). Traditional OCR-based pipelines often combine optical character recognition with layout-dependent heuristics, geometric rules, and manually designed templates to locate and crop title-blocks before field-level text recognition and cleaning (e.g., Ondrejcek et al., 2009). While often effective on standardized layouts, such methods may become less reliable on diverse templates and degraded scans. More recent work has explored hybrid pipelines that integrate detection models with LLMs to improve both localization and structured extraction on complex or noisy drawings (Lombardi et al., 2025).

Multi-model-based Pipelines

To overcome the limitations of rule-based systems, recent work has adopted learning-based pipelines that combine object detection with OCR. In such approaches, detectors such as You Only Look Once (YOLO) (Khanam and Hussain, 2024) or Faster R-CNN (Liu et al., 2017) are used to localize regions of interest, followed by text recognition and post-processing (Faltin et al., 2024). More recently, LLMs and VLMs have been integrated with visual pipelines to improve semantic interpretation and schema mapping (Li et al., 2019; Huang et al., 2022; Lombardi et al., 2025). Although such pipelines can improve accuracy, they often involve multiple interconnected components, which may increase deployment and maintenance effort.

VLMs in Document Understanding

Recent advances in VLMs support visual-text alignment and structured output generation from complex images. Models such as LayoutLM have shown strong performance in document understanding and information extraction by jointly modeling text and visual features (Xu et al., 2020; Xu et al., 2021; Huang et al., 2022). Benchmarks such as DocVQA further validate the usefulness of joint visual-textual representations in complex document tasks (Mathew et al., 2021). More recently, general-purpose VLMs such as GPT-4 and Qwen-VL have extended these capabilities beyond document-specific architectures and support prompt-based visual grounding, multimodal reasoning, and flexible output generation (Achiam et al., 2023; Bai et al., 2025). However, their use for engineering drawings remains limited, particularly for workflows that employ the same open-source VLM backbone family for both title-block localization and structured information extraction. This gap motivates the VLM-based workflow proposed in this paper.

Methodology

Overall Workflow

We propose a VLM-based workflow for title-block detection and structured information extraction using the same Qwen2.5-VL backbone family for both tasks, as shown in Figure 1.

At inference time, the workflow operates in two prompting steps. First, the complete engineering drawing is prompted to localize the title-block via visual grounding. Second, the cropped title-block region is prompted to extract structured information. This design avoids the use of standalone detectors or OCR engines within the proposed workflow.

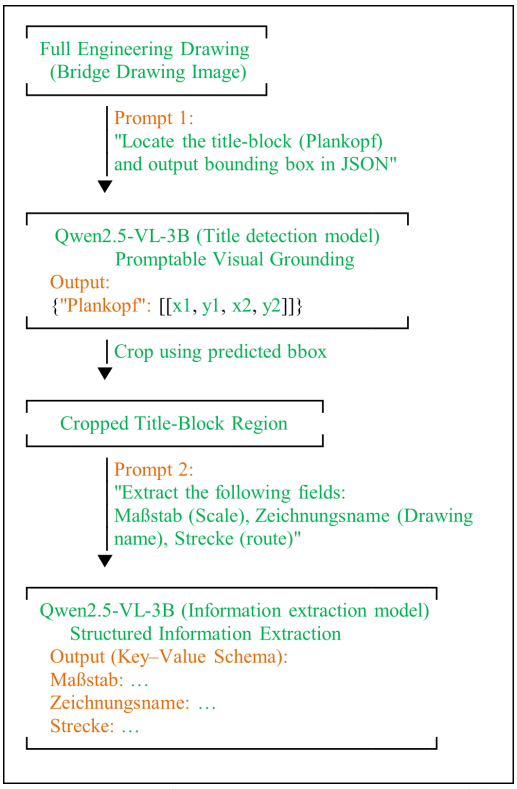

Figure 1: Overview of the proposed VLM-based workflow with separate task-specific stages for detection and structured extraction.

