From Engineering Drawings to Assembly Instructions: A Vision and Language Model Approach

Shokhikha Amalana Murdivien[1], Kyungwan Choi[2], Minji Kim[1], and Jumyung Um[1,2,3*]

[1] *Department of Artificial Intelligence, Kyung Hee University,*
Yongin-si, Gyeonggi-do, 17104, Republic of Korea
amalanadivien@khu.ac.kr
minjikim2000@khu.ac.kr

[2] *Department of Metaverse, Kyung Hee University,*
Yongin-si, Gyeonggi-do, 17104, Republic of Korea
choikyungwan7@khu.ac.kr

[3] *Department of Industrial and Management Systems Engineering, Kyung Hee University,*
Yongin-si, Gyeonggi-do, 17104, Republic of Korea
*jayum@khu.ac.kr**

Abstract

Engineering drawings such as CAD draft sheets are widely used in manufacturing to document product structure, part geometry, and dimensional specifications. While these documents contain valuable information, they are not typically organized to support step-by-step assembly tasks, which can present challenges for non-expert technicians during installation, maintenance, or repair. This paper presents a system that automatically generates structured and human-readable assembly instructions from CAD drafts by combining a vision model, an OCR model, and a language model. The vision model, trained on a constructed synthetic dataset, was able to detect mechanical components with an average precision score of 95.2% on real CAD sheets, while the OCR model successfully extracted dimensional information. These outputs, together with existing description text, were processed by a language model to produce clear and interpretable assembly steps. A synthetic dataset was used to train the vision model, addressing the lack of publicly available CAD annotations. The results demonstrate that the proposed system improves the interpretability and usability of engineering documentation in assembly-related tasks.

1. Introduction

In modern manufacturing environments, consistent and accurate production processes depend not only on advanced machinery and automation but also on reliable technical documentation. Among various forms of documentation, assembly instructions serve as a key resource for guiding technicians through product construction and installation tasks. As production systems become more complex, the demand for clear, interpretable instructions increases, particularly in settings where inexperienced or cross-functional personnel are involved. At the same time, a shortage of skilled maintenance and assembly technicians has been reported in many industries, due in part to the increasing complexity of tasks and the limited supply of trained labor (Bocák, Holubek, & Tirian, 2022). These factors highlight the need for more accessible and structured documentation to support both expert and non-expert technicians.

Engineering drawings, such as Computer-Aided Design (CAD) draft sheets, are commonly used to convey product structure, part geometry, and dimensional specifications. In most industrial settings, CAD models and their corresponding draft sheets are readily available for each product, as part of standard design and production workflows. While these documents contain essential technical information, they are not designed to provide step-by-step guidance for assembly. Interpreting CAD drafts often requires domain expertise, and the documents typically lack information such as tool usage or the selection of compatible hardware. For less experienced users, this can result in slower assembly processes or increased reliance on external support.

This paper presents a system that automatically generates structured and human-readable assembly instructions from

CAD draft sheets by combining a vision model, an Optical Character Recognition (OCR) model, and a large language model (LLM). The vision model detects mechanical components in the drawings, while the OCR model extracts dimensional data. These outputs, along with existing textual descriptions, are processed by a language model to generate step-by-step assembly procedures. To overcome the lack of annotated data for training, a synthetic dataset was generated using available CAD models. By combining these components, the proposed system improves the usability of engineering documentation and supports more accessible and consistent assembly workflows.

The rest of this paper is structured as follows: Section 2 reviews prior research on AI-based documentation and instruction generation. Section 3 details the proposed system architecture and implementation. Section 4 presents experimental results and analysis, and Section 5 concludes the paper with a discussion of findings, limitations, and future directions.

2. LITERATURE REVIEW

Assembly documentation supports not only initial product build, but also maintenance, repair, and remanufacturing, where correct sequencing and component handling directly affect reliability, safety, and service quality. In industrial practice, technicians rely heavily on engineering drawings such as CAD draft sheets to understand geometry, part relationships, and dimensional specifications. Yet these drawings are not organized as step-by-step procedures, and they often omit explicit tool usage, hardware selection, and assembly order, making them difficult for less-experienced personnel to interpret efficiently. These limitations have motivated research on transforming unstructured engineering documents into task-oriented assembly guidance through computer vision, OCR, and language model.

