CADAGENT: A Multi-Agent System for Manufacturing Process Classification from 2D Engineering Drawings

Jaerim Choi
MOAI Inc. / GSEP, Seoul National University
jrim@moai.co / jaerim@snu.ac.kr

Abstract

Korean foundational ("Ppuri") manufacturers, 90% of them under fifty employees, begin every order with a senior engineer reading a 2D drawing and deciding which manufacturing process should produce the part. That decision takes one to three engineer-days and rests on tacit knowledge senior workers cannot easily hand down to juniors. CADAGENT is a multi-agent system for the drawing-to-process step. Material, Treatment, and Geometry specialists each propose a candidate process from a disjoint slice of evidence; a Conflict-Resolver orchestrator handles their disagreements with explicit logging; an Audit agent escalates low-confidence or conflicted decisions to the engineer. On 100 real engineering drawings from a deployed industrial conveyor project, the decomposition reproduces engineer-verified rule decisions on every drawing, inheriting the rule classifier's 96.0% external pilot accuracy (Wilson 95% CI 90.2–98.4%). The ablation gains are monotonic ($92.0\% \rightarrow 98.0\% \rightarrow 100\%$). The orchestrator surfaces 7 inter-agent conflicts; audit escalates them plus 10 low-confidence cases, so the engineer reviews 17 of 100 drawings rather than all 100. We analyze the SS400+HARD-Cr boundary case (n=4) that drives that review load and discuss what transfers from a single-customer pilot to broader deployment.

1 Introduction

A Korean foundational ("Ppuri") manufacturer of fifty employees, with one or two senior engineers, begins every order by reading a 2D drawing and deciding which manufacturing process should produce the part. The decision takes one to three engineer-days. It rests on tacit knowledge, and it worsens as senior engineers retire. AI assistance is the obvious move. The hard part is that the one engineer who can verify the agent is also the one it is meant to free up.

Single-model approaches fail the same way. A trained process classifier [7] or cost predictor [8] returns a softmax that hides whether material and surface-treatment evidence pointed in the same direction. A zero-shot VLM [6] reads drawings fluently and gets wrong the same boundary cases a junior engineer gets wrong, handing back one label and one number with no caveat. For an SME engineer, an unflagged wrong answer is harder to recover from than a flagged one. CADAGENT is designed to make disagreement visible rather than hide it inside a single confidence number.

Contributions. (i) A multi-agent decomposition (M, T, G specialists + Conflict-Resolver + Audit) for the drawing-to-process step. (ii) A 100-drawing pilot: rule baseline reaches 96.0% (Wilson 95% CI 90.2–98.4%); a zero-shot Claude VLM on the same set reaches $\sim$92% on PNG-only input and stumbles on special materials (ACETAL, TEFLON). (iii) An SS400+HARD-Cr boundary analysis (n=4) that a single model averages over silently, plus an honest account of what transfers to the planned $\sim$80 K evaluation.

AI Agents for Discovery in the Wild (AID-Wild), Workshop at ACM CAIS 2026.

