FORGE: Fine-grained multimodal evaluation for manufacturing quality assurance

Xiangru Jian[1,*], **Hao Xu**[2,3,*,†], **Wei Pang**[4,*], **Xinjian Zhao**[4], **Chengyu Tao**[5], **Qixin Zhang**[6]
Xikun Zhang[7,†], **Chao Zhang**[1], **Guanzhi Deng**[8], **Alex Xue**[1], **Juan Du**[9],
Tianshu Yu[4], **Garth Tarr**[2], **Linqi Song**[8,10], **Qiuzhuang Sun**[3], **Dacheng Tao**[6]
[1]University of Waterloo, Canada [2]University of Sydney, Australia
[3]Singapore Management University, Singapore
[4]The Chinese University of Hong Kong, Shenzhen, China
[5]Hunan University, China [6]Nanyang Technological University, Singapore
[7]Royal Melbourne Institute of Technology, Australia [8]City University of Hong Kong, China
[9]The Hong Kong University of Science and Technology (Guangzhou), China
[10]City University of Hong Kong Shenzhen Research Institute, China
[*]Equal contribution [†]Corresponding author

Abstract

The manufacturing sector is increasingly adopting Multimodal Large Language Models (MLLMs) to move from low-level perception toward higher-level reasoning and decision support, yet current evaluations rarely reflect the fine-grained precision that real-world manufacturing demands. Progress is further hindered by data scarcity and a lack of fine-grained domain semantics in existing datasets. To bridge this gap, we introduce FORGE,[1] a benchmark for three core quality-assurance tasks of discrete manufacturing: workpiece verification, structural surface inspection, and assembly verification. We first construct a high-quality multimodal dataset spanning real-world 2D images and 3D point clouds, annotated with fine-grained domain semantics such as standardized model numbers. We then evaluate 18 state-of-the-art MLLMs on these three tasks, revealing significant performance gaps. Counter to conventional understanding, our bottleneck analysis on five representative models indicates that visual grounding is not the primary limiting factor. Instead, insufficient domain-specific knowledge is the key bottleneck. Beyond evaluation, we show that our structured annotations can serve as an actionable training resource: supervised fine-tuning of a compact 3B-parameter model on our data yields up to 90.8% relative improvement in accuracy on held-out manufacturing scenarios, providing preliminary evidence for a practical pathway toward domain-adapted manufacturing MLLMs.

[1]Dataset, evaluation harness, and fine-tuning scripts are released under the MIT License at https://github.com/AI4Manufacturing/FORGE and https://huggingface.co/datasets/AI4Manufacturing/forge.

1 Introduction

The manufacturing sector, a critical pillar of the global economy, generates massive amounts of heterogeneous data from production lines and relies on complex decision-making systems (Gautam et al., 2025). The complexity and volume of such data strongly require advanced technologies to address the challenges of integrating and interpreting fragmented multimodal data. Crucially, as modern manufacturing paradigms increasingly rely on data-driven decision-making and sophisticated human-machine collaboration, there is an urgent need for intelligent systems capable of higher-level cognitive tasks (Gautam et al., 2025; Fan et al., 2024; Wang et al., 2024; Yuan et al., 2025).

Vision models have been widely applied, primarily functioning as perception modules that focus on information extraction, such as object localization and anomaly detection (Defard et al., 2021; Roth et al., 2022; Tao et al., 2023; Wen et al., 2024). These models typically operate within a modular, pipelined architecture, generating task-specific outputs (*e.g.*, defect type and location) that are subsequently passed to higher-level Manufacturing Execution Systems (MES) (Saenz de Ugarte et al., 2009; Yuan et al., 2025) for decision-making. Yet, these vision models are fundamentally limited by their inability to reason and execute autonomous control (Zhang et al., 2024; Fan et al., 2025; Lin et al., 2025).

More recently, Large Language Models (LLMs) and MLLMs have demonstrated remarkable generalization and cross-task transfer capabilities across diverse domains, including multimodal media understanding (Jian et al., 2026b), topology analy-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 18762–18839
October 24-29, 2026 ©2026 Association for Computational Linguistics

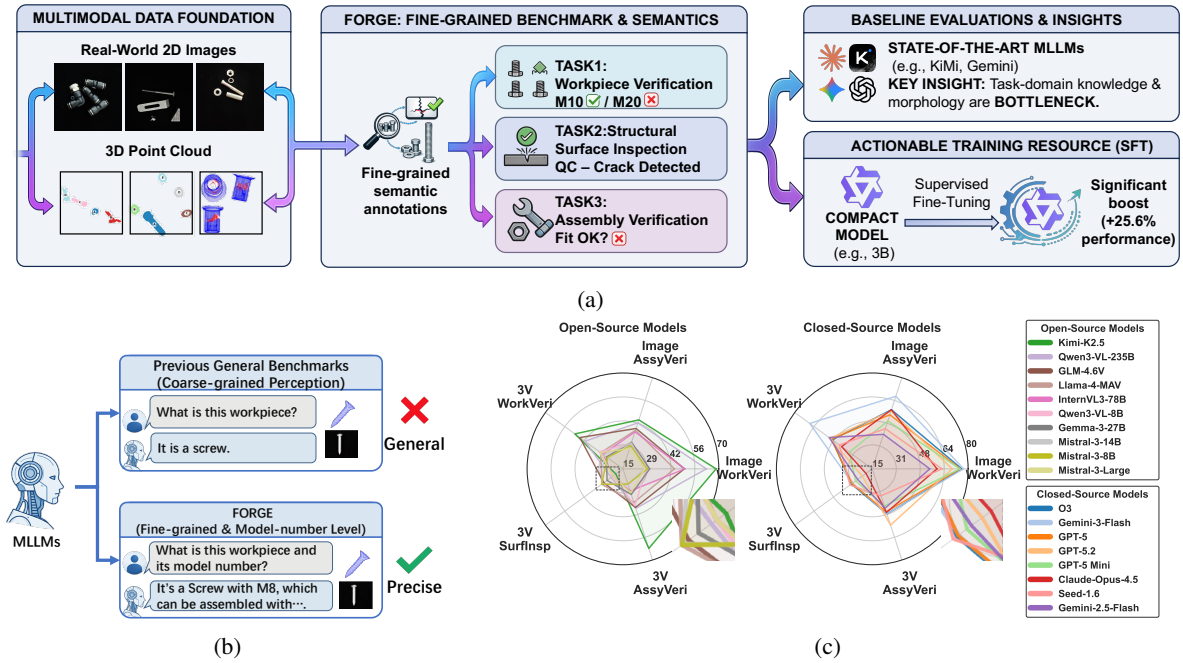


Figure 1: **Benchmark Overview.** (a) The main pipeline of FORGE. (b) Compared to previous benchmarks, FORGE primarily evaluates the understanding, reasoning, and decision-making capabilities regarding fine-grained domain semantics in manufacturing scenarios. (c) The performance of all models across all given tasks. From left to right: open-source models and closed-source models.

sis (Xu et al., 2026b; Zhao et al., 2026), scientific research (Pang et al., 2025), GUI navigation (Jian et al., 2026a; Nayak et al., 2025) and many other fields (Li et al., 2025). Despite this proven versatility, their application in manufacturing remains nascent. The integration of MLLMs presents a critical, yet underexplored, research avenue for manufacturing scenarios (Lee and Su, 2023; Picard et al., 2025; Zhang et al., 2025). MLLMs can overcome the architectural bottleneck of traditional vision models by integrating heterogeneous data streams with advanced reasoning capabilities. This unique synthesis enables MLLMs to bridge the gap between low-level perception and high-level reasoning, supporting a transition in manufacturing intelligence from information extraction toward decision support (Du et al., 2025). Motivated by this potential, we pose the following core inquiry: **Can MLLMs understand and explain the fine-grained judgements required by quality-assurance tasks that are characteristic of the manufacturing domain?**

Recent studies have begun to address this question, as summarized in Table 1. Prior work has examined the performance of MLLMs on specific manufacturing tasks, such as visual anomaly detection (Jiang et al., 2024), engineering documentation comprehension (Doris et al., 2025), and broader manufacturing cognition (Yi et al., 2025). However, current evaluations rarely assess MLLMs' understanding of fine-grained domain semantics, and existing benchmarks only partially reflect the precision demanded by the real-world manufacturing domain (Gao et al., 2024; Boysen et al., 2009; Zhao et al., 2025). Progress in this domain is currently hindered by several fundamental challenges: **(i) Data Scarcity Gap.** Current manufacturing datasets are constrained by limited scale and diversity, so many studies rely on simulated or CAD-based data (Khan et al., 2025; Xu et al., 2026a). **(ii) Lack of Fine-Grained Domain Semantics.** Many current manufacturing datasets merely treat manufacturing workpieces as generic visual subjects. They fail to integrate explicit, fine-grained domain semantics (*e.g.*, exact model numbers following ISO/DIN/GB conventions, such as DIN912 M6×35) that are essential to the rigorous demands of real-world manufacturing. **(iii) Limited Fine-Grained Evaluation Frameworks.** Existing benchmarks rarely provide the *model-number-level* supervision needed to assess whether MLLMs meet the precision requirements of manufacturing quality assurance.

To address these challenges, we introduce **FORGE**, a benchmark tailored to the quality-assurance stage of discrete manufacturing. First,

Table 1: **Comparison of existing manufacturing MLLMs benchmarks with FORGE.**

Benchmarks	Data Modality		Data Source	Granularity †			Statistics	
	Image	Point cloud	Real / Synthetic ‡	Scenario	Workpiece	Model number	Availability	Samples
MMAD (Jiang et al., 2024)	✓	✗	Real-world	✓	✓	✗	✓	39,672
MME-Industry (Yi et al., 2025)	✓	✗	Real-world	✗	✗	✗	✓	1,050
DesignQA (Doris et al., 2025)	✗	✗	Synthetic	✓	✗	✗	✗	1,451
FailureSensorIQ (Constantinides et al., 2025)	✗	✗	Real-world	✓	✗	✗	✓	8,296
EngDesign (Guo et al., 2025)	✗	✗	Synthetic	✓	✗	✗	✓	1,717
FORGE	✓	✓	**Real-world**	✓	✓	✓	✓	**12,972**

‡ *Real / Synthetic* refers to the origin of the workpiece data. In FORGE all images and point clouds are captured from physical workpieces; the surface defects of SURFINSP are synthesized on those real scans (Table 2).
† For Granularity, *Workpiece-level* focuses on manufacturing workpieces; *Scenario-level* focuses on specific manufacturing scenarios; *Model-number-level* requires not only identifying the workpiece type but also the model number.

we collect, construct, and annotate a multimodal manufacturing dataset comprising image and point cloud data of representative workpieces across diverse **model numbers** (*e.g.*, nuts ranging from M10 to M20), thereby capturing the **fine-grained domain semantics** of the real-world manufacturing domain. Furthermore, we adopt three evaluation tasks aligned with the quality-control chain of a production line, namely **material sorting**, **quality inspection**, and **assembly recognition**. Therefore, FORGE investigates these fundamental capabilities that underpin manufacturing-oriented visual understanding, while more advanced aspects of the manufacturing life-cycle are left for future investigation. Beyond benchmarking, we further propose a dedicated dataset for domain-specific fine-tuning. MLLMs fine-tuned on this dataset achieved substantial performance gains on unseen manufacturing scenarios, providing preliminary evidence that the dataset helps bridge the domain knowledge gap in manufacturing settings.

In summary, our major contributions include:

- **High-Quality Multimodal Manufacturing Dataset.** We present a fine-grained manufacturing dataset spanning real-world 2D images and 3D point clouds. The dataset provides rich multimodal annotations designed to support systematic evaluation and development of MLLMs for manufacturing perception and reasoning.

- **Real-World Manufacturing Cognitive Tasks.** Based on the collected fine-grained manufacturing dataset, we design three core manufacturing tasks, **Workpiece Verification (WORKVERI)**, **Structural Surface Inspection (SURFINSP)**, and **Assembly Verification (ASSYVERI)**, which demand MLLMs to perform fine-grained visual discrimination (in Figure 1b) and rule-based reasoning (*e.g.*, verifying assembly compatibility).

- **Extensive Benchmarking and Critical Insights.** We evaluate 18 state-of-the-art MLLMs under three evaluation settings. Our experiments reveal significant performance gaps in microscopic surface analysis and manufacturing task reasoning, and our bottleneck analysis points to morphology understanding and domain knowledge, rather than visual grounding, as the principal limitations in these settings.

- **Actionable Training Resource.** Beyond evaluation, we demonstrate that our structured annotations can serve as training data for domain-specific fine-tuning. Supervised fine-tuning (SFT) (Ouyang et al., 2022) of a compact open-weight model on our training split yields substantial accuracy improvements on held-out manufacturing scenarios unseen during training, providing preliminary evidence that the dataset can help close the domain knowledge gap identified by our analysis.

2 Related Work

Evaluating MLLMs in manufacturing scenarios is a critical yet nascent area. MMAD (Jiang et al., 2024) establishes a standardized framework for visual anomaly detection and evaluates fine-grained perceptual ability on datasets such as MVTec-AD (Bergmann et al., 2019). However, manufacturing demands more than visual pattern recognition, as it requires domain-specific knowledge and reasoning. Recent works expand into specific cognitive domains: MME-Industry (Yi et al., 2025) addresses manufacturing cognition and safety regulations; DesignQA (Doris et al., 2025) focuses on technical blueprints and standards; EngDesign (Guo et al., 2025) targets design synthesis and constraint trade-offs; and FailureSensorIQ (Con-

stantinides et al., 2025) evaluates reliability engineering and failure diagnosis.

Despite this progress, existing methods rarely validate adherence to highly structured and standardized manufacturing settings, which impose more stringent precision demands (Gao et al., 2024; Boysen et al., 2009; Zhao et al., 2025; Maji et al., 2013; Yadav and Jayswal, 2018; Jiao et al., 2007; Gray et al., 1993). Paradigms such as mixed-model assembly lines (Boysen et al., 2009), flexible manufacturing systems (Yadav and Jayswal, 2018), product family design (Jiao et al., 2007), and tool management (Gray et al., 1993) all require not only classifying the general workpiece category but also distinguishing specific variants, demanding precise matching of fine-grained attributes such as model numbers. Current frameworks, constrained by the scarcity of multimodal data and limited fine-grained semantic annotations, fail to meet these requirements. We propose **FORGE**, built on a **multi-modal** dataset with **fine-grained annotations** that include various **workpiece attributes (*e.g.*, model numbers)**. Three evaluation tasks are designed to assess MLLMs against these strict standards, reflecting the fine-grained precision that real-world manufacturing quality assurance requires.

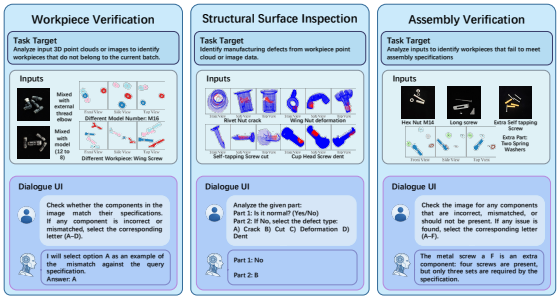


Figure 2: Task descriptions, input data, and dialogue examples of FORGE.

3 FORGE

3.1 Dataset Curation

Dataset Collection. We constructed a comprehensive dataset comprising raw data from authentic manufacturing components. The dataset comprises two subsets. **(i) 3D Point Cloud Subset**: Contains high-fidelity geometric data covering 14 workpiece categories across 90 distinct models. This subset supports tasks such as WORKVERI, SURFINSP, and ASSYVERI. **(ii) Image Subset**: Consists of approximately 3,000 images capturing four distinct manufacturing scenarios (*e.g.*, expansion screw assemblies), including both normal and abnormal samples. All data were captured with fine-grained geometric and visual details.

Table 2: Provenance of FORGE. Only SURFINSP defect regions are synthesized; three-view images are a deterministic rendering, not generated content.

Component	Origin
Image subset (4 scenarios)	Real photographs
Raw point clouds (14 categories)	Real 3D scans
WORKVERI/ASSYVERI point-cloud assemblies	Composed real scans
SURFINSP defects	**Synthesized** on real scans
Three-view images	Rendered from real scans

Dataset Processing. The raw data, comprising both 2D images and 3D point clouds, required preprocessing before being utilized in FORGE. For 2D images, ground-truth labels were established through a two-step process: automated contour and coordinate extraction, followed by manual refinement. For 3D point clouds, strategies varied by task. For WORKVERI and ASSYVERI, we synthesized batch samples by stitching 4-5 individual point clouds with random orientations, automatically generating labels during assembly. For SURFINSP, we simulated four typical manufacturing defects (Crack, Deformation, Dent, and Cut) using morphology-based algorithms and non-rigid deformation to ensure realism. Table 2 summarises what in FORGE is physically measured and what is produced computationally: every photograph and every raw point cloud is an original capture of a physical workpiece, the WORKVERI and ASSYVERI assemblies are compositions of such captures, and the only synthesized content is the defect region of SURFINSP samples. A per-component version with acquisition details is given in Table 10.

The primary objective of FORGE is to evaluate general MLLMs, as their advanced reasoning and versatility make them promising for manufacturing decision support. A central challenge in this evaluation is that these models typically lack native 3D encoders. As validated by our bottleneck analysis (**C** in Section 4.4), directly feeding raw 3D data via text-based coordinate serialization is highly ineffective. To bridge this modality gap, we adopt a multi-view projection strategy: all 3D point cloud samples are rendered as three-view (3V) images, i.e. front, side, and top orthogonal projections, which is also the convention used by engineering drawings to specify manufacturing geometry. This approach preserves the essential

geometric structure across complementary viewpoints while remaining fully compatible with standard visual inputs. FORGE therefore evaluates 3D understanding *through the projection interface available to general-purpose MLLMs* rather than native 3D perception, and we measure how this interface compares with the alternatives: on workpieces present in all three input channels, specialized 3D–language models given the raw clouds score below both serialized coordinates and three-view rendering (Table 17), so among the channels currently available to instruction-following models the projection is the most effective one. The final dataset comprises 12,972 samples across all tasks. Detailed specifications of the dataset statistics, specific object categories, provenance, and the data collection process are provided in the Appendix A.3.

3.2 Task Description

FORGE adopts three tasks aligned with the manufacturing-automation pillars of **material sorting**, **quality inspection**, and **assembly recognition** (Zhao et al., 2025; Wang et al., 2018): **Workpiece Verification (WORKVERI)**, **Structural Surface Inspection (SURFINSP)**, and **Assembly Verification (ASSYVERI)**. The straightforward and direct mapping between real factory scenarios and FORGE tasks is shown in Figure 6 (Appendix). For SURFINSP, a workpiece is considered normal only if it is free from all four defect types (Crack, Cut, Deformation, Dent). For WORKVERI and ASSYVERI, we further distinguish two mismatch granularities: *Different workpiece* (coarse-grained, *e.g.*, wrong workpiece type or missing component) and *Different Model Number* (fine-grained variants within the same production line, such as different screw lengths or thread pitches). An example input/output for each task is shown in Figure 2; additional examples appear in Figures 7a, 7b, and 7c of Appendix A.3.4.

Task 1: WORKVERI. WORKVERI is designed to evaluate how effectively MLLMs perform material sorting. Given explicit specifications, MLLMs must analyze 3D point clouds or images to identify workpieces that do not belong to the current batch. This task includes **three** scenarios: one from the image subset, pneumatic connectors (PCS SCENARIO), and two point cloud datasets, cup head screws (CHS SCENARIO) and nuts (NUTS SCENARIO).

Task 2: SURFINSP. The core objective of SURFINSP is to identify manufacturing defects from workpiece point cloud or image. The task involves two steps: (1) defect detection and (2) defect type classification (*e.g.*, Crack, Dent). Our evaluation covers **14** distinct manufacturing components with 3D point cloud.

Task 3: ASSYVERI. ASSYVERI aims to assess the capability of MLLMs in understanding assembly relationships and compatibility constraints. Unlike sorting, this task requires reasoning over a given complex set of assembly rules. MLLMs must analyze inputs to identify workpieces that fail to meet assembly specifications. This task covers **four** scenarios involving three from the image subset: metal expansion screws (MES SCENARIO), plastic expansion screws (PES SCENARIO), CNC fixtures (CNC SCENARIO), and one for point cloud data for the compatibility among metal screws, washers, and nuts (SWN SCENARIO).

This task design establishes a challenging benchmark that goes beyond basic perceptual evaluation, rigorously assessing the fine-grained logical reasoning and judgment capabilities of MLLMs in real-world manufacturing scenarios. Detailed task description and examples are provided in the Appendix A.3.

4 Experiments

We evaluate 18 MLLMs on three manufacturing tasks, report main results with error case analysis, perform a bottleneck analysis to disentangle visual-perception from domain-knowledge limitations, and show the training potential of our benchmark by fine-tuning small models for cross-scenario generalization.

4.1 Evaluated Models

To provide a comprehensive assessment of the current MLLM landscape, we evaluate 18 representative models across open- and closed-source families (Table 6) in Appendix A.2.

4.2 Evaluation Settings and Metrics

Evaluation Protocol. All tasks in FORGE are formulated as multiple-choice questions (MCQs). For WORKVERI and ASSYVERI, each assembly contains 4–6 components. In image-based evaluation, each MCQ option corresponds to a part identified by its normalized center coordinate (*e.g.*, "A. Part at [0.70, 0.44]"), enabling the model to ground each choice to a specific spatial location in

the image. In three-view evaluation, components are annotated with letter labels (A–F) using the Set-of-Mark (Yang et al., 2023) visual prompting strategy. In both cases, the model must select the letter corresponding to the anomalous component. For SURFINSP, the model must classify the surface condition of a given workpiece into one of five categories (*crack*, *cut*, *deformation*, *dent*, or *good*).Additionally, error scenarios in WORKVERI and ASSYVERI are categorized into two difficulty levels: *Different workpiece* (coarse-grained discrepancies, *e.g.*, entirely wrong workpiece types or missing components) and *Different Model Number* (fine-grained inconsistencies involving subtle variations within the same production line, such as different screw lengths or thread pitches).

We evaluate under three progressively informative settings: **i. Zero-Shot.** The model receives only the test image (or three-view rendering) and a task-specific query. No additional examples or references are provided. This setting measures the model's inherent ability to perform the task using only its pretrained knowledge and visual understanding. **ii. Reference-Conditioned (Ref-Cond).** In addition to the test case, the model is provided with reference images of correct, normal assemblies (or defect-free surfaces for SURFINSP). These references establish a visual baseline for what "correct" looks like, enabling the model to detect deviations by comparison. This setting evaluates whether explicit visual references improve anomaly detection. **iii. In-Context Demonstration (ICD).** Building upon Ref-Cond, the model additionally receives complete solved examples presented as multi-turn dialogue pairs, each containing a query, input data, and the correct answer. This setting tests whether full reasoning demonstrations can bridge the gap in domain-specific task understanding. We generally use 2 ICD examples by default, while comprehensive setting details and further examples, including task-specific adjustments, can be found in Appendix E.

Metric. We adopt **exact-match accuracy** as the evaluation metric. For each test case, the model's predicted MCQ letter is extracted from its free-form response and compared with the ground-truth label. Accuracy is computed as the percentage of cases where the prediction exactly matches the correct answer. The random-chance baseline is the weighted average of a random guess across tasks and serves as a critical benchmark.

4.3 Research Results and Analysis

Table 3 summarizes the main results across all tasks, modalities, and evaluation settings. Detailed extended results are provided in the Appendix A.2. From these results, we distill three key findings on current MLLM capabilities and limitations in manufacturing.

A. Current MLLMs demonstrate better understanding of semantics than in morphological analysis. On WORKVERI and ASSYVERI, Kimi-K2.5 dominates the open-source group and Gemini-3-Flash leads the closed-source group across most settings (Table 3, Figure 1c); the picture changes on SURFINSP, where Llama-4-MAV leads the open-source group without demonstrations and the Mistral family takes the top positions under ICD. Despite SURFINSP's relatively simple objectives, it yields the lowest accuracy among the three tasks. This disparity indicates that MLLMs exhibit fundamentally different capabilities for macroscopic part discrimination and microscopic surface morphology analysis, highlighting significant room for improvement in their understanding of physical details and microscopic features of manufacturing workpieces. Accuracy alone also understates how far these models are from inspection use, because it weights both error directions equally. Decomposing SURFINSP into miss rate and false-alarm rate (Section A.2.1) shows that models with nearly identical accuracy fail in opposite, cost-relevant ways: under ICD, Claude-4.5-Opus misses 60.5% of true defects while Qwen3-VL-235B instead flags 61.8% of defect-free workpieces, at accuracies of 23.9% and 22.7% respectively.

B. Limited comprehension of domain knowledge is the bottleneck for current MLLMs. By comparing evaluation results across Zero-Shot, Ref-Cond, and ICD-based settings based on image modality, we observed that in WORKVERI and ASSYVERI, simple Ref-Cond strategies did not consistently yield performance gains and, in several instances, even led to degradation. Conversely, ICD methods incorporating complete reasoning demonstrations improved over Zero-Shot baselines for most models, and in particular for the stronger ones. This suggests that for complex manufacturing tasks, MLLMs do not lack simple sample references, but rather a deep understanding of task logic and reasoning paths, a gap that ICD bridges but Ref-Cond cannot. Notably, Section 4.4 shows that visual grounding is not the bottleneck: MLLMs can

Table 3: Main benchmark results (accuracy %) across three manufacturing tasks. **Bold orange** / underlined blue / light purple indicate best / second / third within each model category (open- or closed-source) per row. **Rand.** = weighted random-guess baseline.