Recent advances show steady progress in the digitization and interpretation of engineering drawings. The work on automatic digitization of raster engineering drawings integrates detection and text recognition for legacy mechanical drawings, enabling structured extraction from 2D documents (Maupou et al., 2024). Dedicated OCR pipelines for engineering drawings further demonstrate robust text detection and table extraction tailored to technical layouts (Villena Toro, Wiberg, & Tarkian, 2023). Beyond text, component- and topology-level understanding has improved through graph- and learning-based representations for 2D drawings (Zhang et al., 2023) and few-shot symbol detection in CAD-like diagrams (Jamieson, Elyan, & Moreno-García, 2024). In parallel, visually rich document understanding (VDU) has benefited from multimodal transformers such as LayoutLMv3 (Huang, Lv, Cui, Lu, & Wei, 2022) and OCR-free approaches like Donut (Kim et al., 2021), while DocVQA benchmarks have established multi-modal evaluation protocols for reasoning over documents that combine layout, text, and figures (Tito, Mathew, Jawahar, Valveny, & Karatzas, 2021).

For perception front-ends, one-stage object detectors such as YOLOv7 provide accurate, real-time detection suitable for structured technical imagery when training data are sufficient (Wang, Bochkovskiy, & Liao, 2023). However, assembling large annotated datasets for engineering drawings remains costly. To mitigate this, recent studies have adopted synthetic data generated from 3D CAD, domain randomization, and rendering pipelines that reduce labeling effort while maintaining strong sim-to-real generalization in manufacturing contexts (Pasanisi, Rota, Ermidoro, & Fasanotti, 2023; Schraml & Notni, 2024; Monnet, Petrovic, & Herfs, 2024; Li, 2023). Finally, integrating perception outputs with LLMs is emerging as a viable route to produce coherent, human-readable procedural instructions from technical sources, including text-to-instruction generation with ontology/RAG support and assessments of LLM-based assistants in manual assembly settings (Holvoet, van Bekkum, & de Vries, 2024; Colabianchi, Costantino, & Sabetta, 2024; Duan et al., 2025).

While these studies demonstrate significant progress in digitizing technical documents and generating procedural content, most still treat visual and textual modalities independently, depend on domain-specific annotated datasets, or assume well-structured instruction templates. Few works tackle the combined challenges of interpreting unstructured CAD draft sheets, where geometric information, dimensional annotations, and symbolic text coexist in dense and ambiguous layouts, and converting them into coherent, human-readable assembly procedures without any manual labeling.

In contrast, the proposed approach introduces a unified perception language framework that not only integrates vision, OCR, and language models, but also establishes a cross-modal reasoning process in which the LLM reconciles inconsistencies between detected geometry and extracted text. Moreover, synthetic datasets generated from 3D CAD renderings serve not merely as data augmentation but as a domain bridging mechanism to overcome the representation gap between photorealistic CAD images and abstract engineering drafts. This combination enables scalable, annotation free generation of interpretable assembly instructions, a direction largely unexplored in prior work.

3. METHODOLOGY

3.1. System Architecture

The proposed system integrates a vision model, an OCR model, and an LLM to convert CAD draft sheets into complete, human-readable assembly instructions. The overall workflow is illustrated in Figure 1. The CAD draft sheets contain detailed geometric depictions of each mechanical component, along with dimensional specifications, tool types, and a

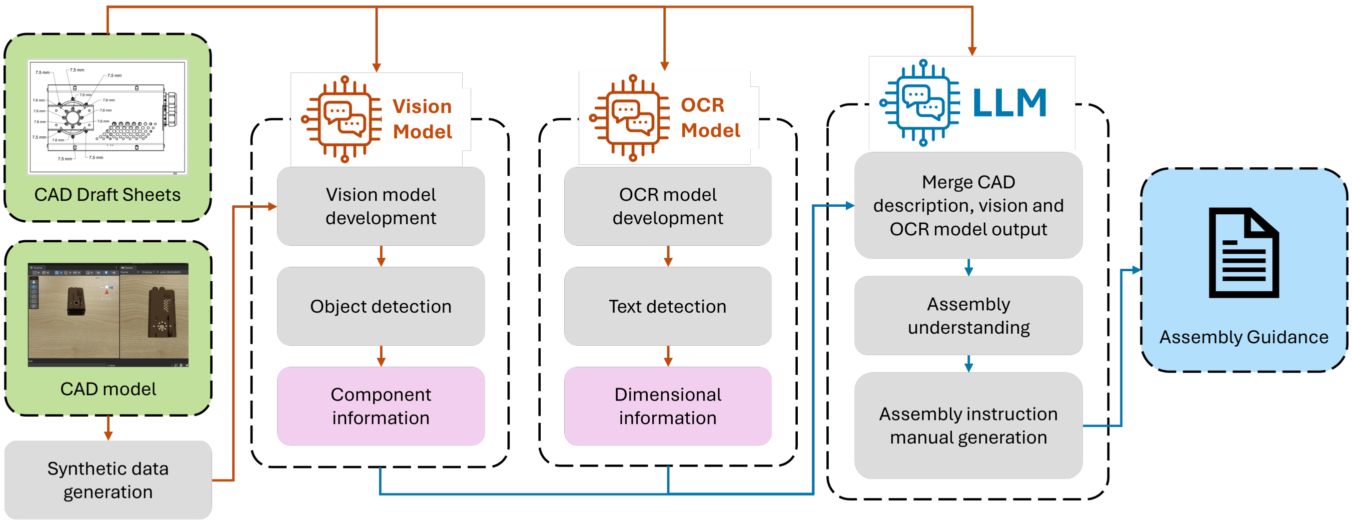


Figure 1. Proposed system architecture

brief procedural description outlining the intended sequence of assembly. While this procedural description is concise and not directly executable as an instruction manual, it serves as a valuable reference for determining the correct order of operations.