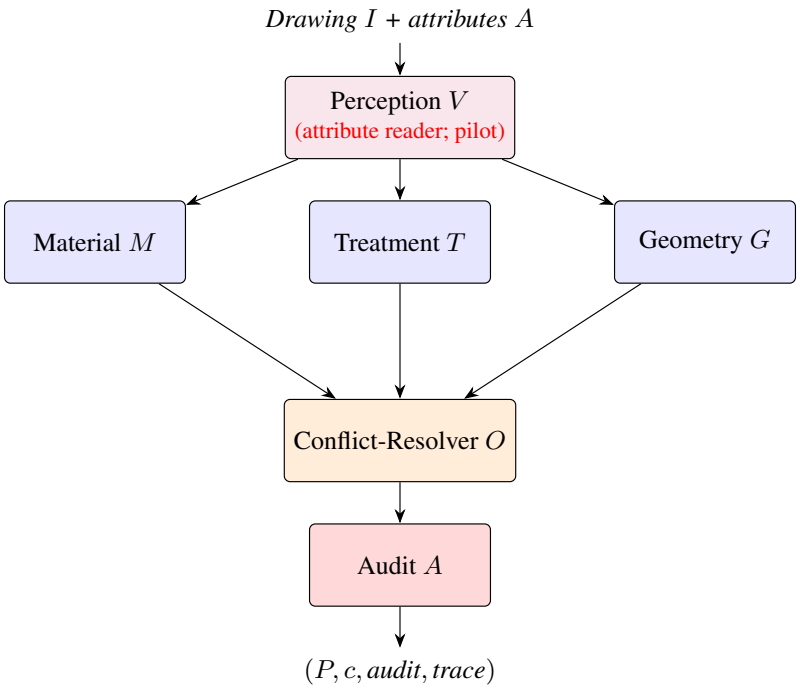

Figure 1: The CADAGENT multi-agent architecture. Specialist agents M, T, G propose process candidates from disjoint slices of evidence; the orchestrator O resolves their conflicts explicitly; the audit agent A gates the final decision behind a confidence threshold and a non-empty conflict log. The figure shows the pilot configuration; a planned RAG retrieval agent (Section 6) is future work and is intentionally not drawn.

2 Related Work

The closest agent systems, CAD-Assistant [3] (VLM + FreeCAD) and DesignAgent [4] (LLM agents for product design), target different stages, not drawing-to-process. Single-model classifiers for process selection [7] and cost prediction [8] train on curated data; manufacturing VLMs [5] explore RAG-based curation. Compound AI [2, 1] and the industry survey of agentic manufacturing [9] place CADAGENT in the human-overseen "bounded autonomy" regime. We differ in (a) decomposing the task across specialist agents with explicit conflict and audit layers, and (b) evaluating on single-customer, single-rater drawings, which is how an SME deployment typically begins.

3 The CADAGENT Multi-Agent System

We organize CADAGENT as five cooperating agents that mirror how an engineer interprets a drawing: read the page, consult several pieces of domain knowledge, reconcile conflicts, escalate when uncertain (Figure 1).

Perception agent V. Reads the drawing I and the title-block attributes A (PARTNAME, MATERIAL, AFTER_TREATMENT). In the deployed design V is a VLM; in the pilot it is a deterministic attribute reader.

Specialist agents M, T, G. Each consults a disjoint slice of evidence and returns $(P_\bullet, c_\bullet)$ or abstains. M maps material grades to default processes (AL6061 to milling, SS400 to fabrication, SUJ2 to turning). T encodes a sharper rule: hard-chrome plating ("HARD-Cr") is a precision finish applied only to machined workpieces, so T proposes milling on HARD-Cr regardless of base material. G inspects part-name patterns: SHAFT, WASHER, BUSHING imply turning; POST, FRAME, BKT imply fabrication.

Table 1: Process distribution across the 100 pilot drawings.

Process	Count	%
Milling (MCT)	40	40.0
Sheet metal	26	26.0
Fabrication	11	11.0
Profile assembly	11	11.0
Special machining	8	8.0
Turning (CNC)	4	4.0

Table 2: Per-configuration ablation on 100 drawings. Fidelity is the fraction matching engineer-verified rule decisions. "Confl." is disagreements detected by the orchestrator O between specialists. "Audit" fires if final $c<0.9$ or the conflict log is non-empty.

Configuration	Fidelity	Confl.	Audit
M	92.0%	—	—
$M+T$	98.0%	—	—
$M+T+G$	100.0%	—	—
$M+T+G+O$	100.0%	7	—
Full $(+A)$	**100.0%**	**7**	**17** (**17.0%**)

Conflict-Resolver orchestrator O. On agreement, O emits the common process at $\max c_\bullet$. On disagreement, O applies a specificity rule: Treatment overrides Material only when $c_M<0.9$; Geometry overrides Material only when $c_M<0.9$ and $c_G\geq 0.9$. A high-confidence Material call cannot be flipped by a weaker specialist (keeping SPCC+PAINTING from being reclassified, Section 5). Every disagreement is written to a conflict log; O does not silently pick a winner.

