Enginuity: Building an Open Multi-Domain Dataset of Complex Engineering Diagrams

Tirthankar Ghosal
Oak Ridge National Laboratory
ghosalt@ornl.gov

Ethan Seefried
Oak Ridge National Laboratory
seefriedej@ornl.gov

Prahitha Movva
Independent Researcher
prahitha.movva03@gmail.com

Naga Harshita Marupaka
Independent Researcher
nagaharshitamarupaka@gmail.com

Tilak Kasturi
Predii
tilak@predii.com

Robert Saethre
Oak Ridge National Laboratory
saethrerb@ornl.gov

Prasanna Balaprakash
Oak Ridge National Laboratory
pbalapra@ornl.gov

Abstract

We propose *Enginuity* - the first open, large-scale, multi-domain engineering diagram dataset with comprehensive structural annotations designed for automated diagram parsing. By capturing hierarchical component relationships, connections, and semantic elements across diverse engineering domains, our proposed dataset would enable multimodal large language models to address critical downstream tasks including structured diagram parsing, cross-modal information retrieval, and AI-assisted engineering simulation. Enginuity would be transformative for AI for Scientific Discovery by enabling artificial intelligence systems to comprehend and manipulate the visual-structural knowledge embedded in engineering diagrams, breaking down a fundamental barrier that currently prevents AI from fully participating in scientific workflows where diagram interpretation, technical drawing analysis, and visual reasoning are essential for hypothesis generation, experimental design, and discovery.

1 Dataset Rationale

Engineering diagrams encode the core knowledge of scientific and technical disciplines but remain inaccessible to AI (refer Appendix A.1) due to proprietary reasons. While current methods achieve 85%+ accuracy on symbol detection, they struggle with relationship extraction—the critical bottleneck where performance drops by 25%+ (13), preventing true diagram understanding. No public dataset exists with >10K real-world engineering diagrams containing both component and structural relationship annotations. This gap prevents AI from participating in scientific workflows requiring visual-structural reasoning and system-level comprehension.

We propose Enginuity—50K annotated engineering diagrams starting with automotive domain (will be later expanded to include other domains) through our partnership with *Predii*, an automotive AI company processing 2B+ repair jobs monthly. Automotive diagrams are ideal: they combine visual structure, text, and functional knowledge in exploded parts diagrams used by technicians globally. Our dataset will enable three core AI tasks: **(1) component detection (2) relationship extraction and (3) diagram VQA.** This will culminate in a CVPR 2026 workshop and shared task, and later into a leaderboard *arena* to test capabilities of frontier AI models.

2 Dataset Building Strategy

To balance openness with real-world relevance, for data collection, we will adopt a two-pronged strategy:

Public-domain automotive diagrams – We will collect and annotate diagrams from declassified government vehicles and older vehicles in the public domain. These resources include both exploded parts diagrams and associated technical procedure manuals that explicitly reference the diagrams in repair workflows. Our industry collaborator will bring in domain experts for the human-labeling.

Industry engagement framework: We will establish a framework for private industry contributors (e.g., OEMs) to contribute "older vehicle parts diagrams" (5-15 years old) without disclosing proprietary information via our industry collaborator. This creates a pathway for researchers to access diverse, realistic datasets while giving industry partners a mechanism to engage with the research community.

Enginuity will contain 50K+ diagrams spanning powertrains, chassis, and body components from 500+ vehicle models. Annotations include hierarchical component relationships, spatial connections, part numbers, and functional roles, standardized to ISO/IEEE ontologies. Our 4-stage annotation pipeline uses AI for initial detection, dedicated teams for refinement, expert validation on 10% samples, and active learning to reduce costs by 65%.

Please refer to Appendices A.2, A.3, A.4, A.5, and A.9 for further details.

3 Defining the AI Task

Understanding and reasoning on scientific/engineering diagrams is a *long-studied, difficult computer vision problem* (7; 9; 11), which requires progress in perception, structured representation, and multimodal reasoning. We believe no single group can solve this in isolation; it demands a sustained community effort supported by a shared benchmark dataset and leaderboard. **Primary task: parsing diagrams into structured graphs linking visual components to part identifiers.** We have a set of concrete subtasks spanning *component and symbol recognition, relationship extraction, functional context interpretation, diagram question answering, multimodal information retrieval, diagram-to-digital-twin alignment* (we elaborate more on these tasks in Appendix A.6 and evaluation metrics in A.7.)