Task	Mod.	Setting	Open-source Models										Closed-source Models								Random
			Gemma-3-27B	Intern VL3-78B	Llama-4-MAV	Mistral-3-14B	Mistral-3-8B	Mistral-3-Large	Qwen3-VL-235B	Qwen3-VL-8B	Kimi-K2.5	GLM-4.6V	Claude-4.5-Opus	Gemini-2.5-Flash	Gemini-3-Flash	GPT-5.2	GPT-5	GPT-5 Mini	O3	Seed-1.6	
WorkVeri	3V	Zero Shot	27.62	32.59	36.36	32.46	30.97	34.78	**52.36**	42.01	50.00	47.42	39.19	56.05	**69.56**	61.29	61.05	37.98	58.22	54.44	25.0
		Ref-Cond	23.59	24.90	20.36	25.00	18.26	20.93	35.67	27.20	36.76	**44.29**	27.88	43.55	**65.25**	28.34	30.02	33.94	29.55	26.87	25.0
		ICD	28.83	30.16	35.69	28.54	28.86	37.53	37.86	26.32	**60.56**	44.90	55.04	53.85	**67.34**	61.69	56.68	30.71	57.91	46.45	25.0
	Image	Zero Shot	25.94	32.59	38.80	33.56	30.61	25.78	64.08	35.41	**66.75**	50.44	59.42	55.78	72.22	72.50	74.72	73.78	**76.18**	67.04	25.0
		Ref-Cond	30.51	53.88	40.65	20.22	24.28	25.50	58.01	24.04	**69.67**	50.33	56.22	51.66	**76.27**	55.48	64.22	72.73	68.32	49.66	25.0
		ICD	34.59	65.26	54.19	27.20	32.07	31.18	66.44	25.28	**68.86**	50.00	60.98	54.55	82.26	79.87	**85.23**	77.01	82.49	70.29	25.0
AssyVeri	3V	Zero Shot	27.18	34.30	38.19	28.16	24.60	31.64	42.33	39.16	**78.26**	39.60	42.07	44.98	47.25	**54.22**	53.90	49.51	52.38	41.69	32.8
		Ref-Cond	32.04	22.01	30.10	19.09	18.45	29.54	31.01	33.77	**57.14**	37.29	50.81	40.13	46.28	**56.96**	30.74	29.55	36.46	26.71	32.8
		ICD	32.04	28.80	37.54	26.06	29.55	27.03	43.26	38.96	**53.54**	37.86	45.45	44.01	47.57	**54.07**	51.95	45.31	52.46	34.30	32.8
	Image	Zero Shot	28.47	33.14	36.42	30.15	29.64	29.12	**36.97**	31.89	33.83	35.46	52.10	39.56	**58.11**	48.18	50.06	43.69	48.93	39.86	29.5
		Ref-Cond	31.70	36.40	32.12	32.12	26.84	23.74	40.82	28.50	**52.08**	37.28	56.36	32.32	**70.61**	56.28	49.77	51.35	60.16	42.49	29.5
		ICD	32.94	42.46	46.13	29.46	30.65	28.67	**50.43**	30.32	47.72	45.33	62.92	46.89	**71.50**	63.99	60.94	52.29	62.34	48.52	29.5
SurfInsp	3V	Zero Shot	20.96	19.64	**25.75**	23.25	22.41	19.64	20.24	22.05	19.07	21.33	22.05	18.31	18.80	20.48	**23.98**	20.36	21.45	21.45	20.0
		Ref-Cond	28.19	18.92	**29.60**	24.94	24.70	20.24	20.96	23.19	22.24	25.91	24.58	28.31	31.69	22.41	35.78	35.78	**36.43**	35.54	20.0
		ICD	18.80	23.37	22.07	36.02	**38.55**	26.39	22.67	26.99	18.67	21.23	23.86	44.34	**49.64**	30.96	41.38	36.39	40.72	38.92	20.0

identify these workpieces, revealing a disconnect between perception and comprehension.

C. Given limited perceptual understanding of 3D spatial contexts, the introduction of additional examples hinders comprehension of MLLMs. Based on Table 3, for the three-view input on WorkVeri and AssyVeri, MLLMs achieved optimal performance under the Zero-Shot setting, whereas performance declined following the introduction of Ref-Cond and ICD. SurfInsp is the exception: its objective does not require cross-part comparison, and there added references and demonstrations help rather than hurt. Given that the fundamental distinction between the three-view and image inputs lies in their multi-angular spatial representations, this counterintuitive phenomenon suggests that the contextual examples provided in Ref-Cond and ICD do not provide effective guidance. Instead, they induced spatial confusion within MLLMs, thereby impeding comprehension of manufacturing domain knowledge. Furthermore, the experimental results regarding point cloud inputs in the Bottleneck Analysis from Appendix C.2 provide additional evidence of current MLLMs' limitations in explicit spatial perception capabilities.

4.4 Bottleneck Analysis

Before the systematic decomposition, qualitative inspection of error cases (Figure 3) reveals two recurring patterns: the case in MES scenario fails due to overreliance on inferred material properties, while the case in CNC scenario shows emerging service-condition awareness (*e.g.*, recognizing wear/chipping); a full case discussion is provided in Appendix C.1.

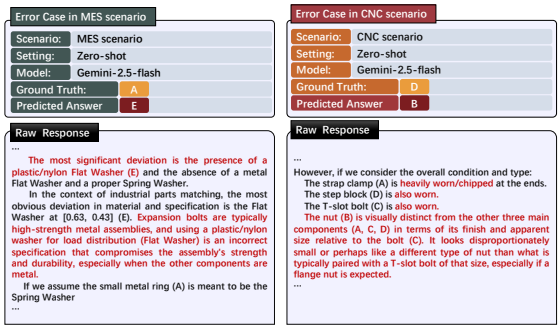


Figure 3: Two error cases of AssyVeri.

Beyond these qualitative observations, our main benchmark requires multi-stage reasoning: a model must first *ground* individual components, optionally *compare* across reference and test images, and finally apply *domain-specific logic* to reach a verdict. When a model fails, it is unclear which stage is the bottleneck. To disentangle these factors, we design three complementary bottleneck analyses, evaluating five representative models: Gemini-3-Flash, GPT-5.2, Qwen3-VL-235B, Seed-1.6, and Mistral-3-8B (details in Appendix C.2).

A. Visual grounding is not the bottleneck. To disentangle perceptual failures from reasoning failures, we isolate each model's visual grounding ability using dedicated probing tasks. In our benchmark, each assembly image is annotated with letter labels (A, B, C, *etc.*) on individual parts, *i.e.*Set-of-Mark (Yang et al., 2023) trick. We test whether models can correctly map between these labels and spatial coordinates in two settings: (i) *Single-image*

Table 4: **Visual grounding Bottleneck Analysis results (accuracy %).** *Single-image*: coordinate-to-letter (C→L) and letter-to-coordinate (L→C) on 500 cases. *Cross-image*: letter-to-letter (L→L) and coordinate-to-coordinate (C→C) on 513 pairs.

Model	Type	Single-Image			Cross-Image		
		C→L	L→C	Avg.	L→L	C→C	Avg.
Gemini-3-Flash	Closed	**98.2**	**99.6**	**98.9**	**88.7**	**79.9**	**84.3**
GPT-5.2	Closed	74.6	97.6	86.1	85.6	75.4	80.5
Qwen3-VL-235B	Open	85.4	98.8	92.1	80.3	72.2	76.3
Seed 1.6	Closed	42.0	99.2	70.6	79.3	71.2	75.2
Mistral-3-8B	Open	66.0	70.6	68.3	62.0	33.9	48.0

grounding: given one annotated image, the model must either locate a part from its letter (**L→C**) or identify a letter from given coordinates (**C→L**); (ii) *Cross-image correspondence*: given two annotated images of assemblies from the same scenario, a part is identified in the first image and the model must find the visually corresponding part in the second image, using either letters (**L→L**) or coordinates (**C→C**). If models perform well on these grounding tasks but poorly on the full benchmark, the bottleneck lies in domain reasoning rather than perception. Presented in Table 4, Gemini-3-Flash achieves 98.9% average on single-image grounding, and four of five models reach at least 97.6% on L→C, the direction most relevant to our benchmark. These near-ceiling results confirm that failures on the full WORKVERI and ASSYVERI evaluations cannot be attributed to poor visual localization. Cross-image comparison is harder (84.3% for Gemini, 80.5% for GPT-5.2) but remains well above chance for the top four models on letter-based matching (at least 79.3%), indicating it is a *contributing* but not *dominant* factor in the performance gap between Zero-Shot and Ref-Cond/ICD settings, as it validates MLLMs' ability to compare between images.

B. Fine-grained identification remains a domain-knowledge bottleneck. To further disentangle *domain-specific reasoning* from explicit *ground* capabilities, we conduct a bottleneck analysis focusing on the missing part scenario. In this setup, each model is provided with an explicit assembly specification (*i.e.*, a comprehensive list of required parts, their counts, and functional descriptions) and must identify the absent component via an MCQ. This directly probes domain knowledge: detecting that "something is missing" requires only counting, but pinpointing *which* part is absent demands understanding the visual and functional distinctions between components. As shown in Table 5 (The column headers indicate the specific missing workpiece type in that scenario), the top four models achieve 74.9–90.7% overall accuracy on images, which is well above the 23.3% random baseline and thus demonstrates that MLLMs can reason about assembly completeness when given structured descriptions. Performance is near-perfect on most part types where components are visually distinctive (*i.e.*, screws, nuts, anchors, wedges). However, a systematic failure is observed in flat washer detection (23.3–60.0% on images, 8.3–74.5% on three-view), where all five models struggle to some extent. Error analysis reveals the models can reliably detect that *a washer* is absent but cannot determine *which* washer, despite the two having distinct physical forms (further discussion and examples in the Appendix C.2). Since the grounding ability analysis (Table 4) confirms that the four strongest of these models can accurately localize individual components, these confusion points indicate insufficient fine-grained manufacturing knowledge of the functional and morphological differences between part variants rather than a perceptual failure. A secondary finding concerns normal-case recognition: Seed 1.6 achieves only 43.3% on normal cases overall (as low as 10% on MES SCENARIO) while maintaining 89.5% on most missing-part subcases, revealing a bias toward predicting a missing component rather than confirming completeness, again a reasoning rather than perception limitation.

Table 5: **Zero-shot missing-part detection.** Superscripts denote scenario: [1]MES SCENARIO (6 options), [2]PES SCENARIO (3 options), [3]CNC SCENARIO (5 options) . Three-view covers SWN SCENARIO only (5 options). FW=Flat Washer, SW=Spring Washer, Sc=Screw, An=Anchor, Nu=Nut, We=Wedge, Norm=No missing part.

Model	Image (240 cases)									Three-View (137)			
	FW[1]	SW[1]	Sc[2]	An[2]	Nu[3]	Sc[3]	We[3]	Norm	All	FW	SW	Norm	All
Gemini-3-Flash	36.7	**100**	**100**	96.7	**100**	**100**	95.0	**98.3**	**90.7**	8.3	84.6	**100**	63.5
GPT-5.2	**60.0**	83.3	90.0	96.7	**100**	**100**	85.0	80.0	85.0	**74.5**	87.2	**100**	**87.5**
Qwen3-VL-235B	23.3	**100**	**100**	**100**	95.0	**100**	**100**	85.0	86.2	29.2	84.6	86.0	65.7
Seed 1.6	26.7	**100**	**100**	**100**	90.0	89.5	**100**	43.3	74.9	41.7	**89.7**	88.0	72.3
Mistral-3-8B	36.7	40.0	44.8	23.3	5.0	0.0	80.0	8.3	27.2	10.4	56.4	4.0	21.2

C. Raw point cloud text is not a viable substitute for visual projection.

Two complementary analyses further examine the bottleneck of 3D input representation, including experiments that feed 3D coordinates directly to MLLMs and experiments with specialized 3D–language models operating on native point clouds (Appendix C.2). Serialized coordinates provide only limited evidence of geometric understanding, achieving at most 22.6% on SURFINSP and 53.6% on WORKVERI. To isolate the effect of the input representation, we further compare three input modalities using the same 3D encoder and identical workpieces. Native point clouds perform the worst (12.9%), followed by serialized coordi-

nates (20.3%), while three-view rendering achieves the best performance (23.4%) under the same encoder, and reaches 28.4% when considering all 18 MLLMs. These results suggest that, despite providing less explicit 3D information, visual projection through three-view rendering offers a more effective interface for MLLMs to exploit geometric information than directly exposing native or serialized 3D coordinates.

4.5 From Benchmark to Training Resource

The preceding analyses identify insufficient manufacturing domain knowledge as the primary bottleneck (Section 4.4). We investigate whether FORGE annotations can also serve as an *actionable training resource* to close this gap. We fine-tune Qwen2.5-VL-3B-Instruct using task-specific *full-parameter* SFT (3 epochs, effective batch size 32, learning rate 1×10^{-5}, bfloat16, ViT frozen) and adopt a *scenario-based* train/eval split: for WORKVERI three-view, training on the CHS SCENARIO and evaluating on the held-out NUTS SCENARIO, and for ASSYVERI image, training on the MES SCENARIO and PES SCENARIO and evaluating on the held-out CNC SCENARIO. This out-of-distribution protocol ensures that observed improvements reflect genuine acquisition of transferable manufacturing reasoning rather than memorization of specific assembly layouts. Full configuration details and additional training results are provided in the Appendix B.

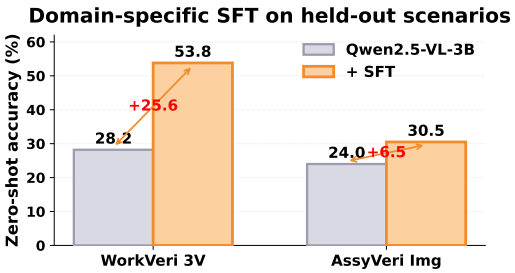


Figure 4: Domain-specific SFT on held-out scenarios (zero-shot acc. %).

As shown in Figure 4, SFT yields a **90.8%** relative improvement on WORKVERI three-view, bringing the 3B model level with Qwen3-VL-235B (54.4%), a model with 78× more total parameters (∼7× more active parameters via its A22B mixture-of-experts architecture), although the strongest closed-source model on this held-out scenario remains ahead (63.6%). On ASSYVERI image, SFT achieves a 27.1% relative gain (24.0 to 30.5), surpassing all reference models except Gemini-3-Flash and GPT-5.2, but a substantial gap to Gemini-3-Flash (53.5%) remains; this is also where the benefit of fine-tuning on FORGE is smallest, as only 627 training examples are available for this scenario. Because these gains are measured on product categories absent from training, they indicate that FORGE annotations encode *transferable* manufacturing knowledge. We therefore regard this as *preliminary* evidence that FORGE is not merely a static benchmark but an **actionable resource**: even modest amounts of domain-specific training data can substantially improve compact models on out-of-distribution manufacturing scenarios.

5 Conclusion

We present FORGE, a fine-grained multimodal benchmark of real-world 2D images and 3D point clouds covering three quality-assurance tasks of discrete manufacturing.Evaluating 18 state-of-the-art MLLMs, we find that current models handle macroscopic part recognition but fall short on model-number-level reasoning and microscopic surface analysis. Our evaluation and bottleneck analysis indicate that visual grounding is not the primary limiting factor; rather, insufficient manufacturing domain knowledge and morphology understanding are the key gaps. Beyond evaluation, we show that FORGE serves as an actionable training resource: domain-specific fine-tuning of a compact 3B-parameter model closes most of the gap to a 78× larger open-weight model on one held-out scenario (53.8 vs. 54.4), while a substantial gap to the strongest closed-source models remains on the other (30.5 vs. 53.5). These findings position FORGE as an evaluation baseline and a practical starting point for closing the domain knowledge gap in manufacturing quality assurance.

Limitations

While FORGE provides a rigorous benchmark for manufacturing scenarios, we acknowledge several limitations. First, rendering 3D point clouds into three-view images inevitably sacrifices depth and geometric topology. Although FORGE targets generalizable MLLMs rather than specialized 3D architectures, establishing a clear evaluation boundary, how to more effectively process and analyze raw point cloud data in manufacturing settings remains an important direction for future work. Second, the SFT experiment in FORGE serves as a prelim-

inary validation for domain adaptation. Because SFT optimizes primarily for accuracy, it does not fully address how to internalize complex manufacturing constraints or elicit deeper reasoning capabilities. Future work should explore more principled training paradigms that explicitly target manufacturing reasoning rather than task-level accuracy. Finally, FORGE focuses on fine-grained requirements within isolated tasks. Real-world manufacturing occurs in inherently multi-scale and highly multimodal environments. Expanding the evaluation framework to encompass these complex, dynamic interactions, as well as other stages of the manufacturing life-cycle such as predictive maintenance, assembly execution, and safety compliance, remains a critical next step. Finally, our conclusions rest on single-run evaluations of proprietary model endpoints, and the bottleneck analyses probe five representative models rather than the full set of 18; both should be kept in mind when interpreting individual numbers.

Acknowledgments

The authors would like to express their sincere gratitude to Mr. Richy Yu Hu from the Machining and Fabrication Workshop at The Hong Kong University of Science and Technology (Guangzhou) for the valuable assistance with data collection. This research is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office. Any opinions, findings and conclusions or recommendations expressed in these materials are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore as well as CyberSG R&D Programme Office, Singapore.

This work was also supported in part by the Research Grants Council of the Hong Kong SAR under Grant GRF 11217823, 11216225, and Collaborative Research Fund C1042-23GF, the National Natural Science Foundation of China under Grant 62371411, InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, the Lee Kong Chian Fellowship, the National Natural Science Foundation of China under Grant No. 72371219, and the Guangdong Project under Grant No. 2024TQ08A432.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600.

Nils Boysen, Malte Fliedner, and Armin Scholl. 2009. Sequencing mixed-model assembly lines: Survey, classification and model critique. *European Journal of Operational Research*, 192(2):349–373.

Christodoulos Constantinides, Dhaval Patel, Shuxin Lin, Claudio Guerrero, Sunil Dagajirao Patil, and Jayant Kalagnanam. 2025. FailureSensorIQ: A multi-choice QA dataset for understanding sensor relationships and failure modes. *arXiv preprint arXiv:2506.03278*.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer.

Anna C Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. 2025. Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *Journal of Computing and Information Science in Engineering*, 25(2):021009.

Kaze Du, Bo Yang, Keqiang Xie, Nan Dong, Zhengping Zhang, Shilong Wang, and Fan Mo. 2025. LLM-MANUF: An integrated framework of fine-tuning large language models for intelligent decision-making in manufacturing. *Advanced Engineering Informatics*, 65:103263.

Haolin Fan, Chenshu Liu, Neville Elieh Janvisloo, Shijie Bian, Jerry Ying Hsi Fuh, Wen Feng Lu, and Bingbing Li. 2025. Mavila: Unlocking new potentials in smart manufacturing through vision language models. *Journal of Manufacturing Systems*, 80:258–271.

Haolin Fan, Hongji Zhang, Changyu Ma, Tongzi Wu, Jerry Ying Hsi Fuh, and Bingbing Li. 2024. Enhancing metal additive manufacturing training with the advanced vision language model: A pathway to immersive augmented reality training for non-experts. *Journal of Manufacturing Systems*, 75:257–269.

Huihui Gao, Xiaoran Zhang, Xuejin Gao, Fangyu Li, and Honggui Han. 2024. A hierarchical coarse-to-fine fault diagnosis method for industrial processes based on decision fusion of class-specific stacked autoencoders. *IEEE Transactions on Instrumentation and Measurement*.

Anuj Gautam, Manish Raj Aryal, Sourabh Deshpande, Shailesh Padalkar, Mikhail Nikolaenko, Ming Tang, and Sam Anand. 2025. IIoT-enabled digital twin for legacy and smart factory machines with LLM integration. *Journal of Manufacturing Systems*, 80:511–523.

Ann E Gray, Abraham Seidmann, and Kathryn E Stecke. 1993. A synthesis of decision models for tool management in automated manufacturing. *Management Science*, 39(5):549–567.

Wei Guan, Jun Lan, Jian Cao, Hao Tan, Huijia Zhu, and Weiqiang Wang. 2025. EMIT: Enhancing MLLMs for industrial anomaly detection via difficulty-aware GRPO. *arXiv preprint arXiv:2507.21619*.

Xingang Guo, Yaxin Li, XiangYi Kong, YILAN JIANG, Xiayu Zhao, Zhihua Gong, Yufan Zhang, Daixuan Li, Tianle Sang, Beixiao Zhu, Gregory Jun, Yingbing Huang, Yiqi Liu, Yuqi Xue, Rahul Dev Kundu, Qi Lim, Yizhou Zhao, Luke Granger, Mohamed Younis, and 46 others. 2025. Toward engineering agi: Benchmarking the engineering design capabilities of llms. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xiangru Jian, Shravan Nayak, Kevin Qinghong Lin, Aarash Feizi, Kaixin Li, Patrice Bechard, Spandana Gella, and Sai Rajeswar. 2026a. CUA-suite: Massive human-annotated video demonstrations for computer-use agents. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

Xiangru Jian, Wei Pang, Zhengyuan Dong, Chao Zhang, and M Tamer Özsu. 2026b. LazyVLM: Neuro-symbolic approach to video analytics. In *IEEE International Conference on Data Engineering (ICDE), Demo Track*.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. 2024. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. *arXiv preprint arXiv:2410.09453*.

Jianxin Jiao, Timothy W Simpson, and Zahed Siddique. 2007. Product family design and platform-based product development: a state-of-the-art review. *Journal of Intelligent Manufacturing*, 18(1):5–29.

Muhammad Tayyab Khan, Lequn Chen, Ye Han Ng, Wenhe Feng, Nicholas Yew Jin Tan, and Seung Ki Moon. 2025. Leveraging vision-language models for manufacturing feature recognition in computer-aided designs. *Journal of Computing and Information Science in Engineering*, 25(10):104501.

Yejin Kwon, Daeun Moon, Youngje Oh, and Hyunsoo Yoon. 2025. LogicQA: Logical anomaly detection with vision language model generated questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 411–432, Vienna, Austria. Association for Computational Linguistics.

Jay Lee and Hanqi Su. 2023. A unified industrial large knowledge model framework in industry 4.0 and smart manufacturing. *arXiv preprint arXiv:2312.14428*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1578–1597.

Shiwei Lin, Chenxu Wang, Xiaozhen Ding, Yi Wang, Boyuan Du, Lei Song, Chenggang Wang, and Huaping Liu. 2025. A VLM-based method for visual anomaly detection in robotic scientific laboratories. *arXiv preprint arXiv:2506.05405*.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A. Rodriguez, Montek Kalsi, Nicolas Chapados, M. Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. 2025. UI-vision: A desktop-centric GUI benchmark for visual perception and interaction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45817–45851. PMLR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2025. Paper2Poster: Towards multimodal poster automation from scientific papers. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

C. Picard, K. M. Edwards, A. C. Doris, B. Man, G. Giannone, M. F. Alam, and F. Ahmed. 2025. From concept to manufacturing: Evaluating vision-language models for engineering design. *Artificial Intelligence Review*, 58(9):288.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. 2024. ShapeLLM: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision (ECCV)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object

detection with region proposal networks. *Advances in neural information processing systems*, 28.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328.

B Saenz de Ugarte, Abdelhakim Artiba, and Robert Pellerin. 2009. Manufacturing execution system–a literature review. *Production Planning and Control*, 20(6):525–539.

Qiuzhuang Sun, Piao Chen, Xin Wang, and Zhi-Sheng Ye. 2023. Robust condition-based production and maintenance planning for degradation management. *Production and Operations Management*, 32(12):3951–3967.

Qiuzhuang Sun, Zhi-Sheng Ye, and Nan Chen. 2018. Optimal inspection and replacement policies for multi-unit systems subject to degradation. *IEEE Transactions on Reliability*, 67(1):401–413.

Chengyu Tao, Xuanming Cao, and Juan Du. 2025. G2sf: Geometry-guided score fusion for multimodal industrial anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20551–20560.

Chengyu Tao and Juan Du. 2025. Pointsgrade: Sparse learning with graph representation for anomaly detection by using unstructured 3D point cloud data. *IISE Transactions*, 57(2):131–144.

Chengyu Tao, Juan Du, and Tzyy-Shuh Chang. 2023. Anomaly detection for fabricated artifact by using unstructured 3D point cloud data. *IISE Transactions*, 55(11):1174–1186.

Chengyu Tao, Hao Xu, and Juan Du. 2024. F2PAD: A general optimization framework for feature-level to pixel-level anomaly detection. *arXiv preprint arXiv:2407.06519*.

Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. 2018. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48:144–156.

Tian Wang, Junming Fan, and Pai Zheng. 2024. An llm-based vision and language cobot navigation approach for human-centric smart manufacturing. *Journal of Manufacturing Systems*, 75:299–305.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. 2024. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2018. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes.

Hao Xu, Juan Du, Andi Wang, and Yingcong Chen. 2026a. Ano-sups: Multisize anomaly detection for manufactured products by identifying suspected patches via vision transformer. *INFORMS Journal on Data Science*.

Hao Xu, Xiangru Jian, Xinjian Zhao, Wei Pang, Chao Zhang, Suyuchen Wang, Qixin ZHANG, Zhengyuan Dong, Joao Monteiro, Bang Liu, Qiuzhuang Sun, and Tianshu Yu. 2026b. Graphomni: A comprehensive and extensible benchmark framework for large language models on graph-theoretic tasks. In *International Conference on Learning Representations*, volume 2026, pages 116120–116234.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. PointLLM: Empowering large language models to understand point clouds. In *European Conference on Computer Vision (ECCV)*.

Anupma Yadav and SC Jayswal. 2018. Modelling of flexible manufacturing system: a review. *International Journal of Production Research*, 56(7):2464–2487.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *Preprint*, arXiv:2310.11441.

Dongyi Yi, Guibo Zhu, Chenglin Ding, Zongshu Li, Dong Yi, and Jinqiao Wang. 2025. MME-industry: A cross-industry multimodal evaluation benchmark. *arXiv preprint arXiv:2501.16688*.

Zhaolin Yuan, Ming Li, Chang Liu, Fangyuan Han, Haolun Huang, and Hong-Ning Dai. 2025. Chat with MES: LLM-driven user interface for manipulating garment manufacturing system through natural language. *Journal of Manufacturing Systems*.

Haiyang Zhang, Wei Yan, Huicong Hu, Xumei Zhang, Qingtao Liu, Hong Xia, Yingguang Zhang, and Yuhao Lin. 2025. An LLM-based knowledge and function-augmented approach for optimal design of remanufacturing process. *Advanced Engineering Informatics*, 65:103206.

Yiheng Zhang, Yunkang Cao, Xiaohao Xu, and Weiming Shen. 2024. LogiCode: An LLM-driven framework for logical anomaly detection. *IEEE Transactions on Automation Science and Engineering*.

Shuxuan Zhao, Sichao Liu, Yishuo Jiang, Bo Zhao, Youlong Lv, Jie Zhang, Lihui Wang, and Ray Y Zhong. 2025. Industrial foundation models (IFMs) for intelligent manufacturing: A systematic review. *Journal of Manufacturing Systems*, 82:420–448.

Xinjian Zhao, Wei Pang, Zhixuan Yu, Xiangru Jian, Xiaozhuang Song, Yaoyao Xu, Zhongkai Xue, Dingshuo Chen, Shu Wu, Philip Torr, and Tianshu Yu. 2026. When vision meets graphs: A survey on graph reasoning and learning. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*.