The CAD drafts are processed in parallel by two detection modules. The vision model identifies each mechanical component, while the OCR model detects and reads text embedded in the drafts, capturing screw specifications and tool information. The procedural description present in the draft is also extracted at this stage and preserved as a sequencing reference for later use.

The outputs from these two models, together with the extracted procedural description, are processed by the LLM, which is guided by a system prompt to produce clear, document-style assembly instructions. The LLM merges the visual and textual information, applies logical reasoning to ensure internal consistency, and resolves incomplete or imperfect detections by inferring plausible details from the available context. By combining component recognition, dimensional and tooling details, and sequence cues from the original draft, the system is able to generate coherent, step-by-step assembly instructions that maintain technical accuracy while improving clarity and usability for human operators.

3.2. Data Acquisition and Preprocessing

The development of the proposed system required the preparation of CAD draft sheets and corresponding training data for the vision model. The CAD drafts were designed using conventional engineering drawing formats, then exported as PDFs and converted to high-resolution images for processing. Since publicly available datasets for this domain are scarce, synthetic data generation was employed to create a sufficiently large and diverse dataset.

Synthetic datasets were generated using Unity Perception. 3D CAD models of basic mechanical components were arranged in randomized positions and orientations to simulate realistic part layouts. Lighting conditions, rotation, and placement of objects were varied to introduce visual diversity. The illumination environment was adjustable, allowing control over the position, angle, and intensity of the light source as desired. Object orientation and placement were also configurable. Object positions were randomly assigned within a predefined spatial range (x, y, z), where each coordinate was uniformly sampled within the specified bounds $[x_{\min}, x_{\max}]$, $[y_{\min}, y_{\max}]$, $[z_{\min}, z_{\max}]$. The rotation of each object was randomly generated over the full range of $[0°, 360°]$ for each axis (x, y, z). The total number of objects in the dataset was 18, consisting of four distinct categories: two types of screws, one base, and one top part attached to the base. An example of the rendered scene is shown in Figure **??**. Unity Perception automatically provided object location metadata for each rendered image, eliminating the need for manual annotation and enabling efficient preparation of training labels.

3.3. Model Training and Implementation

3.3.1. Vision Model for Component Detection

The vision module integrates two complementary detection processes: mechanical component recognition and textual information extraction. Both operate on the same CAD draft input, producing complete visual and textual data for instruction generation.

For component recognition, YOLOv11 was selected due to its favorable trade-off between accuracy, speed, and computational efficiency. YOLOv11 achieves higher accuracy with