Title-Block Detection via Promptable Visual Grounding

Instead of training a standalone object detector, we formulate title-block localization as a promptable visual grounding task within a VLM. Given a full engineering drawing, the model receives the image together with a natural-language instruction (e.g., “locate the title-block” or “locate the Plankopf”) and directly outputs bounding-box coordinates in a structured JSON format.

This formulation avoids detector-specific training pipelines and allows localization targets to be specified flexibly through language. By performing localization within the same VLM backbone family used for the extraction stage, the predicted bounding boxes can be directly used to crop the title-block region for subsequent processing.

To support multilingual engineering documentation, we employ both English (“title-block”) and German (“Plankopf”) prompts during training and inference and observe consistent grounding performance under both formulations.

Structured Information Extraction

Following title-block detection, the cropped region is re-prompted to perform structured information extraction. We formulate this step as a prompt-based generation task, where the model is instructed to extract metadata fields and return them in a fixed key–value schema. In this study, we focus on three representative attributes commonly present in title-blocks of bridge engineering drawings, namely scale (Maßstab), drawing name (Zeichnungsname), and route (Strecke), as an initial step toward broader metadata extraction.

By explicitly constraining both the target attributes and the output format in the prompt, the model is used to perform OCR-like text recognition and semantic interpretation within a single prompted inference step, without relying on rule-based parsing or post-processing. This design supports structured extraction across layout variations and mixed printed–handwritten content commonly found in historical engineering drawings.

Model Fine-Tuning Strategy

To adapt the pre-trained VLM to the domain of engineering drawings, we apply parameter-efficient fine-tuning based on Low-Rank Adaptation (LoRA). While Qwen2.5-VL-3B-Instruct provides strong general-purpose visual and textual understanding, engineering drawings differ substantially from natural images in terms of visual structure, symbol usage, and domain-specific terminology. Fine-tuning is therefore used to better address this domain gap and improve performance in title-block detection and metadata extraction.

During fine-tuning, the vision tower is frozen to preserve stable low-level visual representations learned during large-scale pre-training. Trainable LoRA adapters, together with the multimodal projection layer, are introduced into the language model component to specialize the model for domain-specific tasks. This strategy significantly reduces the number of trainable parameters and the computational cost, while still supporting practical adaptation under limited annotated data. Although both tasks use the same Qwen2.5-VL-3B-Instruct backbone family, they are fine-tuned separately because they differ in input granularity, output format, and supervision type: detection uses full drawings with box-level supervision, whereas extraction uses cropped title-block images with field-level supervision.

Dataset and Annotation

Dataset Overview

For title-block detection, we use 500 full bridge engineering drawings with box-level title-block annotations for training and validation, comprising 450 training drawings and 50 validation drawings. In addition, an independent test set of 64 full engineering drawings, separate from the training and validation data, is used for evaluation.

For structured information extraction, we use a crop-level development dataset of 467 title-block images for training and validation, comprising 420 training samples and 47 validation samples. In addition, evaluation is conducted on two independent crop-level test sets consisting of 50 machine-printed and 50 handwritten title-block crops, respectively. These test sets are separate from the development data.

The dataset exhibits substantial variation in layout, typography, and image quality. Machine-printed title-blocks often contain denser information, tabular structures, and standardized fonts, whereas handwritten title-blocks typically have simpler layouts but introduce challenges such as writing-style variability, fading ink, and scanning artifacts. Together, these categories support evaluation across both contemporary and legacy engineering documents.

Annotation Schema

Each sample is annotated at two complementary levels for detection and information extraction.

```
{
 "image": "<image>",
 "conversation": [
  {
   "role": "user",
   "content": "Locate the Plankopf (title-block) in this
image and output its bounding-box coordinates in JSON
format."
  },
  {
   "role": "assistant",
   "content": "{\"Plankopf\": [[965, 1090, 1268, 1262]]}"
  }
 ]
}
```

Figure 2: Example annotation of title-block detection.

```
{
 "image": "<image>",
 "conversation": [
  {
   "role": "user",
   "content": "Please extract the following information
from the image and output it in the specified
format:\nMaßstab:\nZeichnungsname:\nStrecke:"
  },
  {
   "role": "assistant",
   "content": "Maßstab: 1:50\nZeichnungsname:
Widerlager Achse 20\nStrecke: Hamburg – Berlin, km
0,628"
  }
 ]
}
```

Figure 3: Example annotation of structured information extraction.