Appendix

Contents

A Extended Related Works, Main Results and Task Description **13**

A.1 Related Works 13

A.1.1 CV models in Manufacturing Scenario. 13

A.1.2 MLLMs in Manufacturing Scenario. 14

A.2 Extended Main Results 14

A.2.1 Risk-Sensitive Metrics for SURFINSP. 16

A.2.2 Rationale for the MCQ Formulation. 16

A.3 Task Description 17

A.3.1 Raw Data Collection. . . . 17

A.3.2 Task Description of WORKVERI. 20

A.3.3 Task Description of SURFINSP. 20

A.3.4 Task Description of ASSYVERI. 21

B Training Details and Extended Results **23**

C Extended Error Cases and Bottleneck Analysis **25**

C.1 Qualitative Discussion of Error Patterns 25

C.2 Details of Bottleneck Analysis . . 29

C.2.1 Settings and Configurations. 29

C.2.2 Additional Discussion of Bottleneck Analyses. . . . 31

D Image examples **32**

D.1 The Examples of All Three Tasks . 32

D.2 The Grounding Examples of Image Data (Bottleneck Analysis A) . . . 44

E Data examples **49**

E.1 Examples of WORKVERI 49

E.2 Examples of SURFINSP 54

E.3 Examples of ASSYVERI 56

E.4 Examples of Bottleneck Analysis A 74

E.5 Examples of Bottleneck Analysis B 75

E.6 Examples of Bottleneck Analysis C 76

A Extended Related Works, Main Results and Task Description

A.1 Related Works

A.1.1 CV models in Manufacturing Scenario.

Computer Vision has widely served as the "eyes" of modern manufacturing automation, evolving from traditional image processing to deep learning-based approaches. Existing literature in this domain can be broadly categorized into three primary tasks: surface defect detection, geometric measurement, and robotic guidance.

Surface anomaly detection represents the most extensively studied subfield. Early works used supervised Convolutional Neural Networks (CNNs) to classify defects using large-scale annotated datasets (He et al., 2016; Ren et al., 2015). To mitigate the reliance on defective samples, recent trends have shifted towards unsupervised paradigms, such as reconstruction-based methods (e.g., AutoEncoders, GANs) (Bergmann et al., 2019; Tao et al., 2024) and feature-embedding approaches (Roth et al., 2022), which identify anomalies by measuring deviations from a normal distribution. Recently, to leverage geometric information for more robust detection, researchers have increasingly focused on 3D-aware approaches. *G2SF* (Tao et al., 2025) is a multimodal industrial anomaly detection framework that utilizes geometry-guided score fusion to effectively combine 2D appearance and 3D depth information. Furthermore, to address the challenges of unstructured 3D data, *PointSGRADE* (Tao and Du, 2025) employs graph representations and sparse learning to detect anomalies directly from 3D point clouds.

CV models also play a critical role in geometric metrology and robotic manipulation. Key applications include 6D pose estimation for component assembly (Xiang et al., 2018) and precise dimensional measurement for quality control. These models provide spatial coordinates and geometric parameters, enabling automated systems to execute pre-defined mechanical tasks with high precision.

However, despite their perceptual precision, traditional CV models are fundamentally limited by their inability to reason and execute autonomous control. They typically operate as closed-set systems, mapping pixel inputs to fixed output classes without understanding the underlying semantics or physical causality. Consequently, they struggle to interpret complex, high-level instructions or adapt to unseen manufacturing scenarios that require log-

Table 6: List of evaluated open-source and closed-source MLLMs.

Open-Source / Weights Models		Closed-Source Models	
Provider	**Model**	**Provider**	**Model**
Google	Gemma-3-27B	OpenAI	GPT-5 / 5.2
OpenGVLab	InternVL3-78B	OpenAI	GPT-5-Mini
Meta	Llama-4-Maverick	OpenAI	O3
Mistral	Mi(ni)stral-3-8B/14B/Large	Google	Gemini-2.5/3-Flash
Alibaba	Qwen3-VL-8B/235B	Anthropic	Claude-4.5-Opus
Zhipu AI	GLM-4.6V	ByteDance	Seed-1.6
Moonshot	Kimi-K2.5		

ical deduction rather than mere pattern matching. This limitation underscores the urgent need to introduce MLLMs equipped with cognitive reasoning capabilities.

A.1.2 MLLMs in Manufacturing Scenario.

Recent advances in manufacturing anomaly detection leverage MLLMs to shift from simple perception to complex reasoning. LogiCode(Zhang et al., 2024) prompts LLMs to generate executable Python code for verifying logical constraints, though it relies on manual annotations. LogicQA(Kwon et al., 2025) takes a more scalable, annotation-free approach, using VLMs to generate question-based checklists from standard samples for zero- or few-shot detection. EMIT(Guan et al., 2025) proposes Difficulty-Aware GRPO to align MLLMs with challenging anomaly detection tasks via response resampling and advantage reweighting. Despite their innovations, these works operate within existing datasets and predefined tasks, failing to explore novel formulations that reflect real-world factory complexity.

A.2 Extended Main Results

In this section, we present more detailed results compared to the main text.

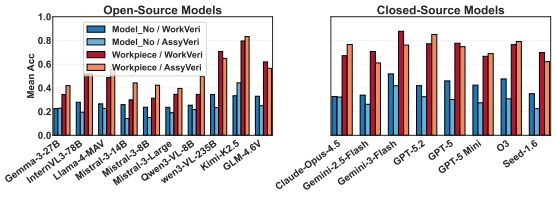


Figure 5: Performance of Open-Source vs. Closed-Source models.

Model-number-level tasks are more challenging for MLLMs compared to workpiece-level tasks. A comprehensive analysis of WORKVERI and ASSYVERI based on Figure 5 reveals a distinct performance disparity: MLLMs consistently outperform on workpiece-level tasks compared to model-number-level tasks. Whether for open-source or closed-source models, performance on model-number-level tasks (blue) is worse than on workpiece-level tasks (red). This substantial gap indicates that while MLLMs have established a certain level of understanding of general manufactured workpieces, there remains significant room for improvement in capturing fine-grained domain specificity. Nevertheless, the capability to conduct fine-grained analysis is crucial in manufacturing scenarios, as manufacturing systems and tasks are discussed in Section 2.

Specifically, Table 7 details the performance of WORKVERI and ASSYVERI across different workpieces and model numbers, while Table 8 breaks down the performance across various anomaly cases. Overall, these findings are consistent with the results discussed in the main text: **Current MLLMs face substantial hurdles, where overall performance remains far from acceptable.** Results across the three tasks reveal that the insufficient capability to internalize and reason about complex manufacturing standards makes this domain an arduous challenge for future MLLM development. Notably, most open-source models perform near the random baseline. While closed-source models such as Gemini-3-Flash and GPT-5 demonstrate superior capabilities, achieving state-of-the-art results in WORKVERI and ASSYVERI, even the most advanced models struggle to surpass 50% accuracy in the highly demanding SURFINSP. This disparity underscores that, although closed-source models exhibit stronger general-purpose capabilities, performance in highly specialized manufacturing scenarios remains a significant challenge.

However, analyzing the results for WORKVERI and ASSYVERI across different workpieces and model numbers in Table 7, we observe a distinct pattern across the three settings. Specifically, for the three-view (3V) modality, both model-number-level and workpiece-level tasks expose a significant bottleneck in 3D spatial understanding. Interestingly, under the 3V setting, both Ref-Cond and ICD exhibit a performance degradation compared to the zero-shot baseline. However, this decrease is significantly more pronounced in workpiece-level tasks than in model-number-level tasks. This observation further corroborates Conclusion C in Section 4.3. The root cause lies in the fact that the similarity among workpieces is substantially lower than that among different model numbers. Conse-

Task	Case	Mod.	Setting	Open-source Models										Closed-source Models							
				Gemma-3-27B	InternVL3-78B	Llama-4-MAV	Mistral-3-14B	Mistral-3-8B	Mistral-3-Large	Qwen3-VL-235B	Qwen3-VL-8B	Kimi-K2.5	GLM-4.6V	Claude-4.5-Opus	Gemini-2.5-Flash	Gemini-3-Flash	GPT-5.2	GPT-5	GPT-5 Mini	O3	Seed-1.6
WorkVeri	Model No	Image	Zero-Shot	23.18	27.04	24.55	25.78	20.09	24.46	36.91	21.12	30.41	**41.84**	30.04	33.48	46.78	48.03	51.93	53.88	**55.70**	38.96
			Ref-Cond	25.97	30.47	21.33	18.61	23.81	19.40	33.64	25.54	**37.04**	27.90	34.33	30.04	**55.79**	32.86	40.52	49.36	45.66	27.07
			ICD	18.03	**41.56**	31.28	24.24	26.84	24.03	35.34	22.41	38.76	29.00	30.90	31.33	65.67	60.87	**71.43**	61.90	66.51	44.21
		3V	Zero-Shot	22.87	20.18	29.15	27.35	26.24	27.91	**37.79**	33.33	35.00	37.50	30.49	39.01	**51.12**	43.50	40.09	31.39	44.09	40.81
			Ref-Cond	20.18	21.17	19.28	**30.94**	19.73	20.36	30.05	25.69	22.64	28.96	29.60	31.39	**49.10**	22.07	25.91	27.35	26.82	22.87
			ICD	25.11	27.03	33.63	28.38	25.45	25.56	32.49	24.22	**36.30**	32.43	41.26	38.12	**49.33**	43.95	45.50	30.04	46.51	36.65
	Workpiece	Image	Zero-Shot	28.90	38.53	54.07	41.90	41.98	27.19	93.12	50.69	**100.00**	57.03	90.83	79.72	**99.54**	99.05	99.08	94.95	97.70	96.79
			Ref-Cond	35.32	78.90	61.54	21.96	24.77	32.09	85.13	22.43	**95.59**	74.31	79.72	74.77	**98.17**	78.74	89.45	97.71	92.65	73.39
			ICD	52.29	90.37	75.36	30.15	37.61	38.89	99.54	28.37	**100.00**	72.35	93.12	79.36	**100.00**	100.00	100.00	93.09	100.00	98.17
		3V	Zero-Shot	31.50	42.49	42.28	36.63	34.80	40.30	**64.07**	49.07	60.71	55.39	46.32	69.96	**84.62**	75.82	78.23	43.38	69.60	65.57
			Ref-Cond	26.37	27.94	21.25	20.15	17.04	21.40	40.16	28.41	45.78	**56.88**	26.47	53.48	**78.39**	33.46	33.33	39.34	31.82	30.15
			ICD	31.87	32.72	37.36	28.68	31.62	47.41	42.06	28.04	**75.11**	55.22	66.30	66.79	**82.05**	76.19	65.81	31.25	66.91	54.41
AssyVeri	Model No	Image	Zero-Shot	19.27	18.94	20.82	17.86	16.76	21.39	15.29	17.52	**27.12**	21.34	33.70	25.87	**40.44**	27.94	29.87	21.17	26.52	19.05
			Ref-Cond	23.50	24.72	21.71	25.14	16.39	16.27	24.71	18.83	**43.56**	24.73	41.71	21.02	**61.86**	38.76	36.13	37.11	46.69	27.74
			ICD	23.72	22.26	30.05	19.17	20.33	19.12	31.55	22.53	**39.28**	35.15	47.63	32.23	**64.30**	48.90	44.89	39.85	46.44	27.88
		3V	Zero-Shot	16.78	26.17	19.46	15.44	14.77	19.63	23.57	19.46	**100.00**	22.07	13.42	20.13	**32.89**	22.97	25.68	21.48	20.00	23.13
			Ref-Cond	26.85	6.04	21.48	0.00	0.67	19.83	21.05	25.52	**42.86**	25.17	26.17	26.17	**30.87**	28.86	18.12	21.62	24.26	14.29
			ICD	**26.85**	18.79	22.82	7.43	20.81	18.25	24.06	26.35	12.77	21.48	30.87	**31.54**	30.87	27.89	26.35	23.49	20.69	22.82
	Workpiece	Image	Zero-Shot	44.92	58.44	65.28	52.70	52.60	42.86	76.08	57.70	**91.84**	58.37	85.02	63.96	89.61	84.67	86.04	83.77	**89.90**	76.87
			Ref-Cond	46.41	57.00	49.80	44.59	45.45	37.01	69.79	46.03	**100.00**	59.61	82.47	52.44	86.27	**89.93**	74.03	76.97	84.04	69.08
			ICD	49.35	78.50	73.98	47.52	49.01	45.60	84.48	44.26	**91.89**	63.52	90.23	72.96	84.36	91.09	90.07	74.51	**92.12**	84.92
		3V	Zero-Shot	36.88	41.88	55.62	40.00	33.75	40.27	58.75	57.50	**77.78**	55.70	68.75	68.12	60.62	**83.12**	80.00	75.62	79.87	58.75
			Ref-Cond	36.88	36.88	38.12	36.88	35.00	36.88	39.61	41.40	**60.71**	48.72	73.75	53.12	60.62	**83.12**	42.50	36.88	47.37	38.12
			ICD	36.88	38.12	51.25	43.40	37.74	34.59	60.40	50.62	**77.50**	53.12	59.12	55.62	63.12	78.12	75.62	65.62	**81.25**	45.00

Table 7: Performance of different cases for WorkVeri and AssyVeri, different workpieces, and different model numbers. **Bold orange** / underlined blue / light purple indicate best / second / third within each model category (open- or closed-source) per row.

Case	Method	Open-Source										Closed-Source							
		Gemma-3-27B	InternVL3-78B	Llama-4-MAV	Mistral-3-14B	Mistral-3-8B	Mistral-3-Large	Qwen3-VL-235B	Qwen3-VL-8B	Kimi-K2.5	GLM-4.6V	Claude-4.5-Opus	Gemini-2.5-Flash	Gemini-3-Flash	GPT-5.2	GPT-5	GPT-5 Mini	O3	Seed-1.6
CRACK	Zero-Shot	26.21	6.21	4.23	**44.14**	37.93	3.45	7.59	30.34	26.57	42.76	21.38	0.69	10.34	6.21	2.07	6.90	4.14	**41.38**
	Ref-Cond	14.48	4.83	9.93	57.24	**58.62**	9.66	5.52	29.86	32.64	44.06	35.17	1.38	40.00	4.14	4.14	6.21	6.25	**48.28**
	ICD	12.41	15.17	2.78	62.76	**66.90**	15.86	3.47	22.76	21.80	11.03	4.14	62.76	**66.90**	19.31	24.14	13.79	28.28	38.62
CUT	Zero-Shot	0.61	1.22	13.21	7.32	1.83	0.00	**18.90**	6.10	16.67	0.61	18.29	1.83	6.71	**34.76**	21.95	25.00	10.98	4.88
	Ref-Cond	1.22	0.00	9.03	8.54	0.00	1.22	**20.12**	11.59	18.87	1.83	18.29	4.27	25.61	31.71	28.05	**44.51**	15.24	5.49
	ICD	3.66	1.83	10.37	23.78	14.63	20.73	4.29	10.98	**28.29**	7.93	13.41	17.68	47.56	37.20	37.20	**59.15**	32.93	19.51
DEFORMATION	Zero-Shot	84.56	95.30	51.75	77.85	81.88	**97.99**	78.52	79.19	49.32	73.83	72.48	**98.66**	85.91	63.09	66.44	71.81	77.18	70.47
	Ref-Cond	65.77	**95.30**	64.29	75.17	64.43	74.50	79.19	56.38	62.16	83.11	57.05	**94.63**	67.11	59.73	40.94	54.36	59.73	76.51
	ICD	44.97	**87.92**	18.12	30.20	34.23	39.60	49.32	46.98	27.21	43.24	2.01	**57.05**	26.17	46.31	25.50	20.81	32.89	53.69
DENT	Zero-Shot	6.34	7.04	3.70	0.70	0.00	**8.45**	4.23	7.04	7.97	2.11	5.63	0.70	1.41	5.63	**16.20**	7.75	7.75	0.00
	Ref-Cond	23.94	5.63	6.62	2.11	9.15	**28.87**	2.82	19.72	8.57	1.41	15.49	4.23	7.04	10.56	18.31	15.49	**23.24**	2.11
	ICD	38.03	8.45	9.15	12.68	7.75	**42.25**	10.56	28.87	11.38	7.75	14.79	28.17	**39.44**	12.68	28.87	7.04	26.06	14.08
GOOD	Zero-Shot	0.00	0.00	**44.89**	0.00	3.04	0.00	1.30	0.43	3.12	0.43	2.61	0.00	0.00	0.87	**16.52**	0.00	12.17	2.17
	Ref-Cond	34.35	0.00	**48.65**	0.00	6.96	0.00	4.78	7.86	0.00	10.04	6.96	34.35	23.04	10.43	**68.70**	48.70	63.48	43.04
	ICD	4.78	11.30	53.04	46.09	**59.57**	18.70	38.16	26.96	8.25	31.30	63.48	53.48	61.74	35.22	**73.36**	62.61	68.26	58.70

Table 8: Per-defect-class accuracy for SurfInsp on 830 cases across 14 workpieces. Each cell is the model's accuracy on test cases whose ground-truth class is the row label (CRACK, CUT, DEFORMATION, DENT, or GOOD). **Bold orange** / underlined blue / light purple indicate best / second / third within each model category (open- or closed-source) per row.

quently, the visual features of workpieces exhibit much higher variance under the inherent spatial and angular shifts of 3V images. This extreme variability severely hinders MLLMs from aligning spatial semantics when integrating reference examples, ultimately leading to a steeper performance drop for Ref-Cond and ICD in workpiece-level tasks.

More granular results for SURFINSP are presented in Table 8. Although the task comprehension requirement for SURFINSP is relatively low, and performance generally improves as the number of references increases across all three settings, the anomaly of DEFORMATION does not conform to this trend. This suggests that even for seemingly simple tasks, varying domain knowledge requirements cause MLLMs to interpret the task differently, leading to distinct performance trajectories across different anomaly types.

A.2.1 Risk-Sensitive Metrics for SURFINSP.

Exact-match accuracy weights all mistakes equally, whereas in industrial inspection the two error directions carry very different costs: a defective workpiece passed as good escapes downstream, while a good workpiece flagged as defective incurs scrap or rework. SURFINSP is the task where this asymmetry is directly defined, because the model makes an accept/reject decision; in WORKVERI and ASSYVERI the model instead selects which candidate is anomalous, so its errors are misidentifications rather than accept/reject decisions.

From the stored per-case predictions of the same 830-case SURFINSP evaluation reported in Table 3 (600 defective, 230 defect-free), we recompute macro-F1 together with the *miss rate* (a defective workpiece predicted *good*) and the *false-alarm rate* (a defect-free workpiece predicted defective) for all 18 models under all three settings. Responses from which no answer can be parsed are counted as errors, pessimistically, in whichever direction the ground truth lies. Twelve of the 54 runs have slightly fewer logged cases than the full set (minimum 750); all rates are computed over the cases logged for each run, and three accuracy cells differ from Table 3 by at most 0.6 points because the stored predictions are re-parsed here. Note that the false-alarm rate is the complement of the GOOD row of Table 8, which reports the same predictions from a per-class perspective.

Table 9 reports the result. Two observations do not follow from accuracy alone. First, in the Zero-Shot setting 15 of the 18 models flag at least 96.8% of defect-free workpieces as defective; their apparently safe near-zero miss rates are an artefact of answering "defective" almost always rather than evidence of sensitivity. Second, models with nearly identical accuracy can fail in opposite, cost-relevant directions: under ICD, Claude-4.5-Opus misses 60.5% of true defects while Qwen3-VL-235B instead over-flags 61.8% of defect-free workpieces, although their accuracies (23.9% and 22.7%) would suggest they are interchangeable. Even the best-calibrated model retains a substantial false-alarm burden (Gemini-3-Flash: 7.5% miss, 38.3% false alarms), which quantifies the distance to quality-critical deployment more sharply than accuracy does. Macro-F1 preserves the ranking reported in the main text, so the conclusions there are unaffected; these metrics add risk-sensitive information rather than revise any result.

A.2.2 Rationale for the MCQ Formulation.

All FORGE tasks are posed as multiple-choice questions. We adopt this formulation to make evaluation judge-free: exact-match scoring of a single letter is deterministic, reproducible, and directly comparable across model families, and it is the protocol used by manufacturing-domain MLLM benchmarks such as MMAD (Jiang et al., 2024). The alternative, free-form scoring, would require either an LLM judge, which introduces its own bias and instability into a benchmark whose purpose is measurement, or manual grading of roughly 13k samples across 18 models and 3 settings, which is not feasible at this scale.

The formulation is nonetheless a simplification of deployed inspection, where a system must detect, localize, and describe faults without a candidate list. Two points bound how much this simplification flatters the models. First, enumerating candidates makes the task *easier* than open-set detection, so the failures we report are a lower bound on the difficulty of the open-ended setting: even with candidates supplied, the best model reaches only 49.6% on SURFINSP against a 20% random baseline. Second, the option sets are constructed to carry no signal about the answer: in WORKVERI and ASSYVERI every workpiece in the assembly appears as one option, identified only by its normalized centre coordinate, so no option is textually distinguishable from the others, and in SURFINSP the same five class labels are offered for every case.

The controlled analyses in Section 4.4 additionally recover capabilities that the multiple-choice

Table 9: **Risk-sensitive metrics for SURFINSP** (full set: 830 cases, 600 defective / 230 defect-free). **Acc** = exact-match accuracy, **F1** = macro-F1 over the five classes, **Miss** = percentage of defective workpieces predicted *good*, **FA** = percentage of defect-free workpieces flagged as defective. Unparseable responses count as errors. Upper block: open-source models; lower block: closed-source models.

Model	Zero-Shot				Ref-Cond				ICD			
	Acc	F1	Miss	FA	Acc	F1	Miss	FA	Acc	F1	Miss	FA
Gemma-3-27B	21.0	14.1	0.0	100.0	28.2	24.6	3.7	65.7	18.8	16.0	3.7	95.2
InternVL3-78B	19.6	11.1	0.0	100.0	18.9	9.7	0.0	100.0	23.4	17.7	0.5	88.7
Llama-4-MAV	25.8	19.6	31.3	55.1	29.6	24.8	15.7	51.4	22.1	16.9	52.9	47.0
Mistral-3-14B	23.3	16.8	0.0	100.0	25.5	19.6	0.0	100.0	36.0	33.5	11.7	53.9
Mistral-3-8B	22.5	15.2	0.7	97.0	25.3	19.8	0.0	93.0	38.6	33.3	14.0	40.4
Mistral-3-Large	19.6	10.3	0.0	100.0	20.2	14.6	0.0	100.0	26.4	26.0	2.5	81.3
Qwen3-VL-235B	20.2	14.3	0.2	98.7	21.0	14.9	0.2	95.2	22.7	17.5	35.2	61.8
Qwen3-VL-8B	22.1	16.7	0.0	99.6	23.2	21.4	0.5	92.1	27.0	26.5	1.8	73.0
Kimi-K2.5	19.1	17.1	6.8	96.9	22.2	19.3	0.3	100.0	18.7	18.4	13.1	91.8
GLM-4.6V	21.3	13.8	0.5	99.6	25.9	18.8	0.2	90.0	21.2	18.2	34.1	68.7
Claude-4.5-Opus	22.1	18.3	0.7	97.4	24.6	23.3	0.5	93.0	23.9	16.1	60.5	36.5
Gemini-2.5-Flash	18.3	7.4	0.0	100.0	28.3	20.0	2.8	65.7	44.3	43.1	4.8	46.5
Gemini-3-Flash	18.8	11.5	0.0	100.0	31.7	29.6	1.5	77.0	49.6	48.2	7.5	38.3
GPT-5.2	20.5	15.4	0.3	99.1	22.4	18.9	0.2	89.6	31.0	30.1	2.0	64.8
GPT-5	24.0	20.3	7.5	83.5	35.8	29.7	18.2	31.3	41.4	37.9	17.5	26.6
GPT-5 Mini	20.4	15.2	0.2	100.0	35.8	31.9	6.7	51.3	36.4	30.6	13.7	37.4
O3	21.5	16.5	5.5	87.8	36.4	30.8	13.7	36.5	40.7	38.0	14.3	31.7
Seed-1.6	21.5	15.3	1.2	97.8	35.5	30.0	2.2	57.0	38.9	36.5	5.5	41.3

format itself does not test: the grounding probes require the model to *predict* coordinates rather than select among them, and the missing-part probe removes coordinates entirely and adds a "no workpiece is missing" option. Full prompts and configurations for these probes are given in Section C.2.1. A fully open-ended track, in which models must localize and describe faults directly, is a natural extension that our released annotations, which include part contours, coordinates, model numbers, and defect types, are sufficient to support.

A.3 Task Description

In FORGE, we primarily designed three tasks related to manufacturing scenarios to evaluate the key capabilities of MLLMs in typical manufacturing contexts such as **material sorting**, **quality inspection**, and **assembly recognition**. The mapping relationship between the real factory scenario and FORGE tasks is shown in Figure 6. This Section systematically introduces the detailed design framework of these three tasks, including task design, construction of testing and evaluation data, evaluation settings, and the corresponding data collection process.

A.3.1 Raw Data Collection.

In this work, to facilitate the evaluation of MLLMs and provide data for our tasks, we collected raw data from authentic manufacturing workpieces. The data acquisition setup was established using a platform equipped with a precision rotary table and a custom-designed fixture. The data collection process was divided into two distinct phases: point cloud acquisition and image data collection.

Availability. All FORGE assets are released under the MIT License: (i) the complete dataset, comprising raw images, raw point clouds, three-view renderings, and all annotations, at https://huggingface.co/datasets/AI4Manufacturing/forge; and (ii) the full evaluation harness, including the per-task evaluation scripts, prompts, and the exact configuration used for every reported number, together with the fine-tuning scripts and the scripts used for the 3D–language baselines and the risk-sensitive metrics, at https://github.com/AI4Manufacturing/FORGE.

Point Cloud Data Acquisition: A handheld 3D scanner was utilized for data collection with a scanning precision configured to 0.02 mm. Before scanning, positioning markers were affixed to the platform to facilitate the establishment of a local coordinate system via marker recognition. To mitigate the surface reflectivity inherent in metallic workpieces, a 3D scanning developer spray was applied as a pretreatment. Once the workpiece was securely positioned, the scanning procedure was initiated. Subsequently, proprietary software was employed to perform post-processing operations, including registration, denoising, mesh reconstruction, and texture mapping, yielding high-fidelity 3D model reconstructions.