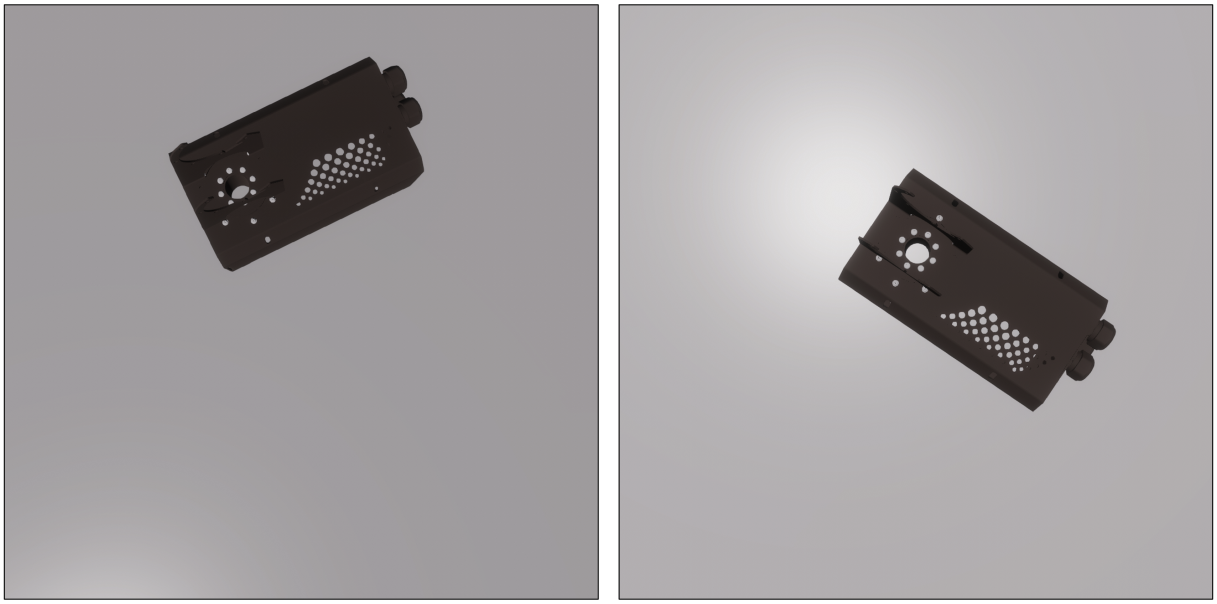


Figure 2. Generated image example using Unity Perception.

fewer parameters and faster inference by introducing efficient modules that enhance feature extraction and attention (Jocher & Qiu, 2024). The model was trained on the synthetic dataset described in Section 3.2, enabling specialization for engineering drawings rather than natural images. At inference, the model processes each draft as a high-resolution image and outputs a list of detected components.

In parallel, textual information is extracted using EasyOCR, chosen for its robustness in handling varied orientations and font styles common in CAD drafts. This stage retrieves dimensional and tool information for screws to ensure accurate integration into the final instructions. The OCR outputs are returned as structured text entries.

The results from component recognition, OCR extraction, and procedural description parsing are combined into a single structured dataset, which serves as the input for the LLM.

3.3.2. Large Language Model for Assembly Instruction Generation

The final stage uses the open-source `gpt-oss:20b` model via Ollama to convert the combined outputs into coherent, human-readable assembly instructions. Inputs include the detected component list, dimensional and tool details, and the brief procedural description from the CAD draft, which serves as a sequencing guide.

A task-specific system prompt directs the LLM to maintain the correct assembly order and associate screw specifications and driver types based on OCR data. When detections are incomplete or inconsistent, the LLM applies contextual reasoning to reconcile discrepancies, ensuring technical accuracy. The output is a step-by-step narrative that preserves the precision of engineering documentation while improving clarity for human operators.

4. EXPERIMENT AND RESULTS

4.1. Experimental Setup

A case study was conducted using the base part of the Igus robolink robot system to evaluate the proposed method. The CAD drafts for this component were created in a conventional engineering drawing format, exported as PDFs, and converted to high-resolution image pages for processing.

The synthetic dataset generated with Unity Perception was split into training, validation, and testing sets to support model development and performance evaluation. The YOLOv11 component detection model was trained from scratch using this dataset, with input drafts resized to the network's native resolution and default hyperparameters applied.

All experiments were carried out on a workstation equipped with an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), an Intel Core i7 processor, and 40 GB of RAM, running the Windows 10 operating system. The implementation was based on PyTorch 2.5.1 with CUDA 12.1 support.

4.2. Results and Analysis

4.2.1. Result of Component Detection Model

The vision model was trained on a synthetic dataset of 40,000 images generated using Unity Perception. Its performance was assessed using standard object detection metrics: precision (the ratio of correct detections to all detections), recall (the ratio of correct detections to all actual objects), and mean average precision (mAP), which summarizes detection accuracy across classes and confidence thresholds.

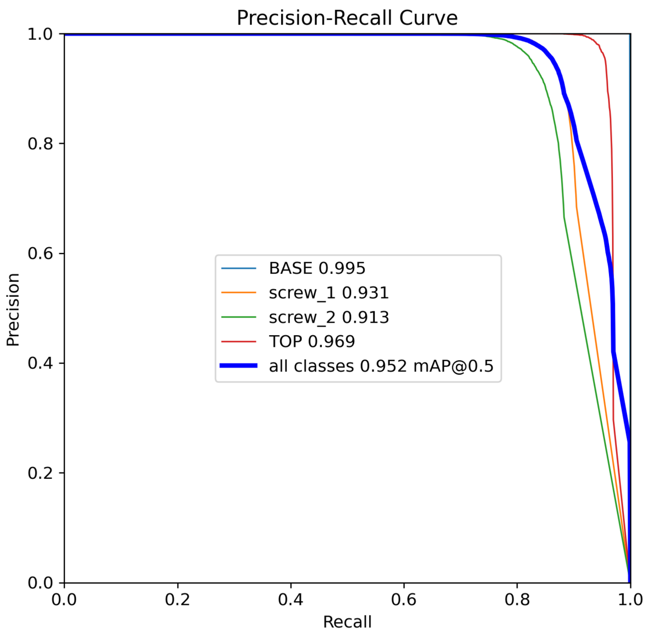


Figure 3. Precision-Recall curves and mAP of the vision model. The curves show that the model maintains high precision across most recall levels, indicating stable and reliable detection performance for all component classes.