4 Potential Acceleration

Enginuity will help to train foundation models for several hard downstream tasks including the ones mentioned in the previous section. The ability to automatically parse and reason over diagrams opens pathways to *scientific acceleration*: **Algorithmic exploration of designs, Cross-domain transfer** - models trained on automotive diagrams generalize to mechanical/process engineering, **Digital twin generation**—automated conversion from 2D diagrams to 3D simulations, and **Knowledge preservation & harmonization** across notation changes spanning decades. Practical *downstream accelerations* include: **Design optimization, Simulation integration, and Automated documentation**. Please refer to Appendix A.10 for further details.

By making such a resource openly available, we will **lower the entry barrier for researchers, catalyze cross-disciplinary innovation, and unite the AI and scientific communities** to tackle one of the most challenging problems in automated scientific discovery. The Enginuity dataset would enable frontier AI models to finally crack the spatial reasoning bottleneck in engineering diagram comprehension by providing the first large-scale (50K+), multi-domain training corpus with rich spatial annotations that bridges the critical gap between isolated symbol recognition and true system-level understanding. This breakthrough would transform AI from merely detecting components to actually comprehending how complex engineering systems interconnect and function, unlocking automated digital twin generation, cross-domain technical reasoning, and real-time engineering design validation that could accelerate the entire Industry 4.0 transformation[1].

[1] https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-industry-4-0-the-fourth-industrial-revolution-and-4ir

5 Budget and Timeline

Our overall estimated budget for data collection & annotations ($150K), infrastructure support ($30K), baselines & evaluation ($20K) would approximately amount to $200K US dollars. By the end of month 12 we will have 50K annotated images on Hugging Face for the research community. If successful, we would anticipate further support from the industry to launch the *arena*.

6 Conclusion

website

7 Acknowledgement

ORNL, Predii, ++ ??, SentinelDevices

References

[1] Alenazi, M. J., and Sterbenz, J. P. Comprehensive comparison and accuracy of graph metrics in predicting network resilience. In *2015 11th international conference on the design of reliable communication networks (DRCN)* (2015), IEEE, pp. 157–164.

[2] Bayer, J. Handwritten schematics - cghd, 2025.

[3] Beutenmüller, F., Dierolf, B., Keckeisen, M., Pausinger, F., and Vaudrevange, P. Topological data analysis in automotive industry. In *International Stuttgart Symposium* (2023), Springer, pp. 44–56.

[4] Bigg, C. Diagrams. *A companion to the history of science* (2016), 557–571.

[5] Boon, M. Diagrammatic models in the engineering sciences. *Foundations of Science 13*, 2 (2008), 127–142.

[6] Elyan, E., Moreno, C. G., and Johnston, P. Symbols in engineering drawings (sied): An imbalanced dataset benchmarked by convolutional neural networks. In *Proceedings of the 2020 International Joint Conference of the 21st EANN (Engineering Applications of Neural Networks), EANN 2020* (2020), vol. 2 of *Proceedings of the International Neural Networks Society*, Springer, Cham.

[7] Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *European conference on computer vision* (2016), Springer, pp. 235–251.

[8] Khan, M. T., Chen, L., Ng, Y. H., Feng, W., Tan, N. Y. J., and Moon, S. K. Fine-tuning vision-language model for automated engineering drawing information extraction. *arXiv preprint arXiv:2411.03707* (2024).

[9] Mani, S., Haddad, M. A., Constantini, D., Douhard, W., Li, Q., and Poirier, L. Automatic digitization of engineering diagrams using deep learning and graph search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020), pp. 176–177.

[10] Moktefi, A. Diagrams as scientific instruments. *Virtual Reality–Real Visuality* (2017), 81–89.

[11] Montalvo, F. S. Diagram understanding: the intersection of computer vision and graphics.

[12] Perini, L. Diagrams in biology. *The Knowledge Engineering Review 28*, 3 (2013), 273–286.

[13] Stürmer, J. M., Graumann, M., and Koch, T. Transforming engineering diagrams: A novel approach for p&id digitization using transformers. *arXiv preprint arXiv:2411.13929* (2024).

[14] Yang, L., Zhang, J., Li, H., Ren, L., Yang, C., Wang, J., and Shi, D. A comprehensive end-to-end computer vision framework for restoration and recognition of low-quality engineering drawings, 2023.

[15] Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. Object detection in 20 years: A survey. *Proceedings of the IEEE 111*, 3 (2023), 257–276.