Image Data Acquisition: Specific experimental scenarios were pre-designed for image data collection. Metal workpieces of corresponding models

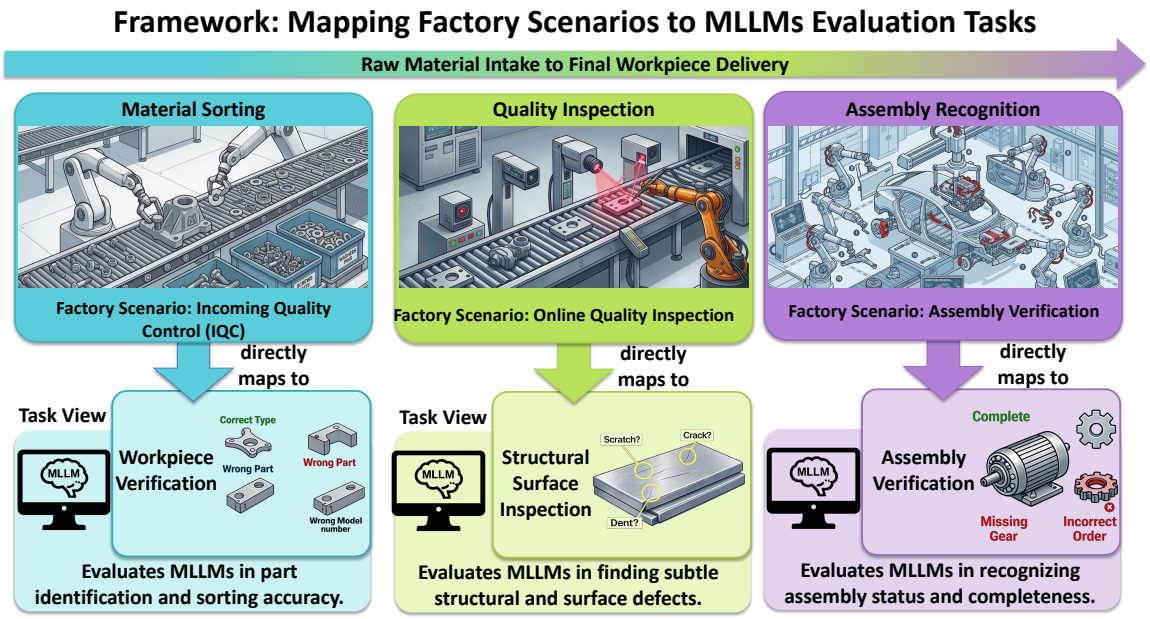


Figure 6: Mapping Factory Scenarios to MLLMs Evaluation Tasks(some images are AI-generated).)

were randomly selected from the sample library and placed in random orientations, while maintaining consistency in shooting angles and distances. Images were captured using a 50-megapixel sensor with a Leica Vario-Summilux optical lens system to ensure high resolution.

Then, the raw dataset is categorized into **Point Cloud Subset** and **Image Subset**, capturing both geometric and visual aspects of manufacturing components.

Component Selection and Representativeness: The workpieces were deliberately chosen from *standardized* industrial components rather than sampled arbitrarily. The 14 categories are standard fastener and fixture families, spanning screws (cup-head, countersunk, self-tapping, T, T-bolt half-thread, wing), nuts (hex, rivet, T, wing), washers (flat, spring), eye bolts, and corner brackets, and their 90 model numbers follow internationally recognized ISO/DIN/GB designations. Standardized fasteners are among the most widely used components in discrete manufacturing, and it is precisely their standardization that makes model-number-level verification a well-posed industrial task. The image scenarios were selected on the same principle, covering metal and plastic expansion-screw assemblies, CNC fixtures, and pneumatic connector assemblies, which are routine in machining and assembly workshops. We do not claim to cover the full space of manufactured components; extending FORGE to further families, such as bearings, gears, or sheet-metal parts, is a natural direction that the released collection and annotation pipeline supports.

Point Cloud Subset: This subset covers 14 component categories and 90 model numbers. Each sample is represented as a 3D point cloud, providing high-fidelity geometric structure information. Through data preprocessing and manual annotation, this subset enables a variety of applications, including WORKVERI, SURFINSP, and ASSYVERI. The detailed collected 3D point cloud data are summarized in Table 11.

Image Subset: This subset comprises approximately 3000 images collected from four manufacturing scenarios, including expansion screw assemblies and positioners. Each scenario contains normal samples and abnormal samples. This subset primarily supports WORKVERI and ASSYVERI under real-world manufacturing conditions. The detailed collected image data are summarized in Table 12.

While these two subsets provide the comprehensive raw material for WORKVERI, SURFINSP, and ASSYVERI, raw data requires specific curation before it can be used for standardized evaluation. Therefore, the detailed data preparation pipelines and specific definitions for each task are presented in the following sections.

Data Processing and Annotation: It is important to note that the raw data, comprising both 2D images and 3D point clouds, could not be directly

utilized for FORGE. Therefore, data preprocessing pipelines were designed for each modality.

For 2D image data, to ensure exceptional label accuracy while maintaining efficiency, we adopted a hybrid strategy combining algorithmic pre-extraction with a two-stage manual fine-annotation. Initially, a Python-based algorithm batch-processed raw 2D images to automatically extract edge contours of the workpiece and generate preliminary spatial coordinates, significantly reducing manual labor. During the first manual annotation stage, annotators conducted a frame-by-frame inspection to correct algorithmic deviations and map the 2D spatial coordinates to the manufacturing workpiece in 2D image data, ensuring highly accurate absolute coordinates. To establish a reliable benchmark test set, experienced personnel conducted a second manual review to identify "mismatch workpieces" that failed production standards. The specific coordinates of these non-compliant workpieces were extracted, yielding highly reliable ground-truth labels for subsequent evaluations.

For WORKVERIand ASSYVERI: Since the collected data consisted of single normal workpieces, we utilized CloudCompare to synthesize batch samples. Specifically, we stitched 4-5 individual point clouds together, applying random orientations and relative positions within a constrained range. Corresponding labels were automatically generated based on the artifact models and workpiece information during this assembly process. For SURFINSP: We focused on synthetic defect generation since real defect samples from production lines are difficult to collect, especially in the fine-grained setting; synthetic generation following industrial defect standards ensures reproducibility and diversity while preserving physical realism. Four typical manufacturing defects were simulated: Crack, Deformation, Dent, and Cut. For each type, we designed a tailored algorithm based on its distinctive morphological characteristics to produce a large set of initial shapes. Non-rigid deformation was then applied to enhance realism and variability. The proportion of defect points per sample was constrained between 5% and 15%. To address data scarcity, we augmented the dataset by applying 20 random rotations (uniformly sampled from $[0°, 360°)$) to each sample. Furthermore, our preliminary tests revealed that feeding raw point cloud files directly into the LLM resulted in sub-optimal performance. Consequently, we adopted a multi-view projection strategy: all point clouds were rendered into 3V images (orthogonal projections) to serve as the actual input for MLLMs.

Data Provenance. Because FORGE combines captured and derived material, Table 10 states, for each component of the benchmark, what was physically measured and what was produced computationally. Every photograph and every raw point cloud is an original capture of a physical workpiece performed by us; no third-party dataset is used. The WORKVERI and ASSYVERI point-cloud assemblies are *compositions of real scans*, i.e. the geometry of every workpiece in an assembly is measured, and only the relative placement is randomized. The only synthesized content in FORGE is the defect region of SURFINSP samples, and the three-view images are a deterministic rendering of the underlying clouds rather than generated content.

On the use of synthesized defects. We generate SURFINSP defects rather than collecting them because real production-line defects are scarce in precisely the fine-grained, per-model-number setting this benchmark requires, and because synthesis provides exact ground-truth defect regions, reproducibility, and balanced coverage of all four defect types across all 14 workpiece categories, none of which an opportunistically collected defect set can guarantee. Each generator was designed from the morphological characteristics of its defect class and applied to real 0.02 mm-precision geometry, with non-rigid deformation added for variability. We do not claim that the resulting defects match the distribution of any particular production line; establishing that correspondence, and complementing FORGE with real production defects, is important future work. Visual examples of the generated defects on real scans are provided in Figures 12–25.

Eval Setting: For all tasks, the evaluation settings are divided into three categories: **Zero-Shot**, **Reference-Conditioned(Ref-Cond)**, and **In-Context Demonstration(ICD)**.In the **Zero-Shot** setting, only the test image and the corresponding textual query are provided as input during evaluation. In the **Ref-Cond** setting, three correct normal cases are additionally included as reference examples for the MLLMs model. In the **In-Context Demonstration** setting, one more example that is similar to the test case, consisting of an image, its corresponding query, and the correct answer, is added on top of the Ref-Cond examples. The final test query is then provided based on these contextual demonstrations. To further systematize

Table 10: **Provenance of each component of FORGE.** All photographs and all raw point clouds are original captures of physical workpieces; the only synthesized content is the defect region of SURFINSP samples.

Component	Origin	How it was produced
Image subset (MES SCENARIO, PES SCENARIO, CNC SCENARIO, PCS SCENARIO)	Real capture	Photographs of physical workpieces, 50 MP sensor with Leica optics; contours and labels from algorithmic pre-extraction plus two-stage manual annotation.
Raw point clouds (14 categories, 90 model numbers)	Real capture	Handheld 3D scans of physical workpieces at 0.02 mm precision, followed by registration, denoising, and mesh reconstruction.
WORKVERI / ASSYVERI point-cloud assemblies	Composition of real scans	4–5 real per-workpiece scans stitched with randomized orientation and position; labels generated automatically during composition.
SURFINSP defects	**Synthesized** on real scans	Per-defect-type morphological generators plus non-rigid deformation applied to real scans; defect region constrained to 5–15% of points; 20 random rotations per sample for augmentation.
Three-view renderings	Deterministic rendering	Orthographic front/side/top projection of the above point clouds; an input-interface choice, not generated content.

Table 11: Simple data summary of raw point cloud data. The dataset contains 14 workpieces.

Workpieces	Model Number	No. of Samples
Corner Bracket	5	25
Countersunk Screw	2	20
Cup Head Screw	22	110
Eye Bolt	5	45
Flat Washer	9	45
Hex Nut	4	33
Rivet Nut	3	30
Self-tapping Screw	6	60
Spring Washer	9	50
T Bolt Half thread Screw	4	40
T Nut	4	20
T Screw	10	50
Wing Nut	3	27
Wing Screw	4	30
Total	**90**	**585**

the evaluation setting, we categorize potential error scenarios into two primary classes: *Different workpiece* and *Different Model Number*. The former refers to coarse-grained discrepancies, including workpiece mismatches, component absence, or other workpiece-level anomalies. The latter addresses fine-grained inconsistencies, specifically focusing on errors arising from distinct model variants despite the workpiece category being correct.

A.3.2 Task Description of WORKVERI.

In manufacturing scenarios, material sorting refers to the process of identifying, selecting, classifying, assembling, matching, and delivering materials from inventory to designated locations in accordance with specific requirements. Among these steps, material identification and verification are the most fundamental and essential. To evaluate the capability of MLLMs in material identification and verification within manufacturing environments, we design a task termed WORKVERI. Given explicit workpiece specifications or model requirements, the task requires MLLMs to analyze 3D point clouds or image data of manufacturing workpieces and identify those that do not satisfy the specified requirements. **Task design and details:** Based on this task, we construct three representative manufacturing application scenarios. Two of them are based on point cloud data, with the research objects being common manufacturing components, namely Nuts and Cup Head Screws. The third scenario is based on image data and focuses on commonly used Pneumatic Connectors (PCs). The illustration of PCS SCENARIO is presented in Figure 9, and the illustration of 3D point cloud scenario (CHS SCENARIO and NUTS SCENARIO) is presented in Figure 10 and Figure 11.

A.3.3 Task Description of SURFINSP.

Quality inspection has long been a fundamental research problem in manufacturing. Its core objective is to systematically inspect and measure workpieces, components, or production processes to ensure compliance with design specifications, process standards, and customer requirements. In this paper, we propose a SURFINSP task to evaluate the capability of MLLMs in identifying manufacturing defects from workpiece point cloud data in manufacturing scenarios. **Task design and de-**

Table 12: Simple data summary of raw image data (scenario-level).

Workpiece	Wrong types	No. of samples
MES SCENARIO	No Spring Washers	56
	Hex Nut M14	52
	Flat Washer M14	51
	No Flat Washers	50
	Two Flat Washers	50
	Two Spring Washers	49
	Spring Washer M8	48
	Cup Head Screw M12 40	46
PES SCENARIO	One missing plastic expansion anchor	52
	One missing plastic expansion Self-tapping Screw	51
	One extra plastic expansion Self-tapping Screw	49
	Screws belonging to model 860 expansion anchors were found mixed in	49
	Screws shorter 30 than the specified length were found mixed in	49
	One extra plastic expansion anchor	47
	Model 860 expansion screws were mistakenly included	46
	Plastic anchors belonging to the expansion screw model 860 were mixed	44
CNC SCENARIO	Missing one nut	57
	Extra one nut	56
	Missing triangular part	56
	retainer block	56
	Missing screw	53
	Long screw	51
	Short screw	50
	Triangular part	48
PCs SCENARIO	2-way pneumatic tube connectors (8 to 6), mixed with model (10 to 6)	67
	Three 2-way pneumatic tube connectors (8mm to 6mm), mixed with T-type 3-way	63
	2-way pneumatic tube connectors (8 to 6), mixed with model (6 to 4)	57
	Three 2-way pneumatic tube connectors (8mm to 6mm), mixed with Y-type 3-way	57
	2-way pneumatic tube connectors (8 to 6), mixed with model (12 to 8)	56
	2-way pneumatic tube connectors (8 to 6), mixed with model (8 to 4)	56
	Three 2-way pneumatic tube connectors (8mm to 6mm), mixed with throttle valve	55
	Three 2-way pneumatic tube connectors (8 to 6), mixed with external-thread elbow	50
Total		**3115**

tails: Specifically, the task involves: (1) determining whether a workpiece contains defects; and (2) further identifying the type of defect. The experimental evaluation covers 14 categories of manufacturing components, and the considered defect types are Crack, Cut, Deformation, and Dent. The point cloud data of different mechanical components are illustrated in Figures 12–25.

A.3.4 Task Description of ASSYVERI.

Assembly recognition is the automatic identification and understanding of assembly relationships, structural hierarchies, and compatibility constraints among workpieces or materials. Compared with material sorting, this task imposes greater requirements on MLLMs, as it requires them to reason about more complex assembly rules and compatibility relationships. To this end, we design an evaluation task termed ASSYVERI, which aims to systematically assess MLLMs' capability in understanding assembly relationships. Given specific compatibility rules, workpiece specifications, or model requirements, MLLMs analyze point clouds or images of manufacturing workpieces and identify those that do not meet assembly requirements.

Task design and details: This task consists of four representative manufacturing scenarios. Among them, three scenarios are based on image data and involve common manufacturing applications, including metal expansion screws(MES SCENARIO), plastic expansion screws (PES SCENARIO), and CNC fixtures (CNC SCENARIO). The remaining scenario is based on point cloud data and focuses on the compatibility relationships among metal screws, washers, and nuts. The illustration of image and point cloud data is in Figure 26–29.

Task	Scenario	Case Type	Error Type	Samples
WorkVeri	CHS scenario	Model No	M16 longer length 100	42
			M18 longer length 100	35
			M12 shorter length 50	34
			M10 shorter length 50	32
		Workpiece	Self tapping Screw	37
			T Bolt Half thread Screw	44
			T Screw 45 M8	36
			Wing Screw	41
	Nuts scenario	Model No	Mixed with M16	75
			Mixed with M20	75
		Workpiece	Rivet Nut	78
			T Nut	81
			Wing Nut	81
	PCs scenario	Model No	Mixed with (10 to 6)	67
			Mixed with (12 to 8)	59
			Mixed with (6 to 4)	59
		Workpiece	Mixed with (8 to 4)	56
			Mixed with external-thread elbow	50
			Mixed with T-type 3-way	63
			Mixed with throttle valve	55
			Mixed with Y-type 3-way	58
SurfInsp			Corner Bracket	48
			Countersunk Screw	40
			Cup Head Screw	234
			Eye Bolt	86
			Flat Washer	62
			Hex Nut	58
			Rivet Nut	58
			Self tapping Screw	116
			Spring Washer	88
			T Bolt Half thread Screw	76
			T Nut	62
			T Screw	160
			Wing Nut	54
			Wing Screw	58
AssyVeri	MES scenario	Model No	Cup Head Screw M12 40	51
			Flat Washer M14	52
			Hex Nut M14	52
			Spring Washer M8	51
		Workpiece	Extra Flat Washers	55
			Extra Spring Washers	52
	SWN scenario	Model No	Cup Head Screw M18 100	52
			Flat Washer M14	15
			Hex Nut M16	37
			Spring Washer M20	46
		Workpiece	Extra Cup Head Screw	32
			Extra Flat Washers	44
			Extra Hex Nut	38
			Extra Spring Washers	46
	PES scenario	Model No	Plastic anchors belonging to expansion screw model 860 were mixed	51
			Screws belonging to model 860 expansion anchors were found mixed in	56
			Shorter screws were found mixed in	50
		Workpiece	Model 860 expansion screws were mistakenly included	51
			One extra plastic expansion anchor	52
			One extra plastic expansion Self-tapping Screw	54
	CNC scenario	Model No	Long screw	51
			retainer block	56
			Short screw	53
			Triangular part	48
		Workpiece	Extra one nut	56
				3559

Table 13: Simple data summary of three tasks (scenario-level).

In summary, by designing three core tasks,WORKVERI, SURFINSP, and ASSYVERI, combined with high-quality multimodal data (point clouds and images), we have constructed a comprehensive and challenging benchmark corresponding to typical manufacturing contexts such as **material sorting**, **quality inspection**, and **assembly recognition**. This design not only addresses the fundamental perceptual requirements of manufacturing scenarios but also aims to rigorously evaluate models' logical judgment and reasoning capabilities under complex manufacturing rules. Building upon this framework, the following sections will systematically evaluate and discuss the reasoning mechanisms and performance of MLLMs in addressing real-world manufacturing challenges. A brief visualization results are presented in Figure 7 and full visualization results are presented in Section D, and complete dialogue examples are provided in Section E.

B Training Details and Extended Results

Training Configurations. We train `Qwen2.5-VL-3B-Instruct` using *full-parameter* supervised fine-tuning with the following configuration: bfloat16 precision, 3 epochs, per-device batch size 1, gradient accumulation steps 8, effective batch size 32 ($1 \times 8 \times 4$ GPUs), learning rate 1×10^{-5}, warmup ratio 0.05, weight decay 0.01, SDPA attention, frozen ViT, gradient checkpointing enabled, and response standardization enabled. Each task-specific model is trained independently. For WORKVERI 3V, training uses 8,337 examples from the CHS SCENARIO (Cup Head Screws). For ASSYVERI image, training uses 627 examples from the MES SCENARIO and PES SCENARIO. Training data is converted to the standard single-turn chat format with the grounding-annotated image as input and the MCQ answer letter as the target response.

Scenario-Based Splitting. Unlike random train/test splits, our scenario-based protocol ensures the evaluation assemblies belong to an entirely different product category than those seen during training. For WORKVERI 3V, the CHS SCENARIO (Cup Head Screws) is used for training and the NUTS SCENARIO for evaluation. For ASSYVERI image, the MES SCENARIO and PES SCENARIO (metal and plastic expansion screw assemblies) are used for training, and the CNC SCENARIO (CNC fixture assemblies) for evaluation. This tests whether the model learns *transferable* manufacturing reasoning principles rather than assembly-specific visual patterns.

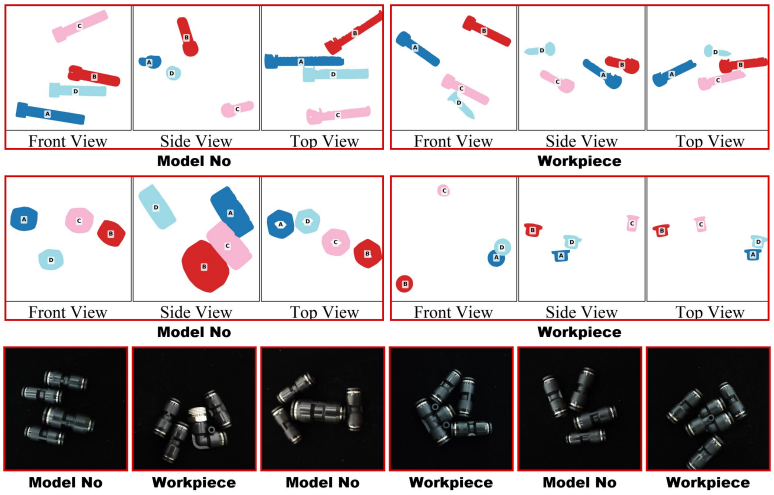
Full Comparison with Frontier Models. Table 14 compares the finetuned Qwen2.5-VL-3B with five reference models evaluated on the same held-out scenarios.

Model	Type	WORKVERI 3V (NUTS SCENARIO)	ASSYVERI Img (CNC SCENARIO)
Gemini 3 Flash	Closed	**63.6**	**53.5**
GPT-5.2	Closed	62.6	37.0
Seed 1.6	Closed	60.0	23.6
Qwen3-VL-235B	Open	54.4	19.7
Mistral 3 8B	Open	35.4	26.0
Qwen2.5-VL-3B (base)	Open	28.2	24.0
Qwen2.5-VL-3B (SFT)	Open	**53.8**	**30.5**

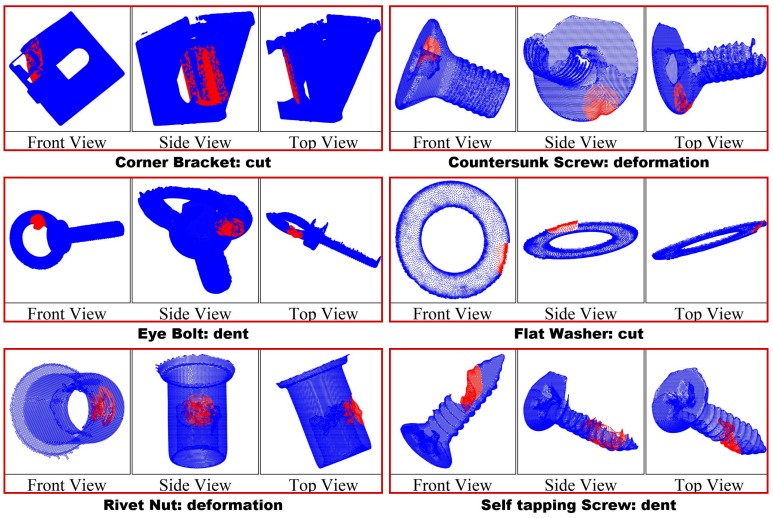
Table 14: **Finetuned 3B model vs. frontier models on held-out scenarios (zero-shot accuracy %).** The finetuned Qwen2.5-VL-3B nearly matches Qwen3-VL-235B on WORKVERI (53.8% vs. 54.4%) and surpasses Seed 1.6, Qwen3-VL-235B, and Mistral 3 8B on ASSYVERI. **Bold** / Underlined / Shaded = best/2nd/3rd among the five reference models.

Task-Level Analysis. The asymmetry between the two tasks is informative. WORKVERI 3V requires distinguishing a single mismatched part among geometrically similar components, a skill that transfers well across scenarios because the training data teaches the model *what constitutes a specification mismatch* in general, regardless of the specific assembly type. In contrast, ASSYVERI image involves diverse error modes (different model numbers and extra parts) across assembly categories with varying visual complexity; a scenario-held-out evaluation is therefore inherently harder because the model encounters part types and spatial layouts it has never seen. Nevertheless, even 627 training examples produce a measurable improvement on a completely novel product category, indicating that FORGE annotations encode transferable manufacturing knowledge.

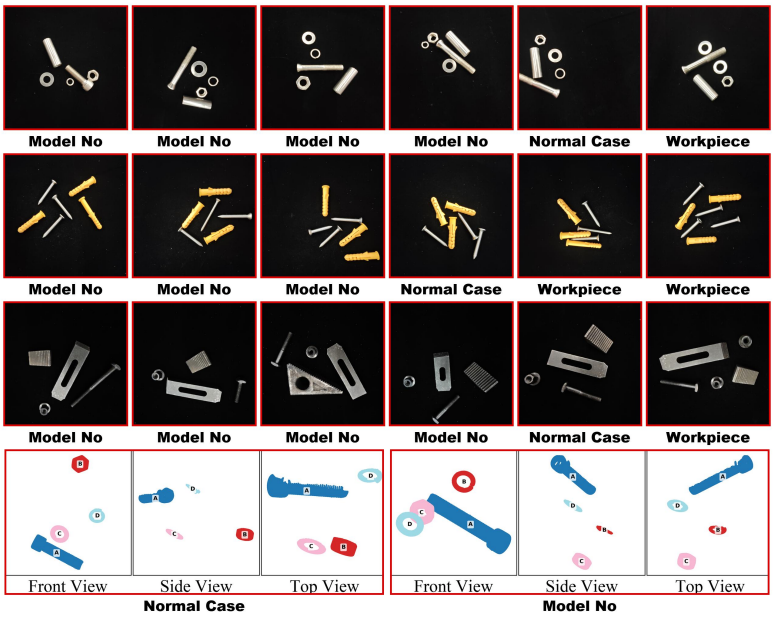
Per-Category Breakdown. On WORKVERI 3V (NUTS SCENARIO), the finetuned model shows a notable shift in behavior: accuracy on Normal cases improves dramatically from 24.3% to 74.8%, while accuracy on Different Model Number cases decreases from 33.8% to 23.8%. This indicates the model learns a more conservative decision boundary, correctly recognizing conforming assemblies far more often, at the cost of some sensitivity to subtle specification mismatches. Since Normal cases constitute 59% of the evaluation set (115/195), the net effect is a large overall improvement. Future



(a) Data examples of WORKVERI



(b) Data examples of SURFINSP



(c) Data examples of ASSYVERI

Figure 7: Qualitative results and analysis. (a) WORKVERI: The top row shows examples of CHS SCENARIO, the second rows are from NUTS SCENARIO, and the bottom row is from PCS SCENARIO. (b) SURFINSP: Randomly selected defect examples across different workpieces. Red highlighting is for illustration only; during evaluation, models use uniformly colored point clouds without explicit defect cues. (c) ASSYVERI: The first three rows correspond to MES SCENARIO, PES SCENARIO, and CNC SCENARIO, respectively, while the last row shows examples of SWN SCENARIO.

work could explore loss weighting or balanced sampling to mitigate this calibration trade-off.

On AssyVeri image (CNC scenario), improvements are more uniform: Different Model Number accuracy increases from 20.5% to 27.0% and Extra Part from 37.0% to 44.4%, suggesting the training signal transfers consistently across error types to a novel product category.