Figure 3 shows the precision–recall curves for each class, along with the mAP at IoU threshold 0.5. The model achieved an overall mAP of 0.952, with individual average precision scores of 0.995 for *BASE*, 0.969 for *TOP*, 0.931 for *screw_1*, and 0.913 for *screw_2*. With an average score of 95.2%, these results demonstrate that the trained model is highly effective in detecting both large mechanical components and smaller fasteners.

When applied to colored CAD renderings (Figure 4), the trained vision model successfully detected all four target component classes, which are BASE, TOP, screw_1 and screw_2, with high confidence scores. The detection results closely aligned with ground truth annotations, both in terms of bounding box localization and correct class assignment.

However, performance declined when the same model was applied to engineering CAD draft sheets (Figure 5). The drafts differ significantly from the training domain in three key ways: (1) the absence of color and shading results in reduced texture cues, (2) the presence of dense dimension annotations and leader lines introduces visual clutter that can be mistaken for object edges, and (3) line thickness and projection style in technical drawings differ from the photorealistic appearance of training renderings. These differences led to missed detections, particularly for small parts such as screws, and

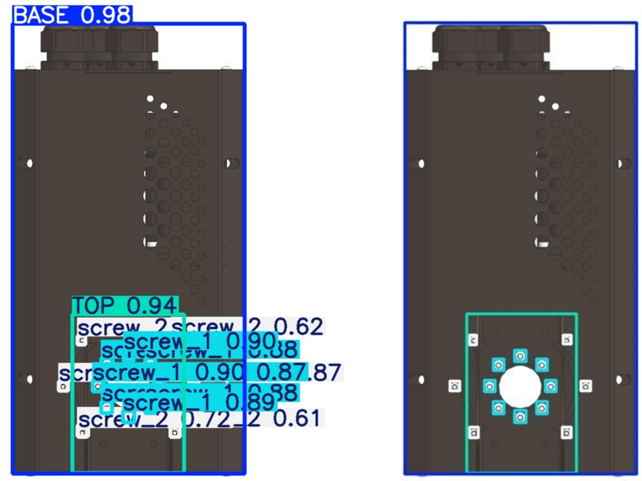


Figure 4. Detection result of rendered CAD image: detected parts (right) and annotated parts (left).

caused incorrect object counts. A possible mitigation is aligning rendered CAD images with draft sheets and transferring detected coordinates between domains. Since rendered CAD images achieved near-perfect detection accuracy, their bounding boxes could be calibrated to the draft sheet coordinate system. This approach would allow missing parts in the draft detection to be supplemented by the CAD results, improving coverage for small or hidden components while maintaining consistency with the draft representation.

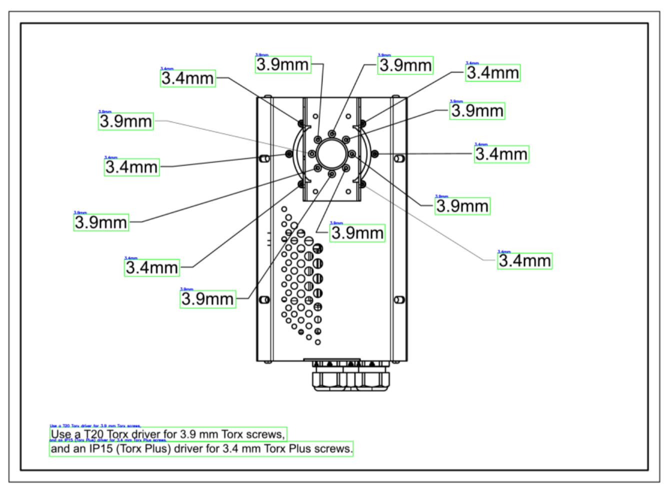


Figure 6. OCR result of a CAD draft sheet: The OCR module detected tool specifications directly from the drawing labels.

Alongside the visual detection process, the OCR module was applied to read dimensional and tooling information directly from the CAD draft sheets (Figure 6). It successfully identified screw size labels such as 3.4 mm and 3.9 mm, as well as driver types including IP15 Torx Plus and T20 Torx. This ensured that precise assembly details, which may not be visu-