First, box-level annotations define the spatial extent of the title-block within the full engineering drawing, as illustrated in Figure 2. For this task, the model is prompted to locate the title-block, and the corresponding ground-truth response provides rectangular bounding-box coordinates in JSON format. These annotations are used

to supervise and evaluate title-block localization performance.

Second, field-level annotations specify the ground-truth values for three target metadata attributes, namely scale, drawing name, and route within the title-block, as illustrated in Figure 3. Given a cropped title-block image, the model is instructed to extract these attributes and return them in a fixed key–value schema.

By representing both detection and extraction annotations as conversational image-text pairs with structured outputs, the dataset enables direct supervision of heterogeneous tasks within the same VLM backbone family. This design reduces the need for task-specific post-processing rules or layout assumptions and facilitates the combined use of the two models at inference time.

Experiments and Results

Experimental Setup

We evaluate the proposed VLM-based workflow on two related tasks: title-block detection and structured information extraction. Both tasks use the same Qwen2.5-VL-3B-Instruct backbone family but are fine-tuned separately because they differ in input granularity and output targets. During preliminary development, we explored a small set of LoRA configurations and selected the final settings based on stable convergence and validation performance. For title-block detection, the final LoRA configuration used rank 32, scaling factor 16, and dropout 0.05. For structured information extraction, the final configuration used rank 32, scaling factor 64, and dropout 0.05. The model was optimized using AdamW with an initial learning rate of 1×10^{-4}, a cosine learning-rate scheduler, and 70 warmup steps. Mixed-precision training with bfloat16 (bf16) was used to reduce memory consumption. For title-block detection, the model was fine-tuned for 8 epochs, and the best-performing checkpoint was selected based on validation loss. For structured information extraction, the final checkpoint was selected after 4 epochs, consistent with the observed validation-loss minimum around the fourth epoch.

Baselines

For title-block detection, we compare the fine-tuned Qwen2.5-VL-3B-Instruct model against two references: (1) YOLOv11-medium as a conventional detector baseline and (2) zero-shot Qwen2.5-VL-3B-Instruct without task-specific fine-tuning. All models are evaluated on the same 64-drawing test set using precision, recall, F1, and mean IoU at an IoU threshold of 0.5.

For structured information extraction, we compare the fine-tuned Qwen2.5-VL-3B-Instruct model against the corresponding zero-shot configuration without task-specific fine-tuning. Both models are evaluated on the same crop-level test sets of machine-printed and handwritten title-block images using Word Accuracy (WA) and Character Accuracy (CA). Per-field results are additionally reported for Maßstab, Zeichnungsname, and Strecke.

Title-Block Detection

We evaluate title-block detection using precision, recall, F1, and mean Intersection over Union (IoU) at an IoU threshold of 0.5. Predictions above the threshold are counted as true positives (TP), lower-overlap detections are treated as false positives (FP), and missed title blocks are treated as false negatives (FN).

Table 1: Comparative Performance for Title-Block Localization

Metric	Fine-tuned Qwen2.5-VL-3B	YOLOv11 -medium	Zero-shot Qwen2.5-VL-3B
Precision	93.8%	91.0%	0.0%
Recall	93.8%	95.3%	0.0%
F1	93.8%	93.1%	0.0%
Mean IoU	0.871	0.951	---
TP	60	61	0
FP	4	6	0
FN	4	3	64
Output Validity Rate	100%	98.4%	100%