Audit agent A. Escalates when final $c<\tau=0.9$ *or* the conflict log is non-empty. High-confidence consensus passes through; any disagreement is surfaced to the engineer together with the per-agent reasoning trace.

Implementation. The agent decomposition is implemented as deterministic Python functions over the title-block fields. V, M, T, G, O, A are functions, not LLM calls. No agentic framework is used, and the confidences are constants chosen once with the domain engineer, not learned. The system is deterministic and single-pass. The companion zero-shot VLM baseline (Section 4) runs Anthropic Claude over the Pipeline B PNG renderings in a separate configuration; it is not part of the agent loop and is not drawn in Figure 1. The planned full system replaces V with the VLM and adds a retrieval agent over historical parts.

4 Pilot Study

Data and pipelines. A single DWG file (industrial conveyor system), processed through two pipelines. **Pipeline A** (DWG $\rightarrow$ DDC $\rightarrow$ XLSX) extracts title-block attributes (PART-NAME, MATERIAL, AFTER_TREATMENT) for rule-based classification; **Pipeline B** (DWG $\rightarrow$ ODA $\rightarrow$ DXF $\rightarrow$ PNG via ezdxf) renders per-drawing images for the VLM baseline. Deduplication and attribute-completeness filtering leave 100 unique drawings (Table 1).

Ground truth. The engineer who supplied the DWG verified all 100 labels; multi-rater κ on a multi-customer corpus is part of the planned follow-up.

Ablation. Five configurations, incrementally adding agents (Table 2). *Decomposition fidelity* is the fraction of drawings on which a configuration reproduces the engineer-verified rule decision. External accuracy is that fidelity times the rule classifier's 96.0% pilot accuracy ($96/100$, Wilson 95% CI 90.2–98.4%).

Result summary. Each added agent helps. Material alone reproduces 92.0%. Adding Treatment lifts this to 98.0% by flipping the four SS400+HARD-Cr cases (Section 5), and adding Geometry

closes the remaining two. The orchestrator surfaces seven conflicts: four SS400+HARD-Cr (Material says fabrication, Treatment wins with milling), one SPCC+PAINTING where Material holds (c_M=0.95 $\gg$ c_T=0.7), and two geometry-vs-material disagreements. Audit escalates all seven plus ten low-confidence cases, so the engineer reviews 17 of 100 drawings rather than all 100. Every disagreement is in that 17.

Zero-shot VLM baseline. A zero-shot Claude VLM on the same 100 drawings via Pipeline B reaches ∼92% on PNG-only input, comparable to Material alone, and fails on special-material parts (ACETAL, TEFLON) whose geometry overlaps with milled steel. Title-block attributes in the prompt lift accuracy to near 100%, but the VLM still emits one label and one number with no signal that material and shape disagreed. Our system records the same disagreement as an explicit conflict and escalates it to the engineer via the audit gate.

5 Boundary-Case Failure Analysis: SS400 with Hard-Chrome Plating

Four of the 100 drawings show the SS400+HARD-Cr pattern. On a representative `REDUCER PLATE` (SS400, HARD-Cr):

M**:** *fabrication*, c=0.8 (SS400 is a structural steel welded into frames).

T**:** *milling*, c=0.85 (HARD-Cr is applied only to machined workpieces).

G**:** abstains; `PLATE` alone is ambiguous.

O**:** records an $M \leftrightarrow T$ conflict; c_M<0.9, specificity rule fires, Treatment wins. Final: *milling* at c=0.8.

A**:** flags the non-empty conflict log; engineer sees both proposals with per-agent rationales.