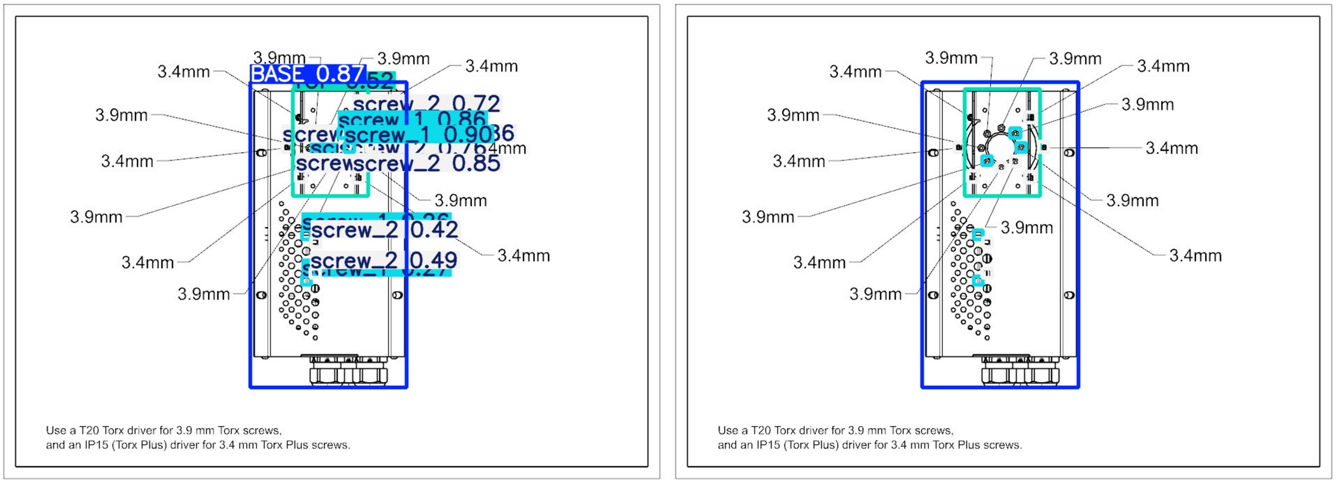


Figure 5. Detection result of CAD draft sheet: detected parts (right) and annotated parts (left)

ally distinguishable from geometry alone, were reliably captured. Even when the vision model missed components due to differences in appearance between CAD models and technical drafts, the OCR output preserved these critical specifications, providing complete and accurate inputs for the instruction generation stage.

4.2.2. Result of Assembly Instruction Generation Model

The final stage employed the open-source `gpt-oss:20b` model via Ollama to transform combined outputs from the vision and OCR modules into complete, human-readable assembly instructions. Even though the CAD draft sheet detection produced incomplete screw counts due to missed detections, as shown in the figure 7, the LLM was able to infer and restore the correct quantities by reasoning over the available textual and visual data. In particular, it leveraged dimensional specifications and driver type details extracted by OCR to ensure consistency across all fastening steps.

Figure 7 also demonstrates the difference between the concise procedural description and the generated assembly instructions. The LLM output expanded the minimal input into a detailed, logically ordered sequence that included specific tooling requirements, fastening order, and verification checks. Notably, even when the raw detection results were incomplete, the final instructions provided accurate screw counts and associated dimensions, ensuring technical correctness suitable for real-world use.

While the results demonstrate that the proposed pipeline can generate accurate and interpretable assembly instructions, the current evaluation is limited to a single representative case. The experiments were designed primarily to verify technical feasibility, specifically, the integration of vision, OCR, and language modules, rather than to measure user experience or comparative performance. A broader validation, including user studies to assess readability and task efficiency, will be conducted in future work.

5. CONCLUSION

This study presented a multi-stage framework that integrates a vision model, OCR, and a large language model to automatically generate detailed, human-readable assembly instructions from CAD draft sheets. The vision model, trained on a synthetic dataset of 40,000 images, achieved an overall mean average precision (mAP) of 95.2%, with individual average precision scores of 0.995 for BASE, 0.969 for TOP, 0.931 for screw_1, and 0.913 for screw_2. While the model performed reliably on colored CAD renderings, its accuracy decreased when applied to engineering draft sheets due to differences in color, annotation density, and visual style, indicating a domain gap that requires further fine-tuning.

The OCR module successfully extracted all dimensional specifications (e.g., 3.4 mm, 3.9 mm) and tooling requirements (e.g., T20 Torx, IP15 Torx Plus), ensuring that critical assembly parameters were preserved even when visual detection was imperfect. The LLM transformed the combined visual and textual outputs into complete, logically ordered instructions, and in some cases, inferred more accurate fastening details than those directly provided by the vision model.

Despite these promising results, the current system has limitations. The vision model's performance on draft sheets remains sensitive to domain differences, and the LLM's reasoning depends heavily on the quality of its input data. Furthermore, the evaluation focused primarily on a single case study, which may limit generalizability to other product types and drawing styles.