Additional training results based on InternVL3-2B Building on the results previously reported with Qwen2.5-VL-3B, we expanded the study to a second base model from a different model family, InternVL3-2B, to assess whether the observed benefits of SFT generalize beyond a single architecture. We repeated the identical scenario-held-out protocol, using the same training sets, training recipe, and zero-shot evaluation harness, changing only the base model. The results provide a broader view of the effect of SFT on held-out manufacturing scenarios. On WorkVeri-3V/NUTS, InternVL3-2B improves from 26.2% to 55.4% (+111.5%), slightly exceeding the fine-tuned Qwen2.5-VL-3B result of 53.8% and matching the zero-shot performance of the much larger Qwen3-VL-235B (54.4%). This replication strengthens our original finding that the improvement is not specific to one architecture. At the same time, the results clarify that the trade-off is not universal: on Different-Model-Number cases, where Qwen2.5-VL-3B decreases from 33.8% to 23.8%, InternVL3-2B instead improves from 25.0% to 33.8%. For AssyVeri-Image/CNC, however, InternVL3-2B shows essentially no change (30.7% to 30.7%), while Qwen2.5-VL-3B improves from 24.0% to 30.5%, indicating a boundary of the current approach. We therefore view these results as broader, but still preliminary, evidence that SFT on FORGE annotations can improve zero-shot generalization to unseen manufacturing scenarios, with the magnitude of the benefit depending on both the model and the amount of available training data.

Held-out Scenario (zero-shot acc. %)	Qwen2.5-VL-3B	InternVL3-2B
WorkVeri-3V / NUTS	28.2 → **53.8** (+90.8%)	26.2 → **55.4** (+111.5%)
Normal cases	24.3 → **74.8**	27.0 → 70.4
Different-Model-Number cases	33.8 → 23.8	25.0 → **33.8**
AssyVeri-Image / CNC	24.0 → **30.5** (+27.1%)	30.7 → **30.7** (±0)

Table 15: **Performance on held-out scenarios (zero-shot accuracy %).** Comparison between the Qwen2.5-VL-3B model and the new InternVL3-2B model. The values show performance before and after finetuning, with the relative improvement reported in parentheses for the overall benchmark scores.

C Extended Error Cases and Bottleneck Analysis

This section supplements the bottleneck analysis (Section 4.4) of the main text with complete transcripts and detailed experimental configurations. Section C.2 then provides the experimental settings and configurations for all three bottleneck analyses (visual grounding, missing-part detection, and raw point cloud text input), followed by additional discussion of the findings.

C.1 Qualitative Discussion of Error Patterns

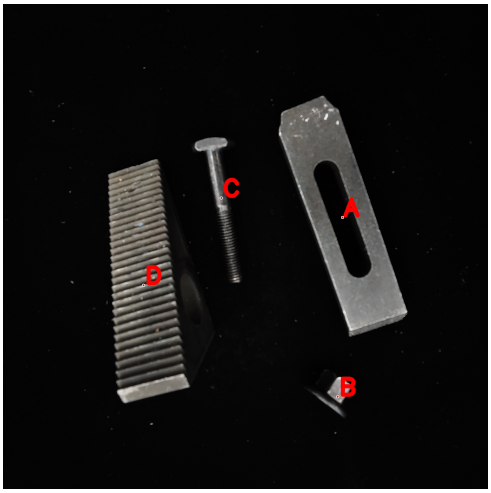
This subsection provides an extended qualitative discussion of the two error patterns illustrated in Figure 3 of the main text, with their corresponding model outputs detailed in Figure 8.

A. Misjudging and over-relying on material properties. Error Case 1 (MES scenario, shown in Figure 8(a)) illustrates this pattern. Here, the MLLM hallucinates material properties from visual textures, misidentifying a metal Flat Washer (E) as "plastic/nylon." It then overcomplicates the reasoning process by relying on this erroneous material inference to select it as the anomalous part (predicting E instead of the correct answer A), even when the specific problem does not require material considerations. Despite the incorrect conclusion, this response demonstrates the model's inclination to autonomously reason about material composition—a capability that could become highly valuable if grounded in accurate perception.

B. Failure in specification recognition, but showing emerging capabilities in service condition assessment. Error Case 2 (CNC scenario, shown in Figure 8(b)) corresponds to point B in Section 4.3. While the MLLM successfully identifies all workpiece types in the CNC fixture assembly, it misjudges their specific dimensions. Specifically, it incorrectly concludes that the Hex Nut (B) is undersized relative to the T-slot bolt, when the actual anomaly is an oversized Step Block (D). Despite this failure in model number/specification recognition, the model's intermediate reasoning reveals emerging capabilities for evaluating service conditions. By casually noting that "the strap clamp (A) is heavily worn/chipped" and "block (D) is also worn," the model demonstrates a potential for graded degradation assessment. Recognizing such wear levels and integrating them into diagnostic reasoning suggests that MLLMs could eventually provide perceptual feedback to support **Predictive**

Maintenance (PdM) (Sun et al., 2023, 2018).

Together, the model outputs in Figure 3 and example images in Figure 8 provide the full reasoning chains behind these error patterns, enabling readers to inspect the exact prompts, intermediate reasoning steps, and final predictions.



(a) Error Case 1: Misjudging Material Properties



(b) Error Case 2: Failure in Specification Recognition

Figure 8: Two error cases with red annotations added for clearer interpretation, illustrating the reasoning chains discussed above.

ASSYVERI CNC SCENARIO Error Cases 1

- **System Prompt:**

  General task:

  - You will be given a single image of an assembly containing multiple industrial parts that match each other well.

  - Your job is to identify which part in the provided test image uses the wrong model number (incorrect size or shape) and thus deviates from the rest.

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (A, B, C, ...), each describing one candidate part location.

  - Choose the letter corresponding to the mismatched part.

  Input format:

  - Query: a single line stating the task.

  - Image: the image to evaluate.

  - Choices: a list of options, each with a normalized coordinate for one part in the image (normalized to [0,1]).

  Output format:

  - Return exactly one letter in the format " Answer: X " at the end of your response.

  - You could include your reasoning before the answer.

ASSYVERI CNC SCENARIO Error Cases 1

- **Setting:** Zero-Shot
- **User Prompt:**

  In CNC machining, the primary role of fixtures is to ensure that the workpiece is accurately positioned and firmly secured during various cutting, milling, or drilling operations.

  A typical CNC fixture is composed of several fundamental elements that work together to deliver reliable clamping performance. These include: (1) the flange nut, (2) the T-bolt half-thread screw, (3) the pressure plate, and (4) the triangular wedge.

  Based on the provided reference images of correctly assembled Computer Numerical Control (CNC) fixture, check the test image for any extra, unnecessary, or incorrectly specified parts. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.70, 0.44]

  B. Part at [0.69, 0.81]

  C. Part at [0.45, 0.40]

  D. Part at [0.29, 0.58]

  Answer:

ASSYVERI CNC SCENARIO Error Cases 1

- **Model:** Gemini-2.5-Flash
- Gt answers: D
- Pred answer:B
- **Raw Response 1:**

  The image displays components typically found in a clamping set for a CNC machine (T-slot clamp set). These usually include a T-slot bolt, a step block (or serrated block), a strap clamp (or clamping strap), and a flange nut.

  The parts present are:

  1. A serrated/step block (left side, [0.29, 0.58]).

  2. A T-slot bolt (center, [0.45, 0.40]).

  3. A strap clamp (right side, [0.70, 0.44]).

  4. A nut (bottom right, [0.69, 0.81]).

  The standard nut used in a T-slot clamping set is a **flange nut** (or sometimes a coupling nut or standard hex nut, depending on the specific kit). A flange nut has a wide flange at one end that acts as an integrated washer, distributing the pressure over a larger area.

  The part at location B ([0.69, 0.81]) is a nut. However, it appears to be a standard **hex nut** or possibly a **T-nut** (though it looks more like a hex nut based on the visible facets and lack of a clear T-shape profile for a T-slot).

  Let's re-examine the parts based on typical clamping sets:

  * A ([0.70, 0.44]): Strap clamp. Correct component.

  * B ([0.69, 0.81]): Nut. This appears to be a standard hex nut, which is often used, but sometimes a flange nut is preferred for better clamping force distribution. However, looking closely at the shape, it seems to be a standard hex nut, which is a common component.

  * C ([0.45, 0.40]): T-slot bolt. Correct component.

  * D ([0.29, 0.58]): Step block/serrated block. Correct component.

  In a typical T-slot clamping set, all four components are necessary. If we are looking for a part with an *incorrect specification* (wrong size or shape relative to the others), we must compare them.

  The T-slot bolt (C) and the strap clamp (A) appear to be sized appropriately for each other (the slot in A looks like it would fit the bolt C). The step block (D) is also standard.

  Let's look closely at the nut (B). It is a small hex nut. If this kit is intended to be a complete, matched set, the nut should match the thread size of the bolt (C). Without knowing the exact thread size, it's hard to confirm a mismatch based on size alone.

ASSYVERI CNC SCENARIO Error Cases 1

- **Raw Response 2:**

  However, if we consider the overall condition and type:

  The strap clamp (A) is heavily worn/chipped at the ends.

  The step block (D) is also worn.

  The T-slot bolt (C) is also worn.

  The nut (B) is visually distinct from the other three main components (A, C, D) in terms of its finish and apparent size relative to the bolt (C). It looks disproportionately small or perhaps like a different type of nut than what is typically paired with a T-slot bolt of that size, especially if a flange nut is expected.

  Let's assume the task requires identifying the part that is visually the most likely mismatch in terms of model/specification.

  The part at B is a small hex nut. Given the size of the T-slot bolt (C), this nut seems undersized for heavy-duty clamping, or it might be a nut for a different, smaller bolt. If the bolt (C) is M12 or M16, the nut (B) looks too small to be the corresponding standard hex nut.

  Comparing the parts again:

  A, C, and D are clearly industrial clamping components. B is a fastener (nut). If B is intended to mate with C, its size relative to the threaded portion of C seems questionable. The diameter of the nut (B) appears significantly smaller than the head of the bolt (C), and possibly smaller than the threaded shaft diameter, suggesting it might be the wrong size or thread pitch for the bolt C.

  Therefore, the nut at B is the most likely candidate for an incorrectly specified part (wrong size/model number) relative to the T-slot bolt (C).

  Answer: B

ASSYVERI MES SCENARIO Error Cases 2

- **Setting:** Zero-Shot
- **Query**

  Based on the provided reference images of correctly assembled Metal expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.77, 0.44]

  B. Part at [0.69, 0.52]

  C. Part at [0.49, 0.47]

  D. Part at [0.30, 0.51]

  E. Part at [0.63, 0.43]

  Answer:

- **Model:** Gemini-2.5-Flash
- GT answers: A
- Pred answer: E

ASSYVERI MES SCENARIO Error Cases 2

- **Raw Response:**

  The standard assembly for a Metal expansion bolt consists of five components: Bolt, Nut, Flat Washer, Spring Washer, and Sleeve.

  Let's identify the parts present in the image:

  1. Sleeve (leftmost cylindrical part): Part at [0.30, 0.51] (D). This is a required component.

  2. Bolt (long threaded part with a conical head): Part at [0.49, 0.47] (C). This is a required component.

  3. Nut (large hexagonal part): Part at [0.69, 0.52] (B). This is a required component.

  4. Flat Washer (large white/translucent ring): Part at [0.63, 0.43] (E). This is a required component.

  5. Small Ring (small metal ring): Part at [0.77, 0.44] (A). This appears to be a small metal washer, likely a Spring Washer (or a second, smaller flat washer). A Spring Washer is a required component.

  The image contains five parts, which matches the expected number of components (Bolt, Nut, Flat Washer, Spring Washer, Sleeve).

  However, upon closer inspection of the components:

  * The part at [0.63, 0.43] (E) is a white/translucent plastic or nylon washer (Flat Washer).

  * The part at [0.77, 0.44] (A) is a small, thin metal ring/washer. This is likely intended to be the Spring Washer, but it is a simple flat washer, not a split/spring washer, and it appears to be significantly smaller than the bolt's diameter, suggesting it might be the wrong size or type.

  * The part at [0.69, 0.52] (B) is the Nut.

  * The part at [0.49, 0.47] (C) is the Bolt.

  * The part at [0.30, 0.51] (D) is the Sleeve.

  The most significant deviation is the presence of a plastic/nylon Flat Washer (E) and the absence of a metal Flat Washer and a proper Spring Washer.

  In the context of industrial parts matching, the most obvious deviation in material and specification is the Flat Washer at [0.63, 0.43] (E). Expansion bolts are typically high-strength metal assemblies, and using a plastic/nylon washer for load distribution (Flat Washer) is an incorrect specification that compromises the assembly's strength and durability, especially when the other components are metal.

  If we assume the small metal ring (A) is meant to be the Spring Washer, then the plastic washer (E) is the incorrect component due to its material.

  Therefore, the part at [0.63, 0.43] is incorrectly specified (wrong material).

  Answer: E

C.2 Details of Bottleneck Analysis

C.2.1 Settings and Configurations.

This section details the experimental setup for each of the three bottleneck analyses introduced in the main text. Unless stated otherwise, all analyses use a temperature of 0.0 for deterministic outputs and evaluate five representative models spanning both open- and closed-source families: Gemini-3-Flash (closed), GPT-5.2 (closed), Qwen3-VL-235B (open), Seed-1.6 (closed), and Mistral-3-8B (open). All models are accessed via the OpenRouter API.

A. Visual Grounding (Table 4 in the main text). This analysis isolates each model's ability to ground and localize individual components, decoupled from downstream reasoning. Each assembly image is annotated with Set-of-Mark (Yang et al., 2023) letter labels (A, B, C, *etc.*) on individual parts.

- **Single-image grounding** (500 sampled cases from 4,257 total). Two complementary sub-tasks are tested: (i) *Coordinate-to-Letter* (C→L): given a normalized coordinate $[x, y] \in [0, 1]^2$ (origin at the top-left corner), the model must identify the letter of the corresponding part; (ii) *Letter-to-Coordinate* (L→C): given a letter label, the model must select the correct coordinates from an MCQ. Images are resized so that the longest side is at most 1,568 pixels.

- **Cross-image correspondence** (513 sampled pairs from 14,992 total). Two images from the same assembly scenario are shown, each with independently assigned letter labels. A part is identified in the first (reference) image, and the model must locate the visually corresponding part in the second (test) image. Two modes are evaluated: (i) *Letter-to-Letter* (L→L) and (ii) *Coordinate-to-Coordinate* (C→C).

All grounding evaluations use zero-shot prompting with `max_tokens` = 512 and reasoning disabled. Grounding image examples are shown in Figures 30–33 in Section D.2

B. Missing Part Detection (Table 5 in the main text). This analysis probes domain-specific reasoning by providing models with explicit assembly specifications. The model receives a complete description of the assembly (part names, quantities, and functional descriptions) and must identify

which component is absent from an MCQ. Each case includes a "No part is missing" option.

- **Image**: 240 test cases across three scenarios (MES SCENARIO with 6 MCQ options, PES SCENARIO with 3 options, and CNC SCENARIO with 5 options).
- **Three-view**: 137 test cases from the SWN SCENARIO (5 MCQ options).

All missing-part evaluations use zero-shot prompting with `max_tokens` = 8,192 and extended reasoning enabled.

Table 16: **Raw point cloud text input Bottleneck Analysis (accuracy %).** 3D coordinates are serialized as integer-scaled text tables and fed directly to MLLMs, bypassing visual rendering. ZS = Zero-Shot, RC = Ref-Cond.

Model	Type	AssyVeri			SurfInsp			WorkVeri		
		ZS	RC	ICD	ZS	RC	ICD	ZS	RC	ICD
Gemini-3-Flash	Closed	25.2	32.7	35.0	22.6	19.8	20.2	**53.6**	**48.1**	**37.1**
Qwen3-235B	Open	25.2	**34.2**	32.7	11.1	10.9	17.7	35.0	39.1	36.7
Random baseline		25.0			20.0			25.0		

C. Visual projection is necessary: the text channel cannot replace it for generic MLLMs. Since our benchmark targets general-purpose MLLMs that lack native 3D encoders (Section 3.1), a natural question is whether the *text* modality can serve as an alternative 3D interface by feeding raw coordinates directly as token sequences.

This analysis investigates whether the text modality can serve as an alternative 3D interface for general-purpose MLLMs. Point clouds are serialized as integer-scaled text tables and fed directly to the model, bypassing visual rendering. Two representative models are evaluated: Gemini-3-Flash and Qwen3-235B (text-only mode).

The serialization procedure is as follows: each part's 3D coordinates are randomly sampled to $n_{\text{points}} = 500$ points, then scaled by 10^d where $d = 2$ (round digits), yielding integer coordinates. The resulting table format is:

```
x, y, z
1896, 750, 31970
3732, -3451, 32051
...
```

For WorkVeri and AssyVeri, assemblies are segmented into individual parts via DBSCAN clustering ($\varepsilon = 3.0$, min_samples $= 50$), each assigned a random letter label (A, B, C, *etc.*). For SurfInsp, a single part's coordinates are provided. All three evaluation modes are tested (Zero-Shot, Ref-Cond, and ICD) with `max_tokens` = 1,024 across WorkVeri (420 cases), SurfInsp (460 cases), and AssyVeri (266 cases). We test this by serializing point clouds as integer-scaled text tables and querying two representative models: Gemini-3-Flash (multimodal) and Qwen3-235B (text-only). As shown in Table 16, both models perform near the 20% random baseline on SurfInsp (surface defect classification), with per-class analysis revealing that models default to a dominant prediction rather than genuinely discriminating defect types. Only WorkVeri shows a moderate signal above chance (Gemini-3-Flash: 53.6% zero-shot), suggesting that coarse-grained shape comparison can partially exploit coordinate distributions. These results confirm that, among the input channels available to general-purpose MLLMs, visual rendering via multi-view projection is a relatively more effective interface for 3D manufacturing data.

D. Specialized 3D–language models on native point clouds. The analysis above compares two input channels that are available to general-purpose MLLMs. A complementary question is whether a model with a *native* 3D encoder does better when it receives the point cloud directly. Because FORGE poses its tasks as free-form textual instructions, a comparable baseline must be an instruction-following, open-domain model; classical industrial 3D pipelines instead emit anomaly scores or fixed labels. We therefore evaluate the two most established publicly available 3D point-cloud–language models, PointLLM-13B (Xu et al., 2024) and ShapeLLM-13B (Qi et al., 2024), feeding each the raw scan (no projection, no serialization) under the same zero-shot two-part protocol and the same answer parser as our main evaluation.

We report these baselines zero-shot only. Both models are built for single point-cloud dialogues and have 2,048-token context windows in which a single cloud already consumes more than 500 tokens, whereas our Ref-Cond and ICD settings present two to seven clouds per conversation. We also tested the largest configuration that does fit, one demonstration cloud plus the test cloud: ShapeLLM produced degenerate output, and PointLLM flipped to flagging almost every workpiece, with accuracy falling to 1.8%. Zero-shot is thus the only setting in which these baselines operate within their intended regime.

Table 17 reports the comparison on the **380 workpieces of SurfInsp that are present in all three 3D input channels**, so that every number is measured on identical physical parts. Comparing

Table 17: **3D input channels compared on identical workpieces.** Zero-shot accuracy (%) on the 380 SURFINSP workpieces present in all three channels (332 defective, 48 defect-free; chance = 20). Native 3D models receive raw point clouds; the other rows use the same underlying scans re-encoded for the stated channel.

Input channel	Model	Acc.
Native 3D	PointLLM-13B	12.9
	ShapeLLM-13B	12.6
Serialized coordinates	Gemini-3-Flash	20.3
Three-view (FORGE)	Gemini-3-Flash	23.4
	best of 18 MLLMs	**28.4**
Always answer "normal"		*12.6*

the best model available in each channel, native 3D input scores lowest (12.9), serialized coordinates come next (20.3), and three-view rendering is highest (23.4 for the same model, 28.4 across all 18 MLLMs); with in-context demonstrations the three-view channel reaches 45.8 against 22.9 for coordinates. The projection interface adopted in FORGE is therefore the most effective of the three channels currently available for 3D manufacturing data, and it is domain knowledge rather than the input channel that governs performance on top of it.

The 3D–language scores should be read as degeneracy rather than as partial competence: on the complete 460-case SURFINSP set, PointLLM and ShapeLLM answer "normal" for virtually every workpiece (98.4% and 99.2% accuracy on defect-free parts, 0–2.3% on every defect class), reaching 26.5% and 26.3% overall, which is indistinguishable from the 26.5% obtained by always answering "normal". Four caveats should be kept in mind when interpreting this comparison: both models use 13B Vicuna-era language backbones; both are instruction-tuned for object-level captioning and classification rather than inspection; FORGE scans carry no colour, so the colour channels these models expect were filled with a constant; and the task prompt was merged into the user turn to respect each model's chat template.

C.2.2 Additional Discussion of Bottleneck Analyses.

Detailed examples of bottleneck analyses A, B, and C discussed in the main text are presented in Sections E.4, E.5, and E.6 in this Appendix, respectively.

Recall that the motivation for the "Missing Part" scenario in Bottleneck Analysis B is to further disentangle domain-specific reasoning from explicit grounding capabilities. As demonstrated in Section E.5, this analysis provides a more direct evaluation of the model's intrinsic task comprehension. Because the System and User Prompts strictly exclude coordinate information and constrain the output format (e.g., "Answer with ONLY the letter of your choice"), the model is forced to rely entirely on logical deduction. Consequently, the raw response reveals an extensive reasoning process where the model continuously verifies its assumptions and corrects its own logical errors. This internal debate is vividly captured by phrases such as: " Wait, the assembly needs four parts: Bolt (B?), Nut (maybe C?), Flat Washer (A?), Spring Washer (D?).", "But wait, the problem is to check if all four are present", and "Wait, but a Metal Expansion Bolt might not require a nut?". This pronounced hesitation and reliance on iterative self-correction underscore the model's fragility when deprived of explicit spatial anchors. It highlights a critical gap between merely recognizing individual components and robustly comprehending their relational assembly logic. Meanwhile, the image presentation of the missing part can follow the images (Figure 26–28) shown in Section D.

The analysis of Section E.6 reveals that in the raw response, LLMs spend a significant amount of time repeatedly analyzing and processing point cloud coordinate data, relying on range-related calculations to determine the relative sizes of the corresponding parts. For instance, "Part C's z span is ~6200, which is 3.26 larger than Part A's (1900), 4.77 larger than Part B's (1300), and 2.38 larger than Part D's (2600)." This suggests that when provided with point cloud coordinates as input, LLMs exhibit the potential to assess relative sizes. However, this method of processing coordinate data struggles to analyze microscopic morphology, which explains why MLLMs achieve a certain level of performance in WORKVERI, whereas performance on the remaining tasks is noticeably inferior. Nevertheless, the responses of MLLMs also reveal certain inherent capabilities for processing point clouds; therefore, analyzing coordinate information and absolute values can provide valuable insights from an alternative perspective.

D Image examples

This section presents randomly sampled images from the benchmark dataset to illustrate the visual characteristics and diversity of each scenario.

D.1 The Examples of All Three Tasks

We show examples for all three tasks: WORKVERI scenarios (Figures 9–11), SURFINSP workpiece types (Figures 12–25), and ASSYVERI scenarios (Figures 26–29). Each figure includes both normal cases and representative anomalous cases.