The same pattern recurs on three further parts (an `X-AXIS ROBOT BASE` pair and a `SCALE BASE`). A monolithic classifier emitting *milling* at the same final confidence would be *numerically* indistinguishable, but it would not show the engineer the disagreement. That is the practical difference. The asymmetric rule cuts the other way on `ANCHOR BKT` (SPCC+PAINTING): c_M=0.95 exceeds the cutoff, so the orchestrator records a conflict but does *not* flip Material, and audit still escalates so the engineer can confirm SPCC dominates the generic painting signal.

6 Discussion: What Transfers from Pilot to Production

Other baselines. No trained classifier baseline on these 100 drawings: with N=100 from one customer, a trained model would overfit (rules already separate the classes) or leak across train and test (drawings in one project share material and naming conventions); full multi-customer comparison is deferred. No LLM-as-judge: ground truth is the engineer's rule-verified labels.

Hallucination and calibration. The deterministic V does not hallucinate; the planned VLM V will. The audit gate handles the first-order failure: a fabricated material call still has to agree with the treatment-rule and the geometry-rule, or it is escalated. The harder failure is silent miscalibration of per-agent confidences, which the multi-customer calibration protocol addresses. Reward hacking is N/A: no learning loop in the pilot.

Limitations and roadmap. 100 drawings from one DWG file reflect a single firm's drafting conventions, and inference latency and per-query cost are not measured. From ∼80,000 transaction records on a Korean B2B platform, quality filtering yields 15,000–25,000 usable drawings, supporting (a) VLM-prompted M, T, G plus a RAG agent over historical parts, (b) four-way comparison against deep learning, zero-shot VLM, and rule-only baselines, and (c) a multi-rater protocol with inter-rater κ.

On a real manufacturing floor, what matters is not a single model with a high accuracy number, but an AI system, divided across several specialist models, that can actually be deployed. That distinction has practical consequences for SME manufacturers, and it is what CADAGENT is designed to demonstrate.

References

[1] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, and A. V. Vasilakos. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv:2501.09136*, 2025.

[2] Y.-A. Lee, G.-T. Yi, M.-Y. Liu, J.-C. Lu, G.-B. Yang, and Y.-N. Chen. Compound AI systems optimization: A survey of methods, challenges, and future directions. In *Proc. EMNLP*, pages 28748–28763, 2025.

[3] D. Mallis et al. CAD-Assistant: Tool-augmented VLLMs as generic CAD task solvers. In *Proc. ICCV*, 2025.

[4] P. Chen, Y. Cai, Z. Zhou, J. Yao, J. Li, W. You, and L. Sun. DesignAgent: An LLM-based multi-agent system to assist early-stage product design and evaluation. *Journal of Engineering Design*, 37(3):945–980, 2026.

[5] H. Fan, C. Liu, N. E. Janvisloo, S. Bian, J. Y. H. Fuh, W. F. Lu, and B. Li. MaViLa: Unlocking new potentials in smart manufacturing through vision language models. *Journal of Manufacturing Systems*, 2025.

[6] C. Picard et al. From concept to manufacturing: Evaluating vision-language models for engineering design. *Artificial Intelligence Review*, 58:288, 2025.

[7] L. Xie, Y. Lu, T. Furuhata, S. Yamakawa, W. Zhang, A. Regmi, L. Kara, and K. Shimada. Graph neural network-enabled manufacturing method classification from engineering drawings. *Computers in Industry*, 142:103697, 2022.

[8] A. B. Arıkan et al. Machine learning-based manufacturing cost prediction from 2D engineering drawings via geometric features. *arXiv:2508.12440*, 2025.

[9] K. M. Edwards, M. Bauer, C. Jacquillat, A. J. Hart, and F. Ahmed. Agentic AI in engineering and manufacturing: Industry perspectives on utility, adoption, challenges, and opportunities. *arXiv:2604.09633*, 2026.