Table 1 compares the fine-tuned Qwen2.5-VL-3B-Instruct model with the YOLOv11-medium baseline and the zero-shot Qwen2.5-VL-3B-Instruct baseline on the 64-drawing test set. The fine-tuned Qwen model achieves 93.8% precision and 93.8% recall, with an F1 score of 93.8% and a mean IoU of 0.871. This performance is comparable to YOLOv11-medium (F1=93.1%), although YOLO yields a higher mean IoU (0.951). YOLO produced a small number of duplicate or mislocalized predictions, which explains why the number of detections exceeded the number of test images. In contrast, the zero-shot Qwen model does not localize title blocks successfully in this setting, despite producing syntactically valid structured outputs. This highlights the importance of domain-specific fine-tuning for engineering drawings.

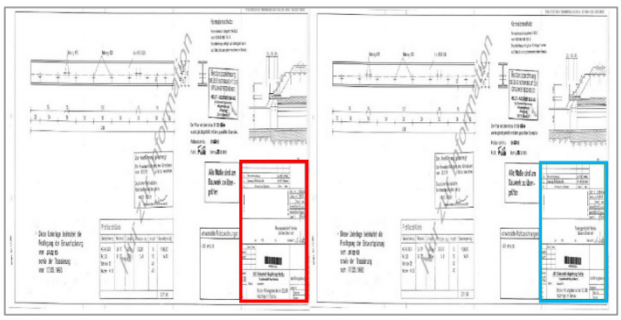

Figure 4: Representative result of title-block detection. Comparison between ground-truth annotation (red) and model prediction (blue).

For comparison, YOLO detections were converted into the same structured output format. The fine-tuned Qwen achieved a 100% output validity rate, meaning that valid structured outputs were generated for all test samples. As shown in Figure 4, the predicted bounding box closely

aligns with the ground-truth annotation in a representative drawing. Most detection failures occur in cases with unconventional title-block placement, severe visual degradation, or ambiguous boundaries between the title block and adjacent tabular regions. In a smaller number of cases, visually similar tabular areas are mistaken for title blocks, or the predicted box only partially covers the true title-block extent, leading to IoU values below the acceptance threshold.

Structured Information Extraction

To evaluate structured information extraction, we report both Word Accuracy and Character Accuracy, which capture complementary aspects of extraction quality. The metrics are computed directly on the extracted key-value outputs by comparing each predicted field value with the corresponding ground-truth value, without additional post-processing, normalization, or special handling of empty fields. Word Accuracy reflects semantic-level correctness and sensitivity to word-boundary errors, while Character Accuracy provides a more fine-grained assessment of character-level robustness.

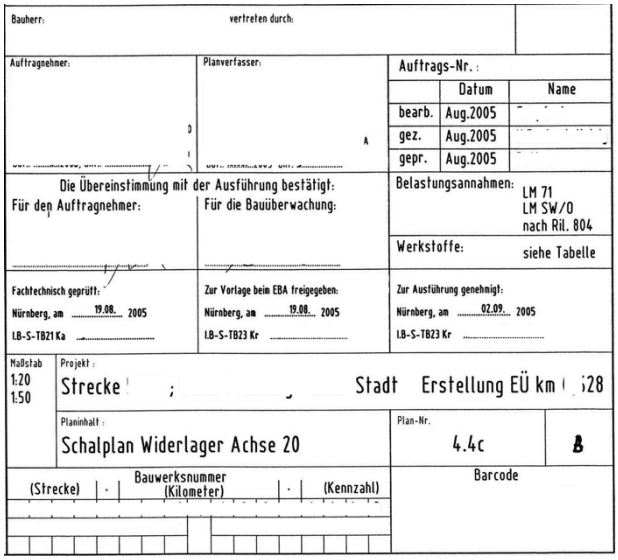

Figure 5: Example of machine-printed title-block. (Sensitive project-specific information has been anonymized for confidentiality.)