Future work will focus on fine-tuning the vision model for CAD draft detection, incorporating real-world assembly videos as an additional input source to capture richer con-

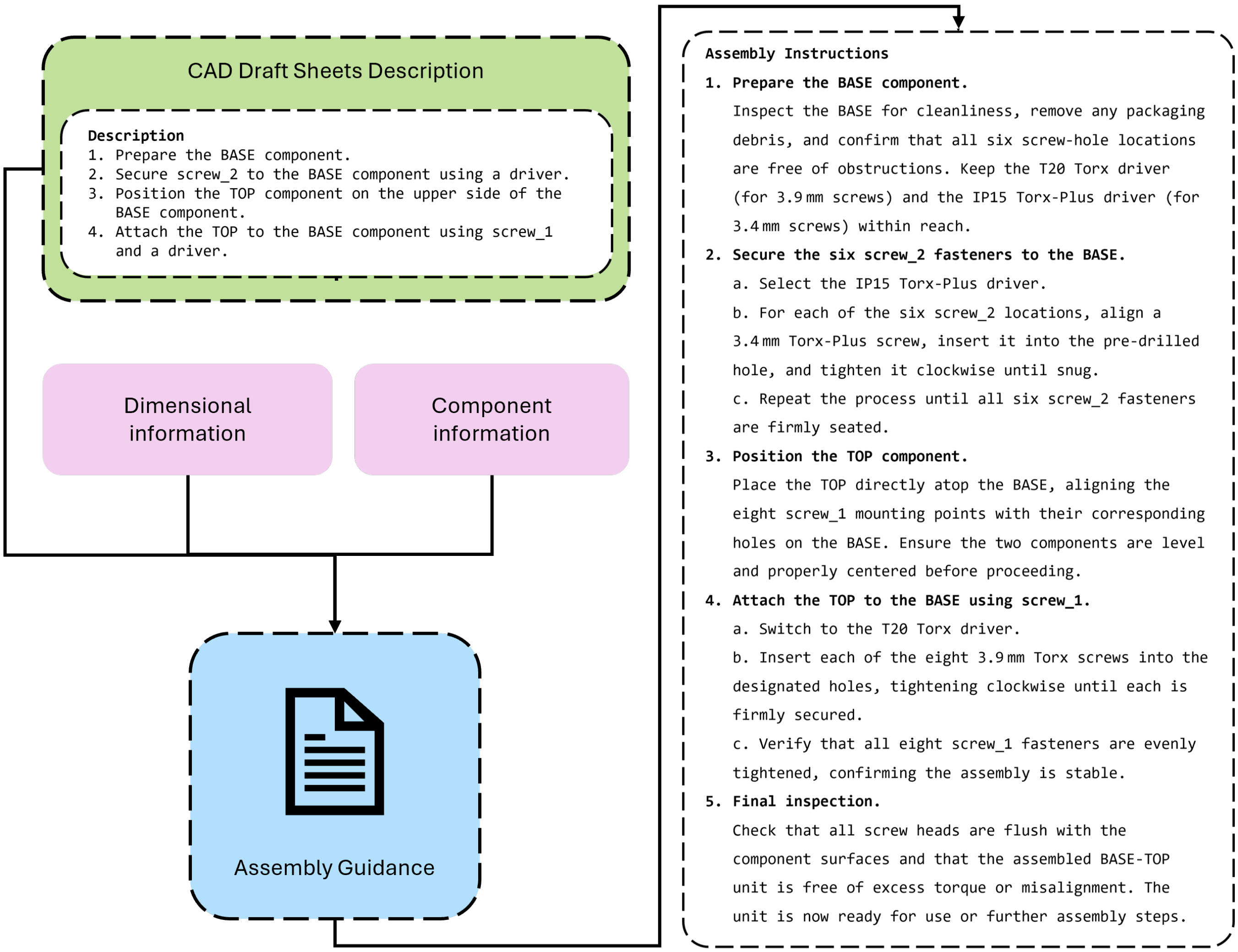


Figure 7. Comparison between the original CAD draft description and the assembly guidance generated by the `gpt-oss:20b` model. The LLM-generated instructions were derived from recognized dimensional and component information extracted through the vision module, transforming descriptive text into structured, step-by-step assembly guidance.

textual information, and enhancing the system's versatility across varied assembly scenarios. User studies will be conducted to assess how effectively the generated instructions improve comprehension and assembly speed compared to traditional CAD drafts. Another promising direction is to align CAD renderings with draft sheets so that high-confidence detections from the renderings can be transferred to the drafts, filling in missing details and reducing domain-related errors. These developments aim to strengthen cross-format robustness, improve detection accuracy, and broaden the applicability of the proposed framework to diverse manufacturing and maintenance documentation tasks.

Acknowledgment

This study is supported by Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea Government(MOTIE) (RS-2025-13672968, Development of international standards for digital product passports for vision data-based manufacturing and recycling processes for a circular economy) and supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. NRF RS-2023-00219052)

References

Bocák, R., Holubek, R., & Tirian, G. (2022). New approach in assembly, disassembly and maintenance processes i4. 0 by displaying in augmented reality environment. In *Journal of physics: Conference series* (Vol. 2212, p. 012015).

Colabianchi, S., Costantino, F., & Sabetta, N. (2024). Assessment of a large language model based digital intelligent assistant in assembly manufacturing. *Computers in Industry*, *162*, 104129.