A Supplementary Material

A.1 Introduction

Engineering diagrams serve as the universal visual language of scientific and technical disciplines, encoding critical knowledge about system architectures, process flows, circuit designs, molecular structures, and experimental setups that form the backbone of scientific research and engineering practice. These diagrams represent decades of accumulated scientific knowledge in a structured, standardized format that researchers rely on for design, analysis, communication, and innovation in different fields of science and technology. In essence, these diagrams are blueprints and instruments of science (4; 10; 12), ensuring that the components of an experiment fit and function together, reducing costly design errors and accelerating assembly. However, due to the inherent complexity of these diagrams, even the frontier AI models struggle to comprehend them [2]. To train modern AI applications, these complex diagrams needs human labeling. However, most such diagrams remain locked in proprietary silos, out of reach for researchers. The lack of a large, open, expert-labeled, cross-disciplinary dataset of complex engineering diagrams is a major bottleneck (3), limiting the ability of AI to generalize, generate, or link diagrams to digital twins.

Hence, we propose *Enginuity*–a large-scale, open, multi-domain dataset of complex engineering diagrams to bridge the gap and help train our models for this hard task. Additionally, we want to catalyze breakthroughs in AI-assisted knowledge extraction from complex diagrams, multimodal information retrieval, scientific reasoning, complex diagram parsing, and design generation by building a community around the dataset and uniting researchers across disciplines. This initiative will culminate in a shared task challenge and workshop that we are already planning for CVPR 2026. If successful, we would like to build a *arena* in lines of *lmsys*[3] where the frontier AI models will be tested on this hard benchmark. With the assistance of our automotive industry collaborator (also coauthor of this proposal) who is bringing in the wealth of domain knowledge of automotive diagrams and data, we are confident to launch this initiative; although automotive diagrams will form a significant portion of Enginuity 1.0, but we are not only limited to it. **Why automotive diagrams matter?** Automotive vehicles provide a uniquely compelling case study. Modern repair workflows are highly dependent on exploded parts diagrams, which allow technicians to identify replacement parts by linking a visual diagram to a set of associated components, assemblies, one-time-use parts, and manufacturer-specific part numbers. In practice, technicians often use natural language queries (e.g., "front-left brake caliper") to navigate these diagrams, bridging between visual structure and symbolic identifiers. The tight coupling of visual, textual, and functional knowledge mirrors challenges across many scientific domains, making automotive diagrams an ideal testbed for advancing AI methods in multi-modal scientific reasoning (5).

A.2 Comparison to Existing Datasets

- **Small or narrow in scope** – Current datasets focus on limited domains such as P&ID, electrical schematics, or specific CAD formats.
- **Proprietary or restricted** – Many high-value diagrams remain locked within corporate, government, or private archives, inaccessible due to IP, competition, or regulation.
- **Domain-locked** – Public datasets rarely span the full breadth of scientific diagrams, leaving cross-disciplinary AI models underdeveloped. (8)
- **Fragmented & siloed** – Data is often trapped in discipline-specific formats and communities, with no interoperability or shared standards.

[2] https://www.businesswaretech.com/blog/benchmarking-ai-on-tables-and-engineering-drawings-results-findin

[3] https://lmsys.org/

- **Legacy diagram barriers** – Decades of archival diagrams use shifting notations and conventions, hindering knowledge transfer between veteran experts and newer technologists. Future models trained on this dataset could harmonize visual languages and bridge generational gaps.

Current datasets are too narrow to support the proposed accelerations:

- **Handwritten engineering diagrams (circuits)** (2) – Focused exclusively on small-scale electrical circuits, lacking diversity in both domain and diagram type.
- **SiED: Symbol Classification Engineering Diagrams** (6) – Provides isolated symbol classification but does not address higher-order structure, system-level relationships, or domain generalization.

By contrast, our dataset is **multi-domain, multi-scale, and system-focused**, capturing the full spectrum from individual components to assemblies and process flows. This breadth directly enables generalizable AI models capable of supporting scientific discovery, industrial design, and digital twin development in a way that current resources cannot.

A.3 Data Creation Pathway

The dataset will be sourced through collaborations with automotive industry partners, providing technical schematics and parts lists across diverse car models and years. These materials will span thousands of diagrams covering powertrains, chassis, electronics, and body components.

Annotation will combine semi-automated preprocessing with expert review, leveraging (*Anonymous*) AI platform to align technical manuals with exploded parts diagrams. Human-in-the-loop annotation by technicians will ensure accuracy and domain fidelity.

Key annotation targets include:

- Part–number relationships linking visual components to identifiers.
- Hierarchy tagging of components, sub-components, assemblies, and one-time-use parts.
- Specifications and usage types for functional context.
- System–subsystem clustering of diagrams for structured retrieval.
- Diagram question answering to support multi-modal reasoning tasks.

This approach ensures the dataset is both scalable and high-quality, directly reflecting real-world automotive repair workflows.