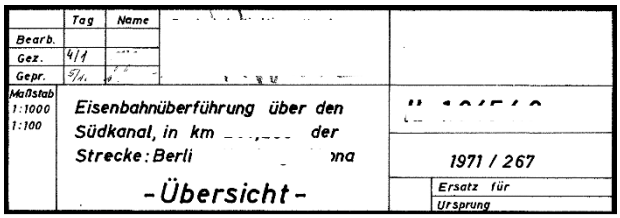

Figure 6: Example of predominantly handwritten title-block. (Sensitive project-specific information has been anonymized for confidentiality.)

We evaluate extraction performance on two independent crop-level test sets constructed from ground-truth title-block crops, as described in the Dataset Overview. One set contains 50 machine-printed title-blocks with dense layouts (Figure 5), and the other contains 50 predominantly handwritten title-blocks with relatively simple layouts but diverse handwriting styles (Figure 6). These results reflect crop-level extraction performance and are reported independently of the upstream detection stage. All results reported in this section correspond to the final configuration selected based on validation performance, as described in the Experimental Setup.

On the machine-printed title-block test set, the fine-tuned model achieves a WA of 0.78 and a CA of 0.76. On handwritten title-blocks, higher performance is observed, with WA reaching 0.83 and CA reaching 0.86. These results indicate promising extraction accuracy for both modern and historical title-block images.

Table 2: Comparative performance for structured information extraction on handwritten title-block crops (Field 1: Maßstab; Field 2: Zeichnungsname; Field 3: Strecke).

Metric	Fine-tuned Qwen2.5-VL-3B	Zero-shot Qwen2.5-VL-3B
Field 1 WA	0.9474	0.6023
Field 1 CA	0.9495	0.5865
Field 2 WA	0.7216	0.1868
Field 2 CA	0.8479	0.4320
Field 3 WA	0.8545	0.2836
Field 3 CA	0.8671	0.3536

To assess the effect of domain-specific adaptation in more detail, we further compare the fine-tuned Qwen model with the corresponding zero-shot configuration at the field level. Tables 2 and 3 report per-field WA and CA for the three target attributes (Maßstab, Zeichnungsname, and Strecke) on handwritten and machine-printed title-block crops, respectively.

Table 3: Comparative performance for structured information extraction on machine-printed title-block crops (Field 1: Maßstab; Field 2: Zeichnungsname; Field 3: Strecke).

Metric	Fine-tuned Qwen2.5-VL-3B	Zero-shot Qwen2.5-VL-3B
Field 1 WA	0.9900	0.5721
Field 1 CA	0.9892	0.5510
Field 2 WA	0.7132	0.1716
Field 2 CA	0.7510	0.3217
Field 3 WA	0.7398	0.2520
Field 3 CA	0.7117	0.3061

Per-field analysis reveals a clear difference in extraction difficulty across the three target attributes. Maßstab is consistently the easiest field, achieving the highest WA and CA in both handwritten and machine-printed title-blocks. This is likely due to its short and highly regular format. Zeichnungsname is the most challenging field, especially at the word level, which can be attributed to its greater length, frequent line breaks, and higher sensitivity to surrounding layout complexity. Strecke shows intermediate difficulty, although its performance decreases substantially in machine-printed title-blocks,

suggesting that dense tabular layouts and closely positioned neighboring text increase ambiguity in field boundaries. Across all three fields, the fine-tuned model substantially outperforms the corresponding zero-shot baseline, highlighting the importance of domain-specific adaptation for engineering drawings.

The higher overall extraction accuracy on handwritten title-blocks is consistent with the per-field comparison. Although Maßstab achieves slightly better results in machine-printed title-blocks, both Zeichnungsname and Strecke are extracted more accurately from handwritten samples. This indicates that, for these semantically richer fields, layout simplicity and reduced visual clutter may be more important than text regularity alone. By contrast, machine-printed title-blocks often contain denser tabular structures and more closely positioned text elements, which increase ambiguity in field boundaries and make extraction more difficult, especially for drawing names and route-related content.

Maßstab (scale): 1:20, 1:50
Zeichnungsname (drawing name): Erstellung EÜ Schalplan Widerlager Achse 20
Strecke (route): …Stadt km …28

Figure 7: Representative result of structured information extraction for a machine-printed title-block, corresponding to the title-block shown in Figure 5.