Normal Case


Normal Case


Normal Case


Normal Case


mixed with model (10 to 6)


mixed with model (8 to 4)


mixed with model (12 to 8)


mixed with model (6 to 4)


mixed with Y-type 3-way


mixed with throttle valve


mixed with external-thread elbow


mixed with T-type 3-way

Figure 9: Random examples from PCs SCENARIO

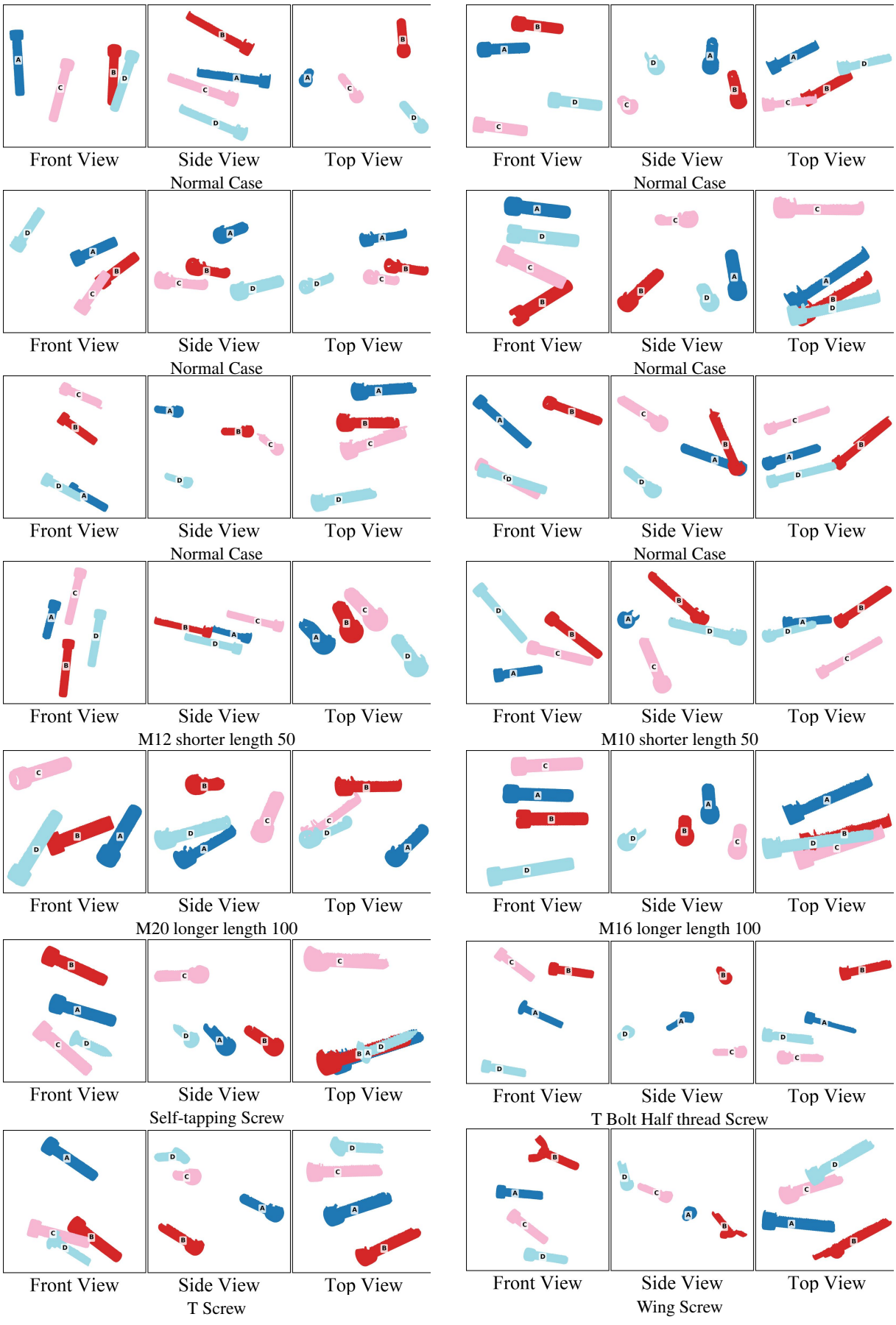


Figure 10: Random examples from CHS SCENARIO

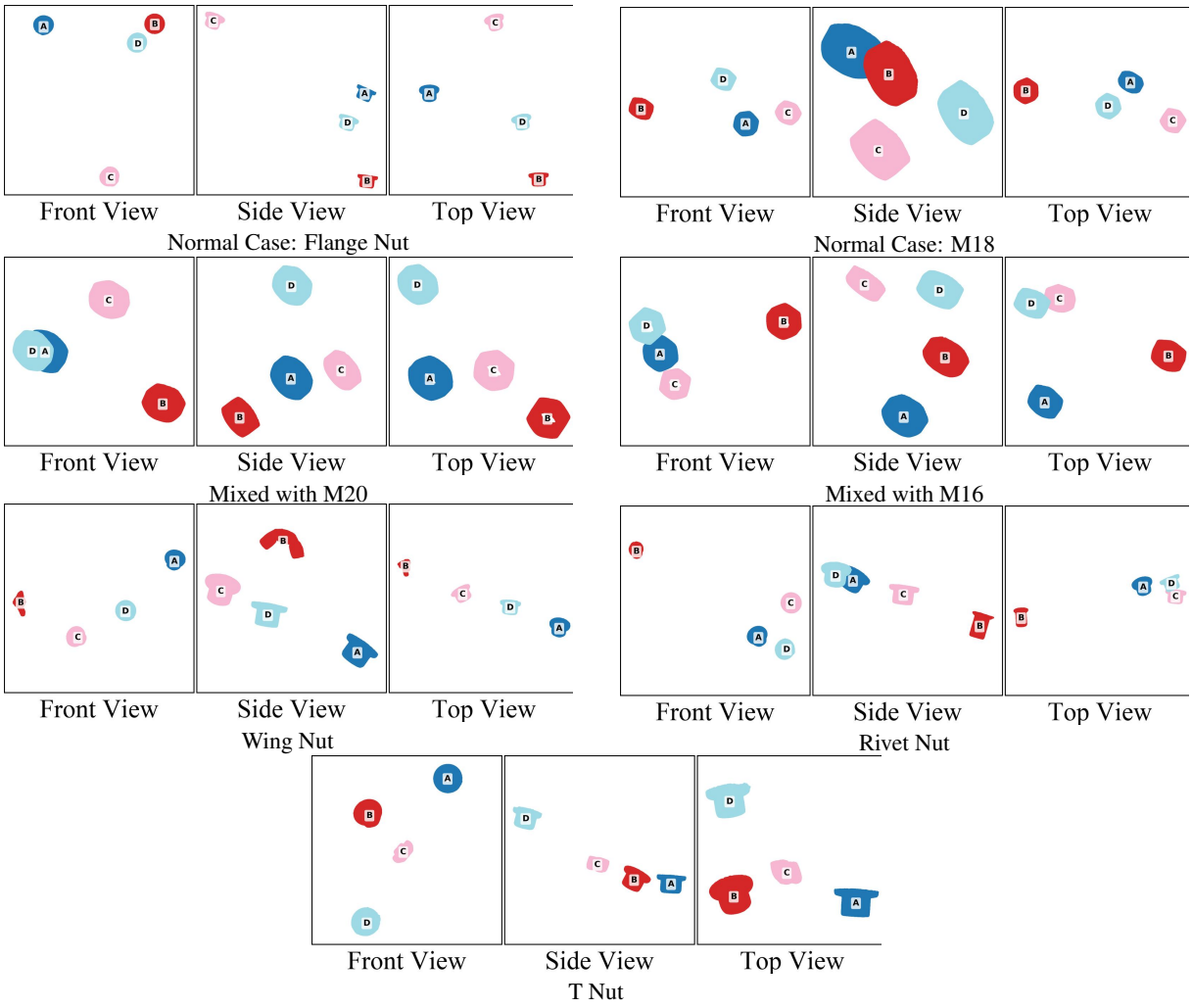


Figure 11: Random examples from NUTS SCENARIO

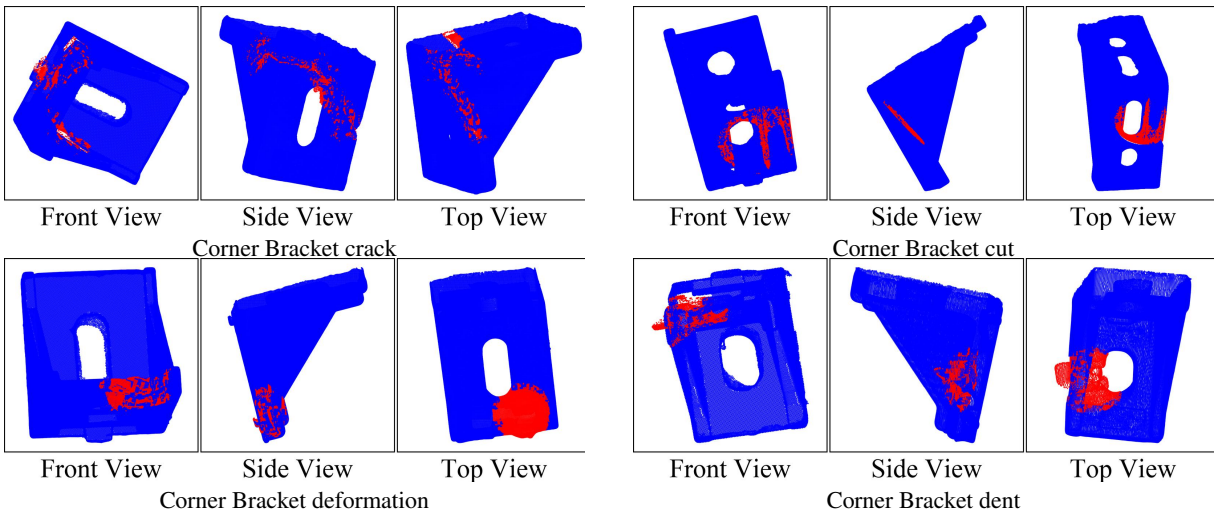


Figure 12: Random examples from Corner Bracket

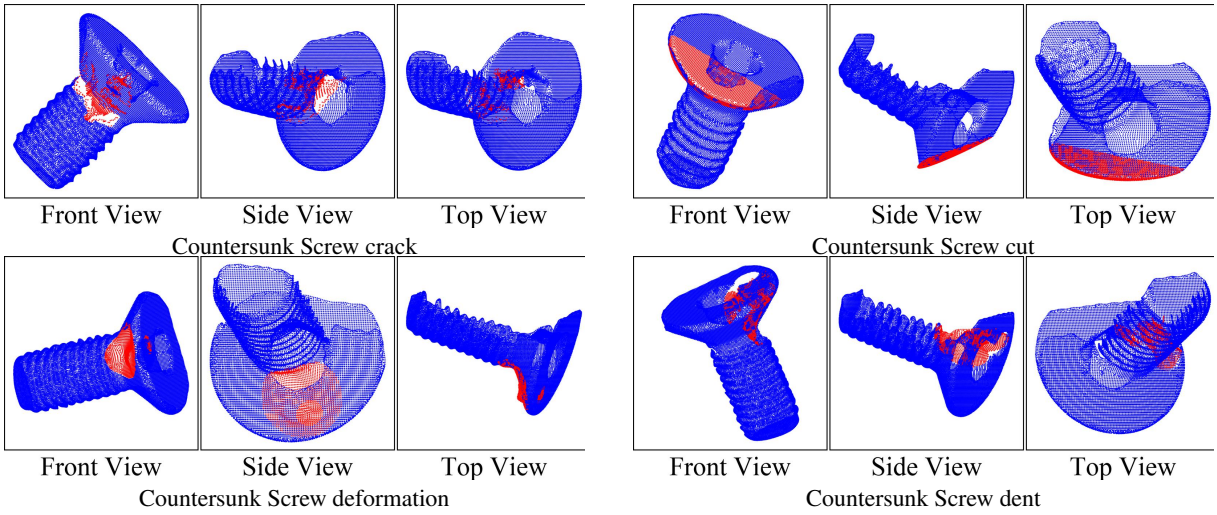


Figure 13: Random examples from Countersunk Screw

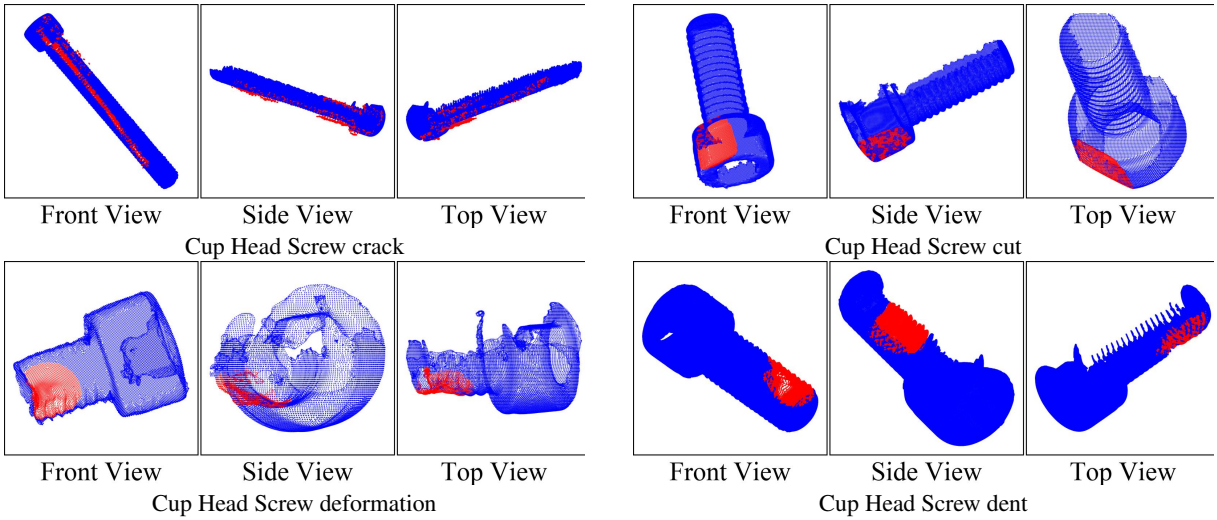


Figure 14: Random examples from Cup Head Screw

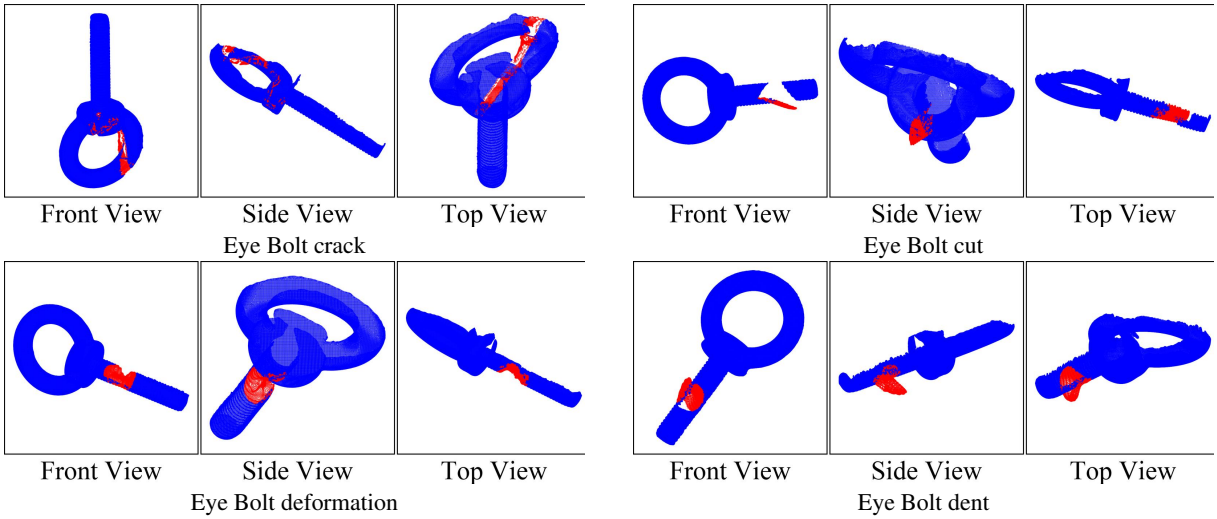


Figure 15: Random examples from Eye Bolt

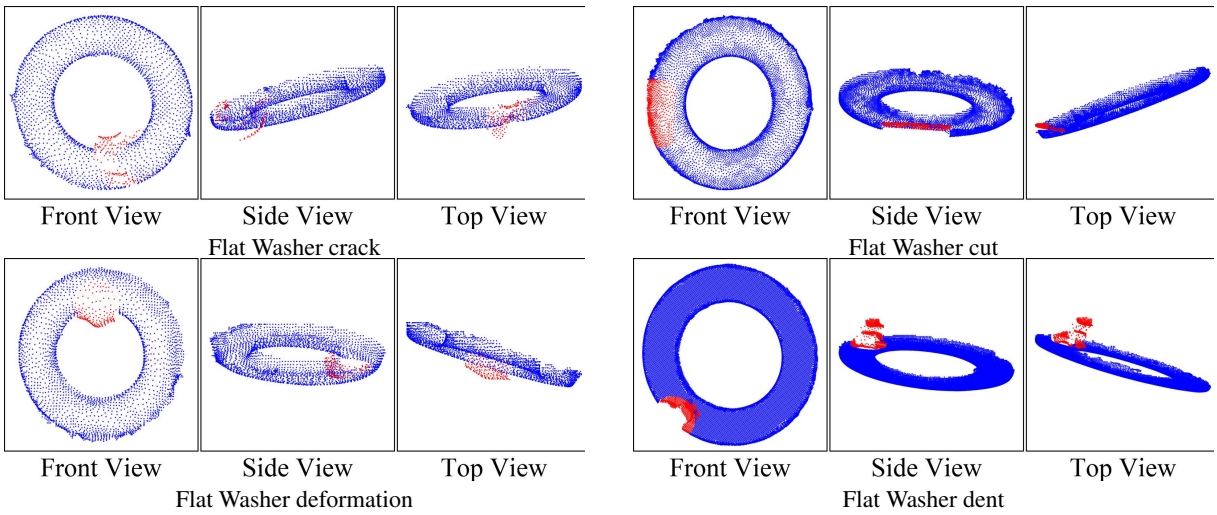


Figure 16: Random examples from Flat Washer

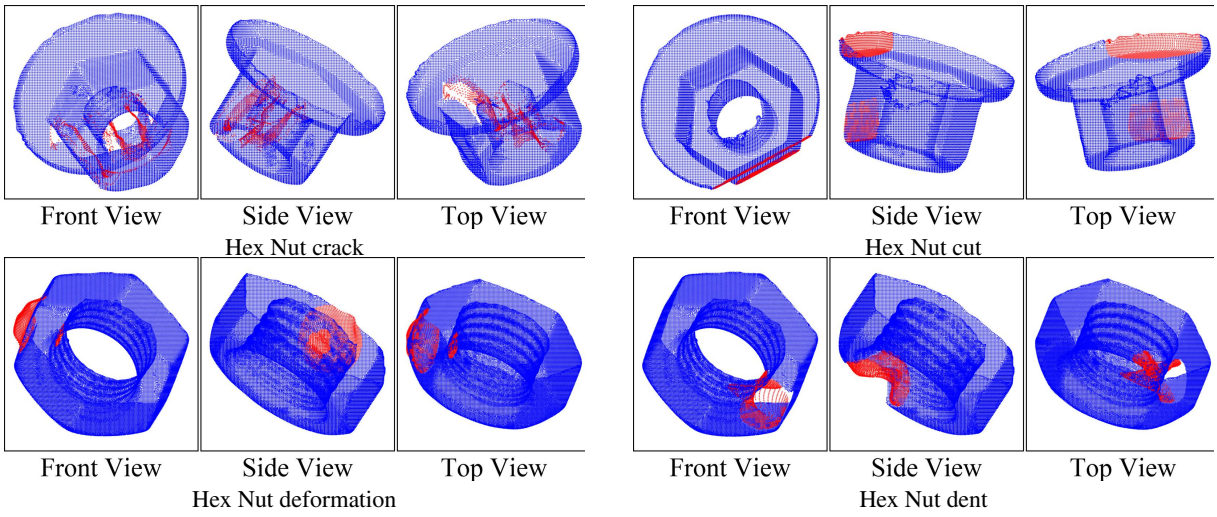


Figure 17: Random examples from Hex Nut

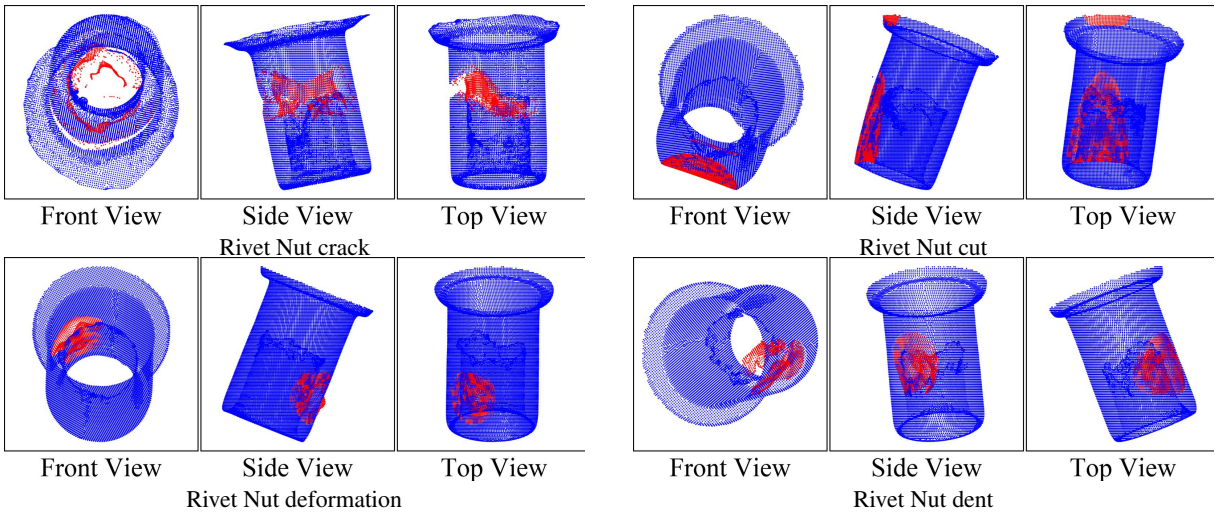


Figure 18: Random examples from Rivet Nut

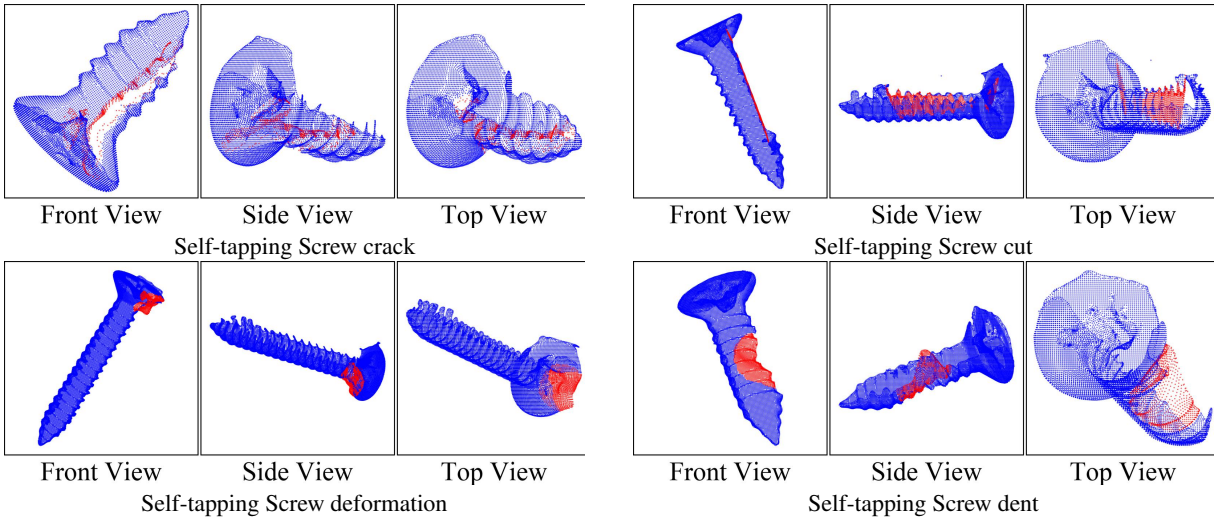


Figure 19: Random examples from Self-tapping Screw

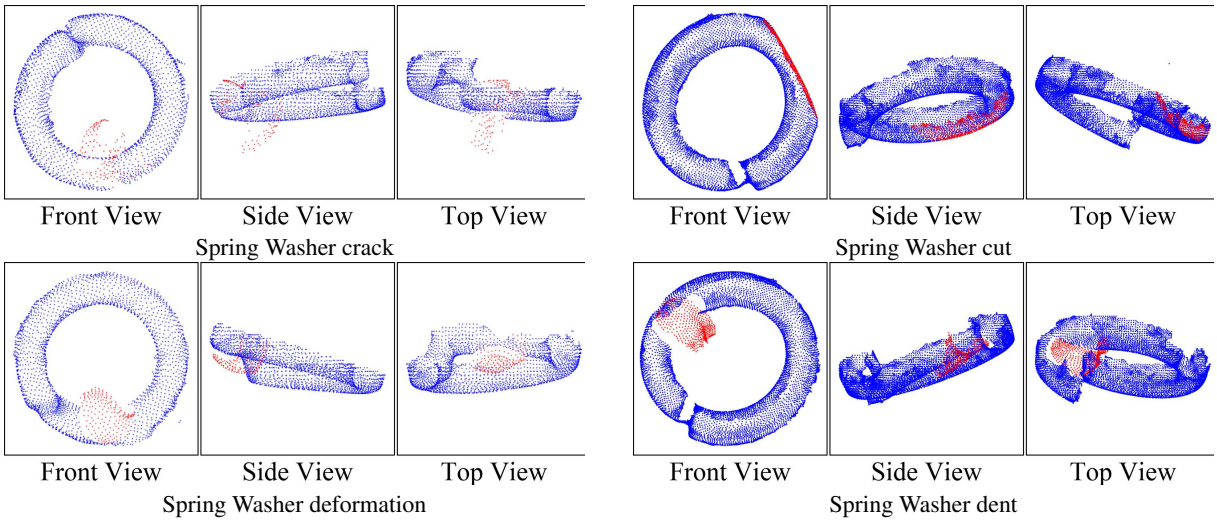


Figure 20: Random examples from Spring Washer

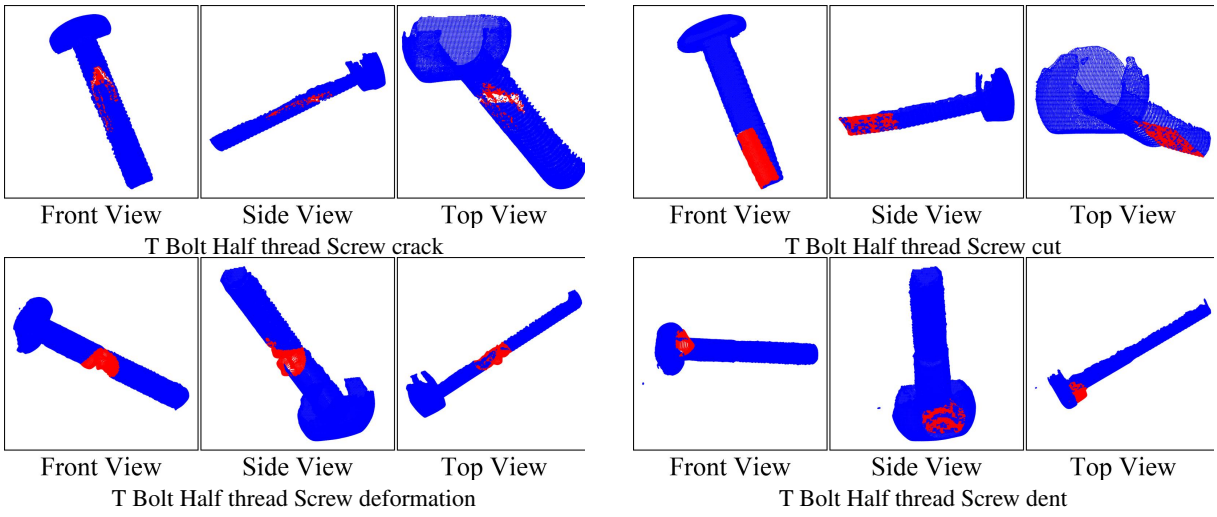


Figure 21: Random examples from T Bolt Half thread Screw

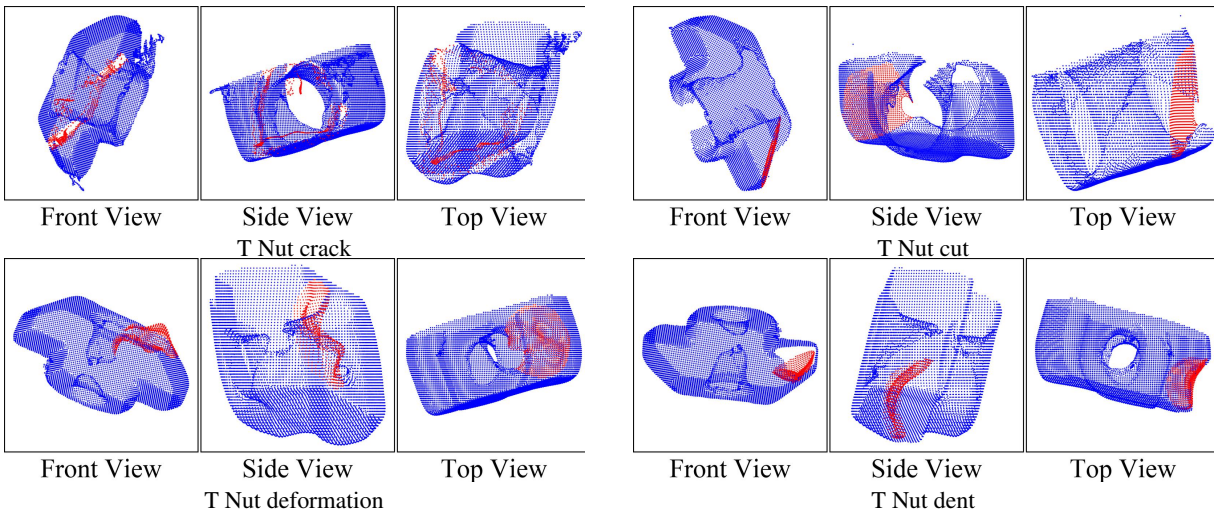


Figure 22: Random examples from T Nut

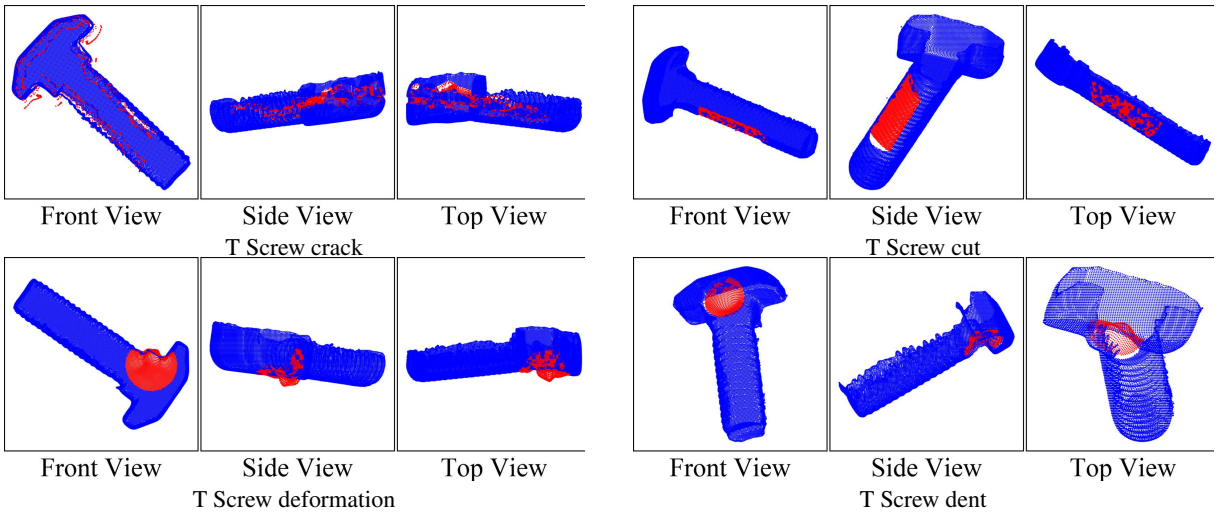


Figure 23: Random examples from T Screw

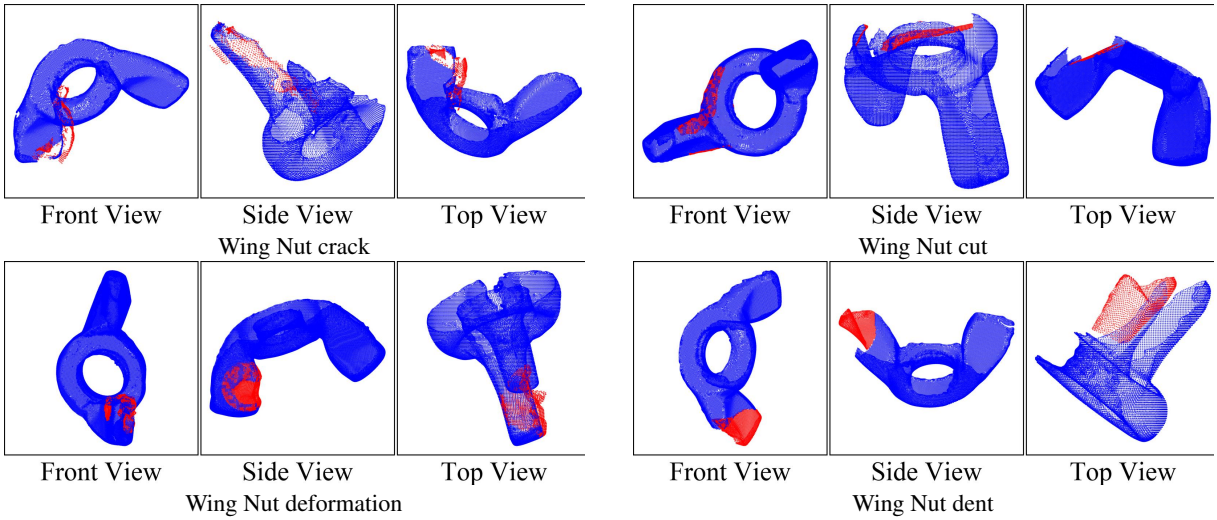


Figure 24: Random examples from Wing Nut

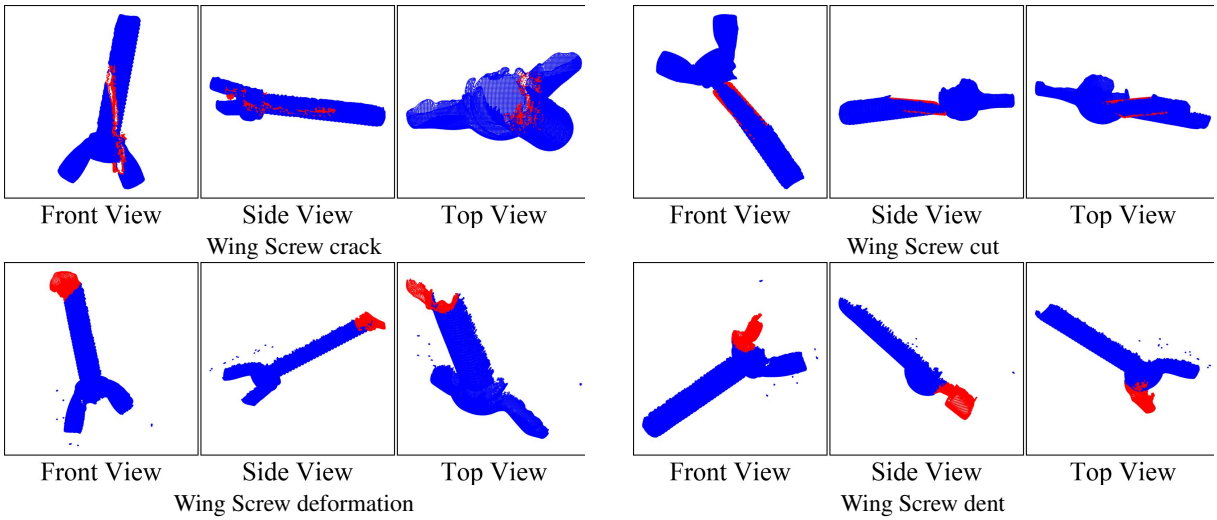


Figure 25: Random examples from Wing Screw



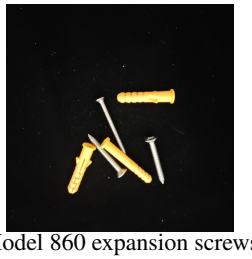
Figure 26: Random examples from MES SCENARIO


Normal Case


Normal Case


Normal Case


Normal Case


Model 860 expansion screws were mistakenly included


One extra plastic expansion anchor


One extra plastic expansion Self-tapping Screw


Screws belonging to model 860 expansion anchors were found mixed in


Plastic anchors belonging to expansion screw model 860 were mixed


Shorter screws were found mixed in


One missing plastic expansion anchor


One missing plastic expansion Self-tapping Screw

Figure 27: Random examples from PES SCENARIO

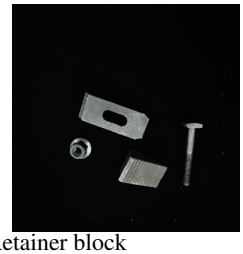

Normal Case


Normal Case


Normal Case


Normal Case


Extra one nut


Retainer block


Short screw


Triangular part


Long screw


Missing triangular part


Missing one nut

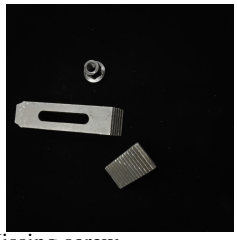

Missing screw

Figure 28: Random examples from CNC SCENARIO

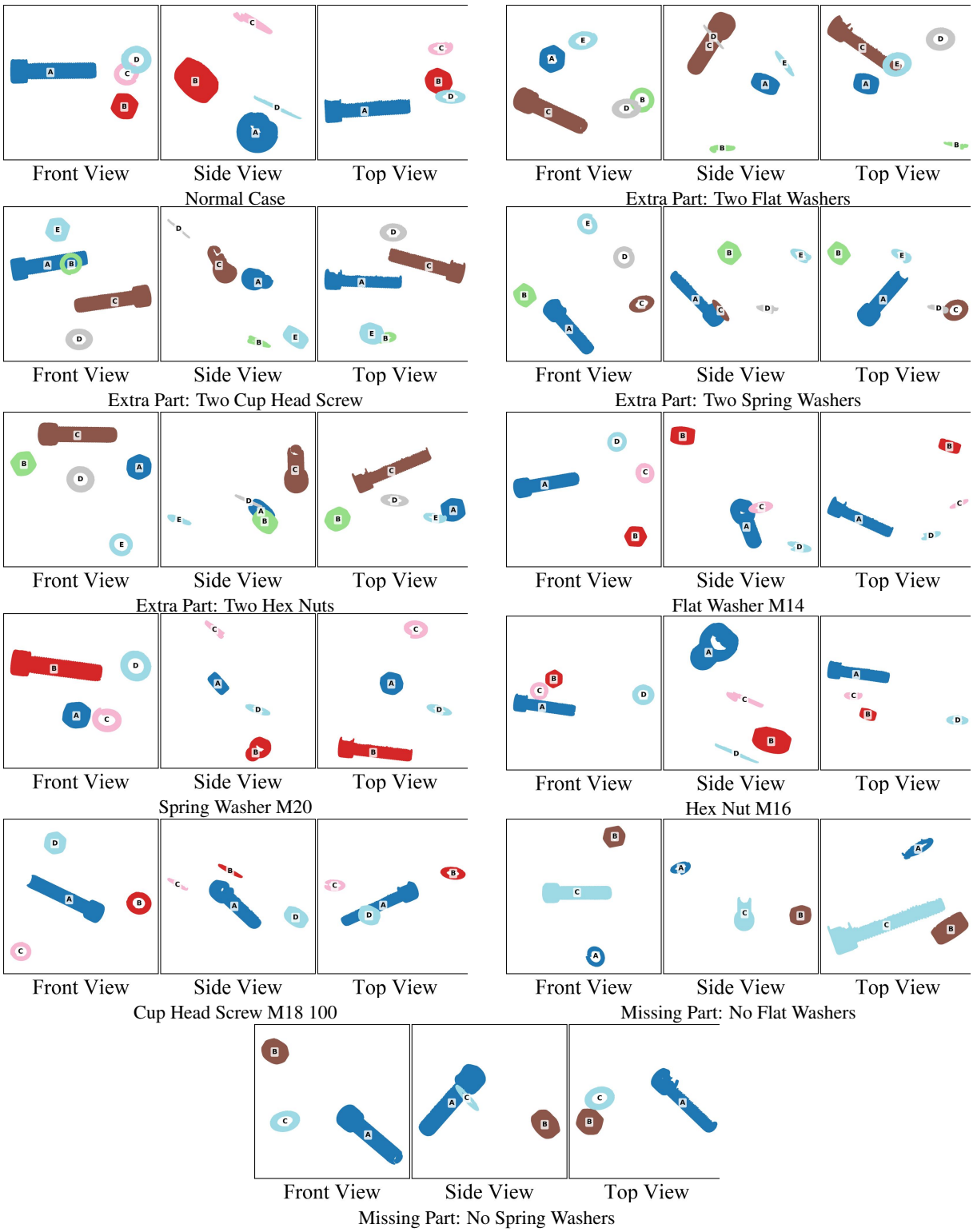


Figure 29: Random examples from SWN SCENARIO

D.2 The Grounding Examples of Image Data (Bottleneck Analysis A)

This subsection shows the Set-of-Mark annotated images used in Bottleneck Analysis A (Section 4.4 of the main text). Each assembly image is overlaid with distinct letter labels (A, B, C, *etc.*) on individual components. These annotated images serve as the visual input for both the single-image grounding tasks (C→L and L→C) and the cross-image correspondence tasks (L→L and C→C). We present examples from all four image-based scenarios: MES SCENARIO (Figure 30), PES SCENARIO (Figure 31), CNC SCENARIO (Figure 32), and PCS SCENARIO (Figure 33).

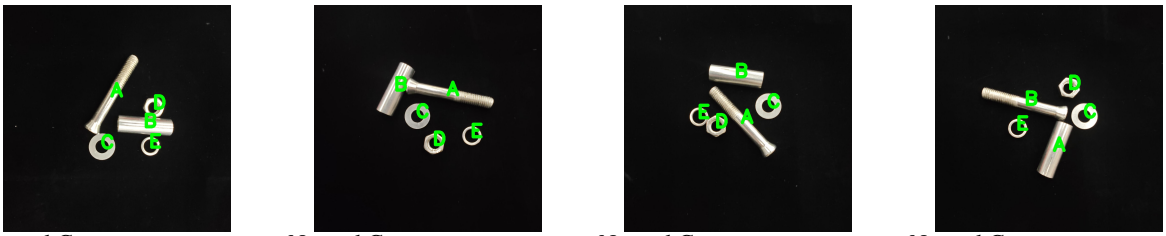


Figure 30: Random grounding examples from MES SCENARIO



Normal Case



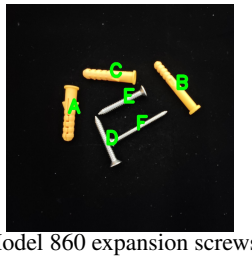
Normal Case



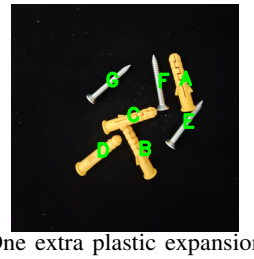
Normal Case



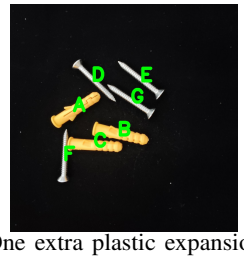
Normal Case



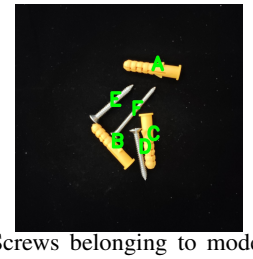
Model 860 expansion screws were mistakenly included



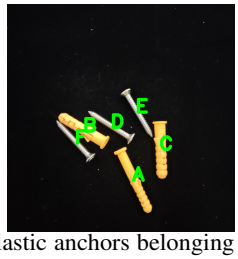
One extra plastic expansion anchor



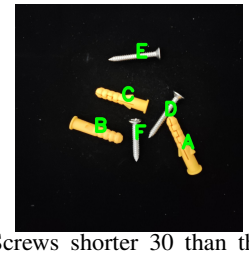
One extra plastic expansion Self-tapping Screw



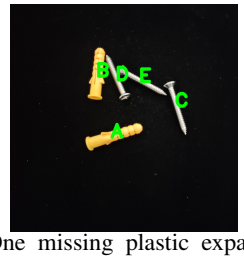
Screws belonging to model 860 expansion anchors were found mixed in



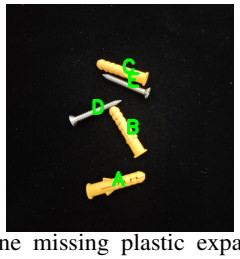
Plastic anchors belonging to expansion screw model 860 were mixed into the ba



Screws shorter 30 than the specified length were found mixed in

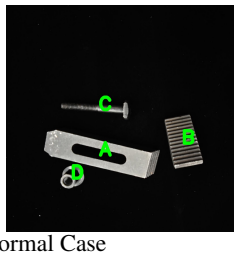


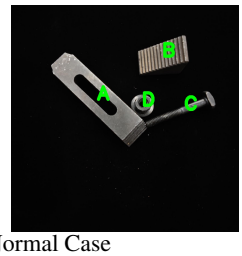
One missing plastic expansion anchor

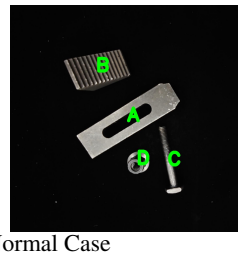


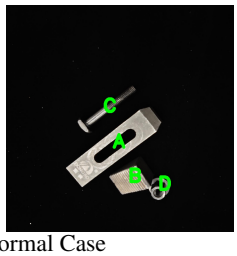
One missing plastic expansion Self-tapping Screw

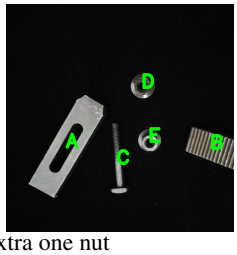
Figure 31: Random grounding examples from PES SCENARIO

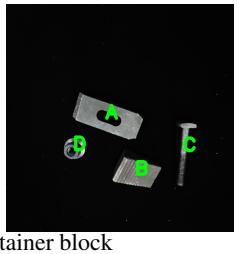

Normal Case

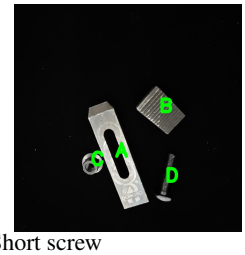

Normal Case

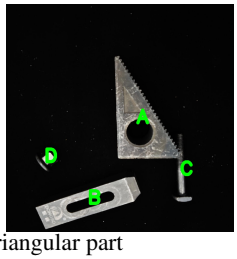

Normal Case

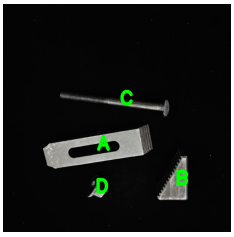

Normal Case

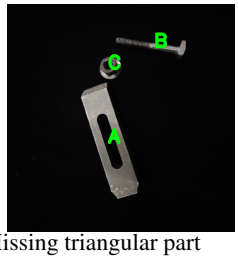

Extra one nut

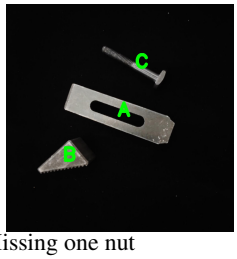

retainer block


Short screw


Triangular part


Long screw


Missing triangular part


Missing one nut

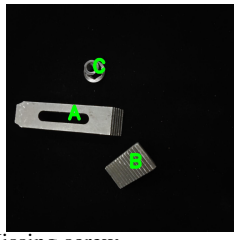

Missing screw

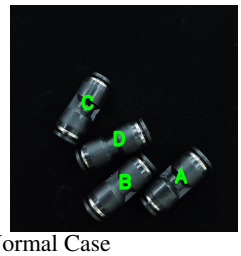
Figure 32: Random grounding examples from CNC SCENARIO



Normal Case



Normal Case



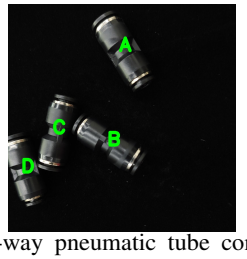
Normal Case



Normal Case



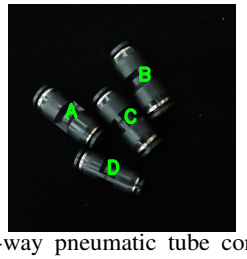
2-way pneumatic tube connectors (8 to 6), mixed with model (10 to 6)



2-way pneumatic tube connectors (8 to 6), mixed with model (8 to 4)



2-way pneumatic tube connectors (8 to 6), mixed with model (12 to 8)



2-way pneumatic tube connectors (8 to 6), mixed with model (6 to 4)



Three 2-way pneumatic tube connectors (8mm to 6mm), mixed with Y-type 3-way



Three 2-way pneumatic tube connectors (8mm to 6mm), mixed with throttle valve



Three 2-way pneumatic tube connectors (8 to 6), mixed with external-thread elbow



Three 2-way pneumatic tube connectors (8mm to 6mm), mixed with T-type 3-way

Figure 33: Random grounding examples from PCs SCENARIO

E Data examples

This section provides complete prompt–response examples for all three main benchmark tasks (Sections E.1–E.3) and for each of the three bottleneck analyses (Sections E.4–E.6). Each example includes the system prompt, user prompt, and the model's raw response, showing the full input–output format used in our evaluation pipeline. For the ICD settings, WORKVERI and ASSYVERI utilize 2 examples by default, whereas SURFINSP uses 10. Note that for certain model providers with strict input limitations on multi-image requests, we appropriately scale down the number of demonstrations for SURFINSP: specifically, 6 for Moonshot Kimi-K2.5, 2 for ByteDance Seed-1.6, and 6 for Google Gemma-3-27B. Additionally, due to space constraints, we only present a single ICD conversation example in the illustrations below.

WORKVERI PCS SCENARIO ICD