Proposed dataset features:

- **Scale** - More than 50K diagrams from automotive systems (*there are over 50,000 distinct vehicle year–make–model–engine combinations in the North American passenger vehicle market alone*), and process engineering.
- **Focus** - Emphasis on physical structure and relationships; excluding electrical schematics.
- **Intra-domain diversity** - Coverage from prototypes to production designs, and from small parts (e.g., gears) to full systems (e.g., drivetrains).
- **Rich metadata & labels** - Domain, Component types, Connections, and Functional roles.
- **Multi-source availability** - Drawn from public archives, industrial standards, and selectively shared industry diagrams.

A.4 Annotation Strategy and Pipeline

A sophisticated, multi-stage human-in-the-loop (HITL) pipeline will ensure both scalability and accuracy while keeping costs feasible. Our approach integrates *Anonymous's* AI platform with active learning to progressively improve annotation quality and efficiency. Funding will directly support the fine-grained annotation and preparation of training, validation, and test sets for the challenge and, more broadly, the long-term academic–industry infrastructure needed to move the field forward.

Stage 1: AI-driven preprocessing. *Anonymous's* domain-specialized LLMs and vector embeddings will act as a machine labeler, performing an initial pass to detect lines, arrows, text regions, and component clusters. Heterogeneous input formats (PDF, DXF, SVG, raster scans) will be standardized into a normalized digital format, with consistent units and vectorized representations for downstream processing.

Stage 2: Dedicated refinement. Low-complexity tasks such as bounding box adjustments, OCR text verification, and alignment of auto-detected components will be handled by a dedicated annotation team in collaboration with our industry partner. This approach ensures consistent quality control, leverages domain-trained annotators, and frees expert scientists to focus on higher-value annotations.

Stage 3: Expert annotation. A domain expert team (e.g., technicians, engineers) will create a "golden set" and review 5–10% of annotations for complex components, assemblies, and functional roles. Their input will seed the validation process and establish quality benchmarks.

Stage 4: Active learning loop. Validated annotations will bootstrap model training. Updated models will then auto-label subsequent batches, focusing expert/crowd review on uncertain or novel cases. This iterative process continuously reduces annotation cost while improving model accuracy.

Annotation schema. Labels extend beyond component metadata to include:

- Object segmentation and bounding boxes
- Attributes (e.g., component type, specifications, usage)
- Relationship and topology graphs (connections, spatial orientation)
- Functional and hierarchical structures (system → subsystem → component)
- Temporal metadata for standards evolution and difficulty scoring

Standardization and Ontology Alignment. All labels will be mapped to or extend existing engineering ontologies (e.g., IEEE, ISO, ISA) to ensure interoperability and reusability across domains.

This tiered pipeline balances automation, crowdsourcing, and expert input, providing a technically feasible and cost-efficient path to generate rich, multi-level annotations at scale.

A.5 Data Partitioning

A key element of the pipeline is the dataset partitioning strategy. The annotated corpus will be divided into four subsets: training, validation, testing, and a withheld testing set reserved for the organized competition. This four-way split is intentional: the training and validation sets support model development and hyperparameter tuning, while the first testing set allows for transparent baseline reporting. The withheld test set, by contrast, is unseen during model development and will be used to score submitted competition entries. This design prevents data leakage and overfitting, ensuring a fair and reproducible evaluation protocol.

To assess generalizability, the withheld test set will additionally include diagrams sourced from private organizations outside the automotive domain. These diagrams, which may reflect distinct conventions in engineering drawing styles, annotation symbols, or system complexity, introduce a domain shift. Incorporating such heterogeneous material ensures that models are not merely optimized for a narrow distribution of automotive schematics but can extend to scientific and engineering diagrams more broadly. This robustness check is essential for evaluating real-world applicability, where trained systems must handle variability in style, notation, and structure.

A.6 Task Suite Details

- **Component and Symbol Recognition** – Detecting and classifying parts, symbols, and visual primitives across heterogeneous diagram styles.
- **Relationship Extraction** – Inferring spatial and logical connections between components to construct a machine-readable graph representation.
- **Functional Context Interpretation** – Reasoning about the role of components and subsystems (e.g., identifying assemblies, one-time-use parts, or failure-prone connections) in order to understand the diagram's operational purpose.

- **Diagram Question Answering (DQA)** – Enabling natural language queries over diagrams (e.g., “Which components must be removed before replacing the brake caliper?”), requiring joint reasoning across visual, symbolic, and textual modalities.
- **Diagram-to-Digital-Twin Alignment** – Mapping diagrams into structured formats compatible with digital twin models for simulation, retrieval, and knowledge transfer.