Maßstab (scale): 1:1000, 1:100
Zeichnungsname (drawing name): Eisenbahnüberführung über den Südkanal Übersicht
Strecke (route): Berli…ona, km …

Figure 8: Representative result of structured information extraction for a predominantly handwritten title-block, corresponding to the title-block shown in Figure 6.

Figures 7 and 8 present representative qualitative examples of structured information extraction on machine-printed and handwritten title-blocks, respectively. In each example, the model correctly identifies and extracts multiple metadata fields, including scale, drawing name, and route, despite variations in layout, typography, and content density.

Overall, the results indicate that a fine-tuned VLM can support OCR-like text recognition and structured information extraction for title-block content without relying on external OCR engines or rule-based post-processing. This highlights the practical potential of the proposed VLM-based workflow for engineering document analysis.

Small-Scale Pipeline Evaluation

We additionally evaluated the workflow on 20 real full bridge engineering drawings on an NVIDIA RTX A5000 GPU with 24 GB of memory. The set reflects challenging and practically relevant legacy-archive conditions, including dark, speckle-noisy scans and mixed machine-printed layout fields with handwritten technical annotations. Title-block detection achieved Precision/Recall/F1 = 85%/85%/85%, with a mean IoU of 0.633. On the corresponding 20 title-block crops, extraction from ground-truth crops achieved a WA of 0.6458 and a CA of 0.7475, whereas extraction from Qwen-predicted crops achieved a WA of 0.6771 and a CA of 0.7712. Per-field trends remained consistent with the main benchmark: Maßstab was the easiest field, Zeichnungsname the most difficult, and Strecke intermediate. Notably, extraction from Qwen-predicted crops did not lead to a performance drop in this small-scale evaluation and even showed slightly higher WA and CA, suggesting that the predicted boxes were sufficiently accurate for downstream extraction and may, in some cases, reduce surrounding visual clutter. The average runtime was 2.97 s per drawing for title-block detection and 8.91 s per title-block crop for structured information extraction, with peak allocated GPU memory below 10 GB in all settings. These results indicate that the workflow is practically feasible for offline archive-style processing on a 24 GB-class GPU.

Discussion and Limitations

The experimental results indicate that the same open-source VLM backbone family can be used for both title-block detection and structured information extraction on engineering drawings. Unlike conventional detector-based pipelines, the proposed VLM-only workflow uses the same VLM backbone for both localization and interpretation, reducing system complexity while showing promising performance across heterogeneous drawing styles. In the present study, the two tasks are fine-tuned separately and combined at inference time.

One factor that may contribute to this behavior lies in the nature of title-blocks themselves. Although title-blocks exhibit substantial variation in layout, typography, and language across organizations and time periods, they are typically semantically well-defined regions that encode standardized metadata. Promptable visual grounding may help the model leverage semantic cues rather than relying solely on geometric or appearance-based patterns, which may contribute to localization performance even under unconventional layouts or degraded scan quality. This differs from traditional object detectors, which may require additional adaptation under distribution shifts or limited training samples.

Another advantage of the proposed approach is the conceptual consistency between detection and extraction through the same VLM backbone and prompt-based interaction paradigm, which may reduce intermediate conversions and error propagation in multi-stage workflows.

The information extraction results further indicate that the model shows promising performance across different title-block styles. Higher accuracy is observed on handwritten title-blocks, which can be attributed to their typically simpler layouts and lower information density compared with modern machine-printed title-blocks. In contrast, printed title-blocks often contain dense tables, multiple scale entries, and additional graphical elements, increasing extraction complexity. These observations

suggest that layout complexity may be an important challenge, in addition to handwriting, for structured extraction in engineering drawings.