Duan, Y., Chen, Z., Hu, Y., Wang, W., Ye, S., Shi, B., ... others (2025). Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the computer vision and pattern recognition conference* (pp. 4026–4037).

Holvoet, L., van Bekkum, M., & de Vries, A. (2024). An approach to automated instruction generation with grounding using llms and rag. In *European symposium on artificial intelligence in manufacturing* (pp. 224–233).

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). 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).

Jamieson, L., Elyan, E., & Moreno-García, C. F. (2024). Few-shot symbol detection in engineering drawings. *Applied Artificial Intelligence*, *38*(1), 2406712.

Jocher, G., & Qiu, J. (2024). *Ultralytics yolo11.* Retrieved from `https://github.com/ultralytics/ultralyt`

Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., ... Park, S. (2021). Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, *7*(15), 2.

Li, D. (2023). *Sim2real: generating synthetic images from industry cad models with domain randomization.*

Maupou, C., Yang, Y., Fodop, G., Qie, Y., Migliorini, C., Mehdi-Souzani, C., & Anwer, N. (2024). Automatic raster engineering drawing digitisation for legacy parts towards advanced manufacturing. *Procedia CIRP*, *129*, 234–239.

Monnet, J., Petrovic, O., & Herfs, W. (2024). Investigating the generation of synthetic data for surface defect detection: A comparative analysis. *Procedia CIRP*, *130*, 767–773.

Pasanisi, D., Rota, E., Ermidoro, M., & Fasanotti, L. (2023). On domain randomization for object detection in real industrial scenarios using synthetic images. *Procedia Computer Science*, *217*, 816–825.

Schraml, D., & Notni, G. (2024). Synthetic training data in ai-driven quality inspection: The significance of camera, lighting, and noise parameters. *Sensors*, *24*(2), 649.

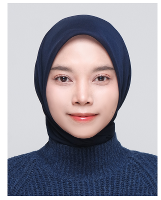
Tito, R., Mathew, M., Jawahar, C., Valveny, E., & Karatzas, D. (2021). Icdar 2021 competition on document visual question answering. In *International conference on document analysis and recognition* (pp. 635–649).

Villena Toro, J., Wiberg, A., & Tarkian, M. (2023). Optical character recognition on engineering drawings to achieve automation in production quality control. *Frontiers in manufacturing technology*, *3*, 1154132.

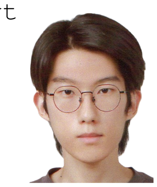
Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7464–7475).

Zhang, W., Joseph, J., Yin, Y., Xie, L., Furuhata, T., Yamakawa, S., ... Kara, L. B. (2023). Component segmentation of engineering drawings using graph convolutional networks. *Computers in Industry*, *147*, 103885.

BIOGRAPHIES



Shokhikha Amalana Murdivien earned a Bachelor's degree in Informatics Engineering Education from Yogyakarta State University, Indonesia. She then completed her Master's degree in Industrial and Management Systems Engineering at Kyung Hee University, South Korea. She is currently pursuing a Ph.D. in the Department of Artificial Intelligence at Kyung Hee University. She is also a member of the university's Smart Factory Laboratory. Her research interests include industrial standards and digital twins, human-robot collaboration, and large language model-based multi-agent systems for autonomous smart manufacturing.



Kyungwan Choi earned a Bachelor's degree in Industrial and Management Systems Engineering from the College of Engineering at Kyung Hee University in 2025. Since the same year, he has been pursuing a Master's degree at the Metaverse Convergence Graduate School of Kyung Hee University. He is a member of the university's Smart Factory Laboratory, where his research focuses on autonomous manufacturing, large language models (LLMs), assembly automation, and virtual environments.

Minji Kim earned a Bachelor's degree in Industrial and Management Systems Engineering from the College of Engineering at Kyung Hee University in 2024. Since the same year, she has been pursuing a Master's degree at the Artificial Intelligence Graduate School of Kyung Hee University. She is a member of the university's Smart Factory Laboratory, where her research focuses on autonomous manufacturing, system design for smart factory and digital product passport.

Jumyung Um received the B.S. degree in mechanical engineering and information communication engineering from Sungkyunkwan University, Suwon, Republic of Korea, in 2003, and the Ph.D. degree in industrial and management engineering from Pohang University of Science and Technology, Republic of Korea, in 2012. From 2012 to 2014, he was a Postdoctoral Researcher with the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. After leaving EPFL, he was a Research Associate with the University of Cambridge, until 2015, and a Senior Researcher of technology initiative with SmartFactoryKL, Germany, in 2018. Since 2018, he has been an Associate Professor with the Department of Industrial and Management System Engineering, Kyung Hee University, Yongin, Republic of Korea. His research interests include the usage of augmented reality for smart operators and plug-nproduce of modular factory systems.