- **System Prompt:**

  General task:

  - You will be given one or more reference images showing the correct part type and model/specification, followed by a query and a test image that should match that specification.

  - Your job is to detect which part in the test image deviates from that specification (or from the majority).

  - Instead of returning a coordinate directly, you will be presented with multiple-choice options (A, B, C, ...), each corresponding to one candidate part location in the test image.

  Input format:

  - References: one or more images showing correct parts/assemblies that match the stated specification.

  - Query: a single line stating the part type and model/specification to check.

  - Test Image: the image to evaluate (containing multiple instances of the part).

  - Choices: a list of options, each giving a normalized coordinate [x,y] in [0,1] for one part in the test image(top-left is (0,0), bottom-right is (1,1)).

  Output format:

  - Return exactly one letter in the format "Answer: X" at the end of your response.

  - You may include brief reasoning before the final answer.

  - Use the reference images only to understand the expected specification; make the final decision based on the Test Image and the provided Choices.

E.1 Examples of WORKVERI

Below are examples of WORKVERI (Workpiece Verification) evaluation data covering both image and three-view modalities. The task requires the model to identify which part in an assembly has the wrong model number or specification.

Examples of PCS SCENARIO

WorkVeri PCs scenario ICD

- **User Prompt1:**

  Pneumatic Connectors (PC) are specialized components designed for use in pneumatic systems, mainly serving as the interface between air tubes and pneumatic devices. They are widely applied in industrial automation and machinery, including production lines, automated assembly equipment, packaging machines, and various systems powered by compressed air. With the growing demand for efficiency and reliability in modern manufacturing, PCs have become highly popular and indispensable in ensuring stable operations.

  The primary function of a PC is to provide secure and convenient connections for pneumatic tubing, ensuring stable and efficient air transmission within the system. For example, the commonly used 2-way pneumatic tube connector creates a reliable passage between two air tubes, allowing smooth airflow and proper system operation. These connectors feature a compact design, easy installation, and quick disassembly, which significantly improves efficiency during maintenance and replacement. As a result, they play a critical role as essential components in any pneumatic system.

  REFERENCE 1 (CORRECT EXAMPLE): These are four correct 2-way pneumatic tube connectors (8mm to 6mm) assembly. All parts meet specifications. Use this as a baseline for comparison.

  REFERENCE 2 (CORRECT EXAMPLE): These are four correct 2-way pneumatic tube connectors (8mm to 6mm) assembly. All parts meet specifications. Use this as a baseline for comparison.

  REFERENCE 3 (CORRECT EXAMPLE): These are four correct 2-way pneumatic tube connectors (8mm to 6mm) assembly. All parts meet specifications. Use this as a baseline for comparison.

  Based on the provided reference images of correctly assembled 2-way pneumatic tube connectors (8mm to 6mm), check the test image for any mismatches in component specifications. Determine whether the components in the test image are correctly specified and matched, and identify any issues such as incorrect model numbers, wrong sizes, or mismatched specifications. If a problem is found, identify which part has the incorrect specification.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.34, 0.43] B. Part at [0.51, 0.77] C. Part at [0.55, 0.54] D. Part at [0.42, 0.24]

  Answer: C

WorkVeri PCs scenario ICD

- **User Prompt2:**

  Based on the provided reference images of correctly assembled 2-way pneumatic tube connectors (8mm to 6mm), check the test image for any mismatches in component specifications. Determine whether the components in the test image are correctly specified and matched, and identify any issues such as incorrect model numbers, wrong sizes, or mismatched specifications. If a problem is found, identify which part has the incorrect specification.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.59, 0.66]

  B. Part at [0.41, 0.27]

  C. Part at [0.32, 0.71]

  D. Part at [0.51, 0.44]

WorkVeri PCs scenario ICD

- **Model:** Gemini-2.5-Flash
- GT answers: D
- Pred answer : B
- **Raw Response:**

  The top connector has a noticeably different body profile (more pronounced faceted/stepped center and lacks the same window/slot detailing seen on the other three), indicating it's a different model/spec than the correct 8mm-to-6mm 2-way reducer connectors.

  Answer: B

WORKVERI PCS SCENARIO Zero-Shot

- **User Prompt:**

  Pneumatic Connectors (PC) are specialized components designed for use in pneumatic systems, mainly serving as the interface between air tubes and pneumatic devices. They are widely applied in industrial automation and machinery, including production lines, automated assembly equipment, packaging machines, and various systems powered by compressed air. With the growing demand for efficiency and reliability in modern manufacturing, PCs have become highly popular and indispensable in ensuring stable operations.

  The primary function of a PC is to provide secure and convenient connections for pneumatic tubing, ensuring stable and efficient air transmission within the system. For example, the commonly used 2-way pneumatic tube connector creates a reliable passage between two air tubes, allowing smooth airflow and proper system operation. These connectors feature a compact design, easy installation, and quick disassembly, which significantly improves efficiency during maintenance and replacement. As a result, they play a critical role as essential components in any pneumatic system.

  Based on the provided reference images of correctly assembled 2-way pneumatic tube connectors (8mm to 6mm), check the test image for any mismatches in component specifications. Determine whether the components in the test image are correctly specified and matched, and identify any issues such as incorrect model numbers, wrong sizes, or mismatched specifications. If a problem is found, identify which part has the incorrect specification.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.59, 0.66]

  B. Part at [0.41, 0.27]

  C. Part at [0.32, 0.71]

  D. Part at [0.51, 0.44]

  Answer:

WORKVERI PCS SCENARIO Zero-Shot

- **System Prompt:**

  General task:

  - You will be given a query naming an industrial part type and its expected model/specification.