From a practical perspective, the proposed VLM-only workflow offers several potential advantages for deployment in AEC environments. The use of a single open-source model reduces engineering overhead, improves reproducibility, and facilitates on-premise deployment, which may be beneficial in contexts involving privacy, regulatory compliance, and long-term maintainability. These properties may make the approach relevant for real-world engineering archives, where heterogeneous document collections and limited annotation budgets are common.

Despite its promising performance, the proposed workflow has several limitations. First, the current study focuses on three representative metadata fields, namely scale, drawing name, and route, which, while practically relevant, do not cover the full range of information present in title-blocks. Extending the schema to additional fields, such as drawing numbers, revision indices, dates, and responsible organizations would further improve applicability.

Second, although promptable visual grounding performs well on most drawings, detection errors still occur in edge cases involving unconventional layouts, severe visual degradation, or ambiguous title-block boundaries. Addressing such cases may require additional contextual prompts, multi-step reasoning, or lightweight layout priors.

Third, the workflow does not explicitly handle non-textual elements commonly found in engineering drawings, such as stamps, signatures, symbols, or dimension annotations. Integrating symbol understanding or multimodal reasoning over graphical primitives remains an open challenge.

Finally, the current dataset is limited to a bridge-engineering context, and broader generalization to other engineering domains, regions, and historical conventions remains to be validated. While the small-scale pipeline evaluation (n = 20) provides encouraging results, systematic validation on a larger full-drawing test set is still needed. Future work will therefore examine the complete workflow on full drawings using predicted title-block localization, extend the approach to other drawing domains such as buildings, Mechanical, Electrical, and Plumbing (MEP) systems, and other infrastructure types, and investigate broader baseline comparisons, including detector-plus-OCR pipelines or document-specialized models. Extending the workflow to new domains will likely require additional domain-specific data at both stages, including full-drawing box annotations for title-block localization and crop-level key-value annotations for structured extraction, as well as adaptation of the target metadata schema, prompt design, and task-specific fine-tuning. Larger-batch throughput analysis beyond the current single-sample sequential runtime setting also remains future work.

Conclusions

This paper shows that a single open-source VLM backbone family can provide a promising basis for unified title-block detection and structured information extraction in engineering drawings. By formulating title-block detection as a promptable visual grounding task and performing metadata extraction through prompt-based generation, the proposed approach combines detection and interpretation within a VLM-based workflow built on the same backbone family without relying on standalone detectors or external OCR engines.

The results across modern and historical drawings indicate that VLM-based approaches show promising performance across heterogeneous layouts and visual conditions, suggesting their potential for broader use in engineering document analysis workflows.

Overall, this work highlights the potential of open-source VLMs as a simplified and reproducible alternative to conventional multi-model pipelines for engineering drawing understanding. It also offers a useful starting point for future research on richer metadata schemas, deeper semantic understanding, and scalable digitization of engineering archives using open-source VLMs.

Data and Code Availability

The study uses the publicly available Qwen2.5-VL model (https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct). The base architecture was not modified. The engineering drawing dataset is confidential but may be shared upon reasonable request for non-commercial research purposes, subject to institutional approval.

Acknowledgments

The authors acknowledge the support of the German Federal Ministry of Transport (BMV) within the project HyBridGen (Project No. 01FV2067A).

Declaration of generative AI and AI-assisted technologies in the Writing Process

During manuscript preparation, the authors used ChatGPT and DeepL under human oversight solely to improve language quality. The authors reviewed and revised the content as needed and take full responsibility for the final version.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., ... & Lin, J. (2025). Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923. https://doi.org/10.48550/arXiv.2502.13923.

Faltin, B., Schönfelder, P., Gann, D., & König, M. (2024). Reconstructing as-built beam bridge geometry from construction drawings using deep learning-based symbol pose estimation. Advanced Engineering

Informatics, 62, 102808. https://doi.org/10.1016/j.aei.2024.102808.

Gölzhäuser, P., Peng, M., Jäkel, J. I., Klemt-Albert, K., & Marx, S. (2023). Approach to generate a simple semantic data model from 2D bridge plans using AI-based Text Recognition. In Proceedings of the 33rd European Safety and Reliability Conference (Vol. 3, pp. 2023-07). https://doi.org/10.3850/978-981-18-8071-1_P424-cd.