By explicitly defining these tasks, the dataset supports a hierarchical research agenda: from low-level perception (symbol detection) to mid-level structure induction (relationship extraction) to high-level reasoning (functional interpretation and question answering). In doing so, it creates a testbed not only for advancing multimodal AI but also for probing the limits of scientific reasoning from visual artifacts.

A.7 Evaluation Metrics

To ensure rigor and comparability, we will define clear baseline metrics for each core task. For symbol and component recognition, we propose to use **Mean Average Precision (mAP)** at standard IoU thresholds, following established conventions in object detection (15). For relationship extraction and diagram graph construction, we propose a **graph accuracy metric**, computed as the proportion of correctly predicted edges and node labels compared to ground-truth graphs (1).

These metrics are intended as initial baselines; we will refine and extend them in consultation with the research community, particularly for higher-level tasks such as diagram question answering (DQA) and diagram-to-digital-twin alignment, where specialized evaluation protocols may be needed and already some established baseline metrics are available to start with. This approach balances concreteness with flexibility, ensuring fair benchmarking while leaving room for iteration as the shared task evolves.

A.8 File Formatting

Finally, a deliberate effort will be made to address the challenge of file format inconsistency. In industrial contexts, schematics are frequently distributed as PDFs, often exported from proprietary CAD tools or drawing software. These PDFs vary in resolution, compression, layering, and embedded metadata, making them difficult to parse systematically (14). Some may contain vector graphics, while others are rasterized scans of paper documents, introducing noise and heterogeneity. Left unaddressed, such inconsistency can bias downstream models and hinder reproducibility. To mitigate this, all diagrams will be converted into a standardized, machine-readable digital format (e.g., high-resolution vector or normalized raster images). By enforcing a uniform representation, we establish consistency across the dataset, enabling reliable annotation, training, and benchmarking. Moreover, this standardization ensures that the resource can be easily extended with new contributions in the future, without inheriting the ad hoc variability of industrial documentation practices.

A.9 Dataset Availability

The curated dataset will be released under an open license and made fully publicly accessible through **Kaggle**, which provides a robust infrastructure for large-scale dataset hosting, versioning, and community engagement. Each release will be accompanied by a detailed **Kaggle Datacard** documenting data provenance, annotation schema, licensing terms, and known limitations, in line with best practices for responsible dataset publication.

To foster reproducibility and accelerate adoption, we will also provide:

- **Baseline models and benchmarks** implemented in common frameworks (PyTorch) with evaluation scripts.
- **Leaderboards** hosted on Kaggle for tracking progress during the organized CVPR competition and beyond.
- **Tutorial notebooks and usage examples** illustrating key tasks such as component recognition, diagram retrieval, and question answering.
- **Community support channels**, including a discussion forum and issue tracker to encourage feedback, error reporting, and collaborative extensions.

Long-term, the dataset will be maintained with clear **versioning protocols** to ensure stability of benchmark splits while allowing for incremental expansion to new domains. This strategy ensures not only a fair competition environment but also a sustainable resource for the broader AI-for-science community.

A.10 Acceleration Potential

The proposed dataset will significantly accelerate research at the intersection of computer vision, scientific reasoning, and engineering design. By providing a large-scale, multi-domain, and richly annotated resource, it enables a new class of scientific questions centered on the understanding, interpretation, and generation of engineering diagrams across multiple domains (*automotive domain* is main focus for **Enginuity v1.0**, but will be expanded to other domains later). Unlike text or tabular data, diagrams encode structural and relational knowledge that directly underpins physical systems. Unlocking this information algorithmically accelerates both basic science and applied innovation.

Scientific Acceleration

- **Algorithmic exploration of designs** – Models can propose, evaluate, and iterate design variations directly from diagrams, reducing the need for costly manual drafting cycles.
- **Cross-domain transfer** – Trained systems can generalize across automotive, mechanical, and process-engineering domains, enabling comparative studies and accelerating design in areas where labeled data is scarce.
- **Knowledge preservation & harmonization** – AI models trained on diagrams spanning decades can bridge generational gaps in notation, style, and conventions, preserving institutional knowledge while making it accessible to new engineers and scientists.

Applied Acceleration

Practical downstream accelerations include:

- **Design optimization** – Tools that automatically suggest structural improvements or highlight inefficiencies in complex assemblies.
- **Simulation integration** – Data-driven conversion of diagrams into machine-readable representations suitable for digital twin simulations, improving prediction accuracy and reducing experimental overhead.
- **Automated documentation** – Generation of consistent metadata and structured component lists from legacy diagrams, enabling interoperability across industry and research archives.