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022, October). LayoutLMv3: Pre-training for document ai with unified text and image masking. In Proceedings of the 30th ACM international conference on multimedia (pp. 4083-4091). https://doi.org/10.1145/3503161.3548112.

Kashevnik, A., Shilov, N., Teslya, N., Hasan, F., Kitenko, A., Dukareva, V., ... & Blokhin, D. (2023). An approach to engineering drawing organization: Title block detection and processing. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3244603.

Khan, M. T., Yong, Z., Chen, L., Tan, J. M., Feng, W., & Moon, S. K. (2025). Automated Parsing of Engineering Drawings for Structured Information Extraction Using a Fine-tuned Document Understanding Transformer. arXiv preprint arXiv:2505.01530. https://doi.org/10.48550/arXiv.2505.01530.

Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725. https://doi.org/10.48550/arXiv.2410.17725.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557. https://doi.org/10.48550/arXiv.1908.03557.

Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., & Shi, G. (2025). A survey of state of the art large vision language models: Benchmark evaluations and challenges. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 1587-1606). https://doi.org/10.48550/arXiv.2501.02189.

Liu, B., Zhao, W., & Sun, Q. (2017, October). Study of object detection based on Faster R-CNN. In 2017 Chinese automation congress (CAC) (pp. 6233-6236). IEEE. https://doi.org/10.1109/CAC.2017.8243900.

Lombardi, A., Duan, L., Elnagar, A., Zaalouk, A., Ismail, K., & Vakaj, E. (2025). Title block detection and information extraction for enhanced building drawings search. arXiv preprint arXiv:2504.08645. https://doi.org/10.48550/arXiv.2504.08645.

Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). DocVQA: A dataset for VQA on document images. In Proceedings of the IEEE/CVF winter conference on Applications of computer vision (pp. 2200-2209). https://doi.org/10.48550/arXiv.2007.00398.

Najman, L., Gibot, O., & Barbey, M. (2001, September). Automatic title block location in technical drawings. In Proceedings of 4th IAPR International Workshop on Graphics Recognition, Kingston, Ontario (Canada) (pp. 19-26).

Ondrejcek, M., Kastner, J., Kooper, R., & Bajcsy, P. (2009). Information extraction from scanned engineering drawings. Image Spatial Analysis Group, National Center for Supercomputing Applications, NCSA-ISDA09-001.

Peng, M., Kang, C., & Marx, S. (2023). Text Recognition for 2D Bridge Plans Using OCR-Algorithms. ce/papers, 6(5), 661-666. https://doi.org/10.1002/cepa.2077.

Peng, M., Qian, H., Marx, S., & Kang, C. (2026). Optimizing 2D Bridge Engineering Drawing Digitization: A Comparative Study of Text Recognition Tools and Development of Lightweight Post-Recognition Structured Information Extraction Methods. Results in Engineering, 110186. https://doi.org/10.1016/j.rineng.2026.110186.

Scheibel, B., Mangler, J., & Rinderle-Ma, S. (2021). Extraction of dimension requirements from engineering drawings for supporting quality control in production processes. Computers in Industry, 129, 103442. https://doi.org/10.1016/j.compind.2021.103442.

Shinde, G., Ravi, A., Dey, E., Sakib, S., Rampure, M., & Roy, N. (2025). A Survey on Efficient Vision-Language Models. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(3), e70036. https://doi.org/10.1002/widm.70036.

Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., ... & Liu, X. (2024, August). DocLLM: A layout-aware generative language model for multimodal document understanding. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8529-8548). https://doi.org/10.18653/v1/2024.acl-long.463.

Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., ... & Zhou, L. (2021, August). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 2579-2591). https://doi.org/10.18653/v1/2021.acl-long.201.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020, August). LayoutLM: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 1192-1200). https://doi.org/10.1145/3394486.3403172.