  - Your job is to detect which part in the provided image deviates from that specification (or from the majority).

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (A, B, C, ...), each describing one candidate part coordinate.

  Input format:

  - A Query: line stating the part type and model/specification.

  - An Image: field containing a single image showing multiple instances of that part.

  - A Choices: Section listing multiple options, each with a normalized coordinate for one part in the image (normalized to [0,1]).

  Output format:

  - Return exactly one letter in the format "Answer: X" at the end of your response.

  - You could include your reasoning before the answer.

WORKVERI PCS SCENARIO Zero-Shot

- **Model:** Gemini-2.5-Flash
- GT answers: D
- Pred answer : A
- **Raw Response 1**:

  Query: 2-way pneumatic tube connectors (8mm to 6mm)

  Image: (Image showing four pneumatic tube connectors)

  Choices:

  A. Part at [0.59, 0.66]

  B. Part at [0.41, 0.27]

  C. Part at [0.32, 0.71]

  D. Part at [0.51, 0.44]

  Analysis:

  The query specifies 2-way pneumatic tube connectors with a size specification of 8mm to 6mm. This means the connector should be a reducing union, connecting an 8mm tube on one side and a 6mm tube on the other.

  I need to examine the four connectors in the image to see if they match this specification. Pneumatic push-to-connect fittings typically have different sizes indicated by the diameter of the tube insertion points. Visually, a reducing union (8mm to 6mm) would have one end visibly larger than the other. A straight union (e.g., 8mm to 8mm or 6mm to 6mm) would have both ends the same size.

  1. **Part B [0.41, 0.27] (Top-left connector):** This connector appears to have two ends of the same size (straight union).

  2. **Part D [0.51, 0.44] (Top-right connector):** This connector also appears to have two ends of the same size (straight union).

  3. **Part A [0.59, 0.66] (Bottom-right connector):** This connector appears to have two ends of the same size (straight union).

  4. **Part C [0.32, 0.71] (Bottom-left connector):** This connector appears to have two ends of the same size (straight union).

  All four parts in the image appear to be straight unions (connecting two tubes of the same diameter, e.g., 8mm to 8mm or 6mm to 6mm).

  The query specifies a **reducing union (8mm to 6mm)**. Since none of the parts show a clear difference in size between the two ends, it is highly likely that all parts are straight unions (e.g., 8mm-8mm or 6mm-6mm) and thus deviate from the required 8mm to 6mm specification.

WORKVERI PCS SCENARIO Zero-Shot

- **Raw Response 2**:

  However, in a multiple-choice question where the goal is to identify *the* mismatched part, and all parts appear to be the same type (straight union) and thus all mismatch the required specification (reducing union), I must look for any subtle difference or assume the question implies finding *any* part that doesn't meet the specification.

  Let's re-examine the parts for any visual cues that might distinguish one from the others, assuming the majority might represent one size (e.g., 8mm-8mm) and one might represent a different size (e.g., 6mm-6mm), or one might actually be the reducing union.

  Upon close inspection, all four connectors look identical in form factor (straight union) and overall size relative to each other. They all appear to be the same model, which is a straight union (equal diameter on both sides).

  Since the query requires an 8mm to 6mm reducing union, and all parts shown are straight unions (equal diameter on both sides), *all* parts are mismatched specifications. In such a scenario, selecting any one of the options is technically correct as they all fail to meet the required specification.

  Let's choose the first option listed, as there is no visual evidence to single out one part as being *more* mismatched than the others.

  If the intent of the question is to find a part that deviates from the *majority* shown in the image, and all parts are visually identical straight unions, then there is no deviation within the image itself.

  If the intent is to find a part that deviates from the *Query specification* (8mm to 6mm reducing union), then all parts deviate. I will select option A as a representative example of the mismatch against the query specification.

  Answer: A

WORKVERI PCS SCENARIO Ref-Cond

- **System Prompt:**

  General task:

  - You will be given one or more reference images showing the correct part type and model/specification, followed by a query and a test image that should match that specification.

  - Your job is to detect which part in the test image deviates from that specification (or from the majority).

  - Instead of returning a coordinate directly, you will be presented with multiple-choice options (A, B, C, ...),

  each corresponding to one candidate part location in the test image.

  Input format:

  - References: one or more images showing correct parts/assemblies that match the stated specification.

  - Query: a single line stating the part type and model/specification to check.

  - Test Image: the image to evaluate (containing multiple instances of the part).

  - Choices: a list of options, each giving a normalized coordinate [x,y] in [0,1] for one part in the test image (top-left is (0,0), bottom-right is (1,1)).

  Output format:

  - Return exactly one letter in the format "Answer: X at the end of your response.

  - You may include brief reasoning before the final answer.

  - Use the reference images only to understand the expected specification; make the final decision based on the Test Image and the provided Choices.

WORKVERI PCS SCENARIO Ref-Cond

- **User Prompt:**

  Pneumatic Connectors (PC) are specialized components designed for use in pneumatic systems, mainly serving as the interface between air tubes and pneumatic devices. They are widely applied in industrial automation and machinery, including production lines, automated assembly equipment, packaging machines, and various systems powered by compressed air. With the growing demand for efficiency and reliability in modern manufacturing, PCs have become highly popular and indispensable in ensuring stable operations.

  The primary function of a PC is to provide secure and convenient connections for pneumatic tubing, ensuring stable and efficient air transmission within the system. For example, the commonly used 2-way pneumatic tube connector creates a reliable passage between two air tubes, allowing smooth airflow and proper system operation. These connectors feature a compact design, easy installation, and quick disassembly, which significantly improves efficiency during maintenance and replacement. As a result, they play a critical role as essential components in any pneumatic system.

  REFERENCE 1 (CORRECT EXAMPLE): There are four correct 2-way pneumatic tube connectors (8mm to 6mm)s. All parts meet specifications. Use this as baseline for comparison.

  REFERENCE 2 (CORRECT EXAMPLE): There are four correct 2-way pneumatic tube connectors (8mm to 6mm)s. All parts meet specifications. Use this as baseline for comparison.

  REFERENCE 3 (CORRECT EXAMPLE): There are four correct 2-way pneumatic tube connectors (8mm to 6mm)s. All parts meet specifications. Use this as baseline for comparison.

  Based on the provided reference images of correctly assembled 2-way pneumatic tube connectors (8mm to 6mm), check the test image for any mismatches in component specifications. Determine whether the components in the test image are correctly specified and matched, and identify any issues such as incorrect model numbers, wrong sizes, or mismatched specifications. If a problem is found, identify which part has the incorrect specification.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.59, 0.66]

  B. Part at [0.41, 0.27]

  C. Part at [0.32, 0.71]

  D. Part at [0.51, 0.44]

WORKVERI PCS SCENARIO Ref-Cond

- GT answers: D
- Pred answer : C
- **Raw Response:**

  Three connectors share the same overall body profile and end/collet sizing consistent with the reference 8mm-to-6mm reducing unions. The lower-left connector has a noticeably different body geometry (more uniform cylindrical Section and different mid-body detailing), indicating a different connector model/spec (likely not the same 8mm to 6mm reducer).

  Answer: C

Examples of CHS SCENARIO

WORKVERI CHS SCENARIO Zero-Shot

- **System Prompt:**

  You are an expert in industrial component quality inspection.

  Metal bolts are fundamental fastening components widely used in mechanical structures, construction projects, transportation equipment, and a variety of industrial manufacturing applications. As essential connecting elements, bolts work together with nuts and washers to provide stability and security between different parts. In modern production, bolts are valued for their strength, durability, and reusability, making them indispensable in ensuring the integrity of complex systems.

  The primary function of a bolt is to create a strong and reliable connection that maintains the strength and stability of the overall structure. For instance, in steel assembly, bolts firmly secure separate plates, preventing loosening or displacement and ensuring long-term load-bearing performance. Their standardized design allows easy installation, quick removal, and efficient maintenance, which greatly improves productivity and reduces downtime. As a result, bolts play a critical role as core components in any application that demands reliable fastening.

  Based on the provided 3V rendered images (showing front, side, and top perspectives) of metal bolts, the task is to compare them with the standard reference images to verify completeness and consistency. The reference images show the expected number of bolts of the same type, all meeting specifications. In the test image, each part is labeled with a letter (A, B, C, D, etc.) and rendered in a distinct color for identification. Your task is to examine the test image and identify any issues such as mismatched components (wrong model number, incorrect dimensions, or different workpiece type). If a problem is identified, return only the letter corresponding to the faulty part (e.g., "B" or "D")."

WORKVERI CHS SCENARIO Zero-Shot

- **User Prompt:** Examine the provided three-view rendered image (front, side, and top views merged). Each part is labeled with a letter (A, B, C, D, etc.) and rendered in a distinct color. Identify which part, if any, has the wrong model number or specification compared to the others. Return only the letter corresponding to the mismatched part.

  Answer:

WORKVERI CHS SCENARIO Zero-Shot

- **Model:** Qwen3-VL-235B
- **Gt answers**: D
- **Pred answer**: D
- **Raw Response:**

  So, let's analyze the three views. First, check each part's consistency across views.

  Looking at part D: In the front view, D is a light blue bolt overlapping A. In the side view, D is a light blue bolt but the head shape seems different. Wait, in the top view, D is a light blue shape that looks like a different type—maybe the head is not matching. Let's check the other parts. A, B, C have consistent head shapes (like hexagonal or standard bolt heads), but D in the top view has a more rounded or different head. Wait, in the front view, D is a bolt, but in the top view, D's shape is different. Let's check the side view: D is a light blue bolt, but maybe the length or head type is wrong. Wait, the key is that all bolts should be the same type. Let's see: A, B, C have the same head style (the top view for A, B, C shows the head as a certain shape). D in the top view is a different shape—maybe it's a different model. Let's confirm:

  - A: In top view, it's a blue shape with a hole (hex head), side view shows the bolt, front view same.

  - B: Red, top view is a red shape with a rounded head (maybe hex), side view, front view.

  - C: Pink, top view is a pink shape, side and front.

  - D: Light blue. In top view, D's shape is different—it looks like a different type of fastener, maybe a screw with a different head. Wait, the other bolts (A, B, C) have heads that in top view are more like hexagonal (A's top view is a hex with a hole), B's top view is a hex, C's top view is a hex. D's top view is a more oval or different shape, which doesn't match. So D is the mismatched part.

E.2 Examples of SURFINSP

Below are examples of SURFINSP (Structural Surface Inspection) evaluation data. The task requires the model to classify a workpiece's surface condition into one of five categories: *crack*, *cut*, *deformation*, *dent*, or *good*. We first list the workpiece

type descriptions provided to the model, followed by a complete prompt–response example.

Task 2 The Introduction of workpieces.

- **Corner Bracket:** A corner bracket is an L-shaped metal fitting used to reinforce right-angle joints. Its function is to securely connect two perpendicular components, ensuring structural stability. It is widely used in furniture assembly, building frameworks, shelving, and machinery support.
- **Countersunk Screw:** A countersunk screw has a conical head that sits flush with the material surface. Its function is to provide a smooth, flat appearance while maintaining strong fastening. It is commonly applied in mechanical assemblies, electronics, and furniture manufacturing.
- **Cup Head Screw:** A cup head screw features a rounded, smooth head and often includes a square neck to prevent rotation. Its function is to hold wooden or metal parts together securely without turning in the joint. It is widely used in woodworking, construction, and heavy machinery connections.
- **Eye Bolt:** An eye bolt has a looped head designed for attaching ropes, cables, or hooks. Its function is to serve as a lifting or anchoring point in load-bearing operations. It is commonly found in lifting equipment, construction, marine, and transportation industries.
- **Flat Washer:** A flat washer is a thin, flat circular disk with a central hole. Its function is to distribute the load of a bolt or nut and protect the connected surface from damage. It is used in mechanical assemblies, construction, and electronic devices.
- **Hex Nut:** A hex nut is a six-sided fastener commonly paired with bolts. Its function is to secure and clamp parts together through threaded engagement. It is widely used in machinery, automotive manufacturing, construction, and furniture.
- **Rivet Nut:** A rivet nut is a threaded insert designed for installation in thin materials. Its function is to provide a strong internal thread where tapping is impossible. It is often used in automotive, aerospace, sheet metal, and electronic enclosures.

Task 2 The Introduction of workpieces.

- **Self-tapping Screw:** A self-tapping screw can cut its own threads into materials without pre-drilling. Its function is to enable fast and easy assembly without requiring separate threading. It is commonly used in metal sheets, plastics, and wood structures.
- **Spring Washer:** A spring washer is a split ring that provides elastic tension under compression. Its function is to prevent loosening of nuts and maintain constant preload. It is widely applied in machinery, automotive, and electrical equipment.
- **T Bolt Half-thread Screw:** A T-bolt half-thread screw has a T-shaped head that fits into profile slots for positioning. Its function is to provide adjustable and secure fastening in modular assemblies. It is typically used in machine frames, automation systems, and aluminum profile structures.
- **T Nut:** A T-nut is designed to fit into T-slots, providing a threaded anchor point. Its function is to enable adjustable fastening within T-slot profiles. It is commonly used in machine tables, aluminum extrusion frames, and assembly lines.
- **T Screw:** A T-screw has a T-shaped head allowing it to slide into slots for secure fixation. Its function is to work with T-nuts for quick assembly and position adjustment. It is often used in industrial equipment, fixtures, and aluminum frame systems.
- **Wing Nut:** A wing nut has two projecting "wings" that allow manual tightening without tools. Its function is to enable quick installation and removal by hand. It is frequently used in devices or setups that require frequent adjustments, such as lighting and lab equipment.
- **Wing Screw:** A wing screw has a head shaped like wings for easy hand operation. Its function is to allow fast manual fastening or adjustment without tools. It is commonly applied in instruments, supports, and lightweight mechanical assemblies.

Examples of SURFINSP Cup Head Screw

SURFINSP Cup Head Screw

- **System Prompt:**

  You are an expert in manufacturing defect analysis based on 3V rendered images of parts.

  Cup Head Screw: A cup head screw features a rounded, smooth head and often includes a square neck to prevent rotation. Its function is to hold wooden or metal parts together securely without turning in the joint. It is widely used in woodworking, construction, and heavy machinery connections.

  Given a 3V image of a manufactured part, your task is to answer two questions:

  Part 1: Is the part normal?

  Answer: Yes or No

  Part 2: If the answer to Part 1 is No, identify the defect type by selecting one letter:

  A) Crack - a material fracture or micro-crack caused by stress concentration or brittleness

  B) Cut - surface scratch or gouge caused by improper machining or external tool impact

  C) Deformation - geometric distortion or warping due to excessive load, heat treatment, or clamping stress

  D) Dent - local surface depression caused by impact or overpressure during handling or clamping

  Important: There is at most one type of abnormality present in each part.

  Answer format:

  Part 1: [Yes/No]

  Part 2: [A/B/C/D] (only if Part 1 is No)

  If uncertain, choose the option that best matches the dominant feature.

SURFINSP Cup Head Screw

- **User Prompt:** Here are 3 normal reference images with no defects. Use these to understand what normal (good) parts look like when identifying abnormalities:

  NORMAL REFERENCE 1 (No Defect)

  NORMAL REFERENCE 2 (No Defect)

  NORMAL REFERENCE 3 (No Defect)

  Analyze the given part:

  Part 1: Is it normal? (Yes/No)

  Part 2: If No, select the defect type:

  A) Crack

  B) Cut

  C) Deformation

  D) Dent

- **Setting:** Ref-Cond
- **Raw response:**

  Part 1: No

  Part 2: B

E.3 Examples of ASSYVERI

Below are examples of ASSYVERI evaluation data covering both image and point cloud modalities. The task requires the model to detect extra, unnecessary, or incorrectly specified parts in an assembly. We show examples from the MES SCENARIO and CNC SCENARIO image scenarios as well as 3D point cloud input.

Examples of MES SCENARIO

AssyVeri MES scenario ICD

- **System Prompt:**

  General task:

  - You will be given one or several reference images of correct assemblies, followed by a query and a test image expected to be the same assembly.

  - Your job is to identify which part in the provided test image uses the wrong model number (incorrect size or shape) and thus deviates from the rest.

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (like A, B, C), each describing one candidate part location.

  - Choose the letter(s) corresponding to the mismatched part(s).

  Input format:

  - References: one or more images showing correct assemblies.

  - Query: a single line stating the task.

  - Test Image: the image to evaluate.

  - Choices: a list of options, each with a normalized coordinate for one part in the test image (normalized to [0,1]).

  Output format:

  - Return exactly the letter of the mismatched part at the end of the answer in a format start with "Answer:" and then letter choice like Answer: C.

  - Also you can return your reasoning based on the provided reference images but be sure to give the final answer only based on the test image.

AssyVeri MES scenario ICD

- **User Prompt 1:**

  Based on the provided reference images of correctly assembled Metal expansion bolts, check the test image for any extra, unnecessary, or incorrectly specified parts.

  A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.35, 0.53]

  B. Part at [0.52, 0.40]

  C. Part at [0.48, 0.24]

  D. Part at [0.65, 0.18]

  E. Part at [0.68, 0.40]

  Answer: A

  Based on the provided reference images of correctly assembled Metal expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.48, 0.53]

  B. Part at [0.53, 0.64]

  C. Part at [0.50, 0.42]

  D. Part at [0.43, 0.32]

  E. Part at [0.66, 0.56]

- **User Prompt 2:**

  Based on the provided reference images of correctly assembled Metal expansion bolts, check the test image for any extra, unnecessary, or incorrectly specified parts.

  A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.35, 0.53]

  B. Part at [0.52, 0.40]

  C. Part at [0.48, 0.24]

  D. Part at [0.65, 0.18]

  E. Part at [0.68, 0.40]

  Answer: A

  Based on the provided reference images of correctly assembled Metal expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.48, 0.53]

  B. Part at [0.53, 0.64]

  C. Part at [0.50, 0.42]

  D. Part at [0.43, 0.32]

  E. Part at [0.66, 0.56]

- **Model:** Gemini-2.5-Flash
- **Gt answers**: C
- **Pred answer**: D
- **Raw Response:** The standard Metal expansion bolt assembly, as shown in the reference images, consists of five components: Bolt, Nut, Flat Washer, Spring Washer, and Sleeve.

  The test image contains the following parts:

  1. **Bolt** (Hex socket cap screw type, different from the references but still a bolt) at approximately [0.43, 0.32] (D).

  2. **Nut** at approximately [0.43, 0.32] (D) - Wait, D is the bolt. The nut is at [0.48, 0.24] in the previous example, or here at [0.48, 0.53] (A).

  3. **Sleeve** at approximately [0.50, 0.42] (C).

  4. **Flat Washer** (large) at approximately [0.53, 0.64] (B).

  5. **Small Washer** (likely a Spring Washer or a second Flat Washer, but much smaller than the others) at approximately [0.66, 0.56] (E).

  Let's re-examine the components based on the standard assembly and the reference images:

  * **Reference Images Components:** Bolt (with a conical end), Nut, Sleeve, Flat Washer (large), and Spring Washer (small, split ring).

  * **Test Image Components:**

  * A. Nut (Correct size relative to the bolt threads). * B. Flat Washer (Large, correct size).

  * C. Sleeve (Correct size).

  * D. Bolt (Incorrect type/head shape Hex socket cap screw head instead of the conical head shown in the references, which is crucial for the expansion mechanism).

  * E. Small Washer (This appears to be a small flat washer, not a spring washer, and its inner diameter seems too small for the bolt/sleeve, or it might be a small flat washer intended to be the spring washer replacement, but it is visually different from the spring washers in the references).

  Comparing the Bolt (D) in the test image to the Bolts in the reference images:

  The reference bolts have a conical or rounded end that engages with the sleeve to cause expansion. The bolt in the test image (D) is a standard hex socket cap screw, which has a flat end and a cylindrical head. This type of bolt head and end shape is incorrect for a standard expansion bolt mechanism where the bolt head/end is designed to push into the sleeve to expand it.

  Therefore, the Bolt (D) is the mismatched part due to its incorrect model/shape (Hex socket cap screw instead of the required expansion bolt type).

  Answer: D

ASSYVERI MES SCENARIO Zero-Shot

- **User Prompt:**

  Metal Expansion Bolt (MEB) is one of the most frequently applied workpieces. Its primary function is to ensure strong anchoring by expanding within the drilled hole of the base material, thereby generating high frictional resistance and mechanical interlock.

  A typical Metal Expansion Bolt consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Together, these components form a complete fastening system that offers durability, reliability, and ease of installation in demanding construction and industrial environments.

  Based on the provided reference images of correctly assembled Metal expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.48, 0.53]

  B. Part at [0.53, 0.64]

  C. Part at [0.50, 0.42]

  D. Part at [0.43, 0.32]

  E. Part at [0.66, 0.56]

  Answer:

ASSYVERI MES SCENARIO Zero-Shot

- **System Prompt:**

  General task:

  - You will be given a single image of an assembly containing multiple industrial parts that match each other well.

  - Your job is to identify which part in the provided test image uses the wrong model number (incorrect size or shape) and thus deviates from the rest.

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (A, B, C, ...), each describing one candidate part location.

  - Choose the letter corresponding to the mismatched part.

  Input format:

  - Query: a single line stating the task.

  - Image: the image to evaluate.

  - Choices: a list of options, each with a normalized coordinate for one part in the image (normalized to [0,1]).

  Output format:

  - Return exactly one letter in the format "Answer: X" at the end of your response.

  - You could include your reasoning before the answer.

- **Model:** Gemini-2.5-Flash
- **Gt answers**: C
- **Pred answer**: E

ASSYVERI MES SCENARIO Zero-Shot

- **Raw Response:**

  The standard assembly for a Metal expansion bolt requires five components: Bolt, Nut, Flat Washer, Spring Washer, and Sleeve.

  Let's identify the parts present in the image:

  1. **Bolt**: The threaded component with a cylindrical head (Socket Cap Screw style). (Approx. [0.50, 0.42]) - Matches C.

  2. **Nut**: The hexagonal threaded fastener. (Approx. [0.48, 0.53]) - Matches A.

  3. **Sleeve**: The long cylindrical expansion tube. (Approx. [0.53, 0.64]) - Matches B.

  4. **Flat Washer**: The large, solid ring. (Approx. [0.43, 0.32]) - Matches D.

  5. **Small Washer**: A smaller, solid ring. (Approx. [0.66, 0.56]) - Matches E.

  The required components are: Bolt, Nut, Flat Washer, Spring Washer, and Sleeve.

  The components present are:

  * Bolt (C)

  * Nut (A)

  * Sleeve (B)

  * Flat Washer (D)

  * Small Washer (E)

  Comparing the required list to the present list:

  * The assembly is missing a **Spring Washer**.

  * The assembly contains two flat washers (D and E), but one of them (E) is significantly smaller than the other (D) and appears to be the wrong size or an extra, unnecessary component, as the standard assembly only requires one Flat Washer and one Spring Washer.

  However, the task asks to identify the part that uses the wrong model number (incorrect size or shape) or is extra/unnecessary.

  If we assume the larger washer (D) is the correct Flat Washer, then the smaller washer (E) is an extra component and is also likely the wrong size for the bolt/sleeve combination. Furthermore, the required Spring Washer is missing.

  In the context of identifying a mismatched part based on size/specification or being extra:

  * Part E ([0.66, 0.56]) is a small flat washer. Given that a standard assembly requires one flat washer (D) and one spring washer (missing), this second, smaller flat washer (E) is an extra, unnecessary component and is likely the wrong size for the bolt/nut combination if it were intended to be the primary washer.

  Therefore, the small washer at [0.66, 0.56] is the problematic part.

  Answer: E

ASSYVERI MES SCENARIO Ref-Cond

- **System prompt:** The same with ASSYVERI Metal Expansion Bolt (MEB) Image In-Context Demonstration.

- **User prompt:**

  Metal Expansion Bolt (MEB) is one of the most frequently applied products. Its primary function is to ensure strong anchoring by expanding within the drilled hole of the base material, thereby generating high frictional resistance and mechanical interlock.

  A typical Metal Expansion Bolt consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Together, these components form a complete fastening system that offers durability, reliability, and ease of installation in demanding construction and industrial environments

  Reference 1: correct Metal expansion bolt assembly.

  Reference 2: correct Metal expansion bolt assembly.

  Reference 3: correct Metal expansion bolt assembly.

  Based on the provided reference images of correctly assembled Metal expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly consists of five key components, each contributing to its performance and reliability: (1)Bolt - the main fastening element that provides tensile strength; (2)Nut - used to tighten and secure the bolt in place; (3)Flat Washer - distributes the load evenly across the surface, reducing the risk of damage to the base material; (4)Spring Washer - prevents loosening caused by vibrations or dynamic loads; (5)Sleeve - the expansion element that deforms outward under tightening, ensuring strong anchorage in the drilled hole. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick one choice corresponding to the mismatched part.

  A. Part at [0.48, 0.53]

  B. Part at [0.53, 0.64]

  C. Part at [0.50, 0.42]

  D. Part at [0.43, 0.32]

  E. Part at [0.66, 0.56]

AssyVeri MES scenario Ref-Cond

- **Model:** Gemini-2.5-Flash
- **Gt answers**: C
- **Pred answer**: C
- **Raw Response:**

  In the reference assemblies, the **bolt** is a countersunk-style expansion bolt with a tapered head and the characteristic shank shape for driving the sleeve. In the test image, the bolt at **[0.50, 0.42]** is a different type (looks like a standard socket-head/cap screw), so it's the wrong model/shape for this expansion bolt set.

  Answer: C

Examples of PES scenario

AssyVeri PES scenario ICD

- **System Prompt:**

  General task:

  - You will be given one or several reference images of correct assemblies, followed by a query and a test image expected to be the same assembly.

  - Your job is to identify which part in the provided test image uses the wrong model number (incorrect size or shape) and thus deviates from the rest.

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (like A, B, C), each describing one candidate part location.

  - Choose the letter(s) corresponding to the mismatched part(s).

  Input format:

  - References: one or more images showing correct assemblies.

  - Query: a single line stating the task.

  - Test Image: the image to evaluate.

  - Choices: a list of options, each with a normalized coordinate for one part in the test image (normalized to [0,1]).

  Output format:

  - Return exactly the letter of the mismatched part at the end of the answer in a format start with "Answer:" and then letter choice like Answer: C.

  - Also you can return your reasoning based on the provided reference images but be sure to give the final answer only based on the test image.

AssyVeri PES scenario ICD

- **User Prompt:** Plastic Expansion Bolts (PEB) stand out as one of the most commonly used variants. Their primary function is to provide a simple yet effective way to anchor screws into walls, panels, or other base materials where direct fastening is not feasible. By expanding inside the drilled hole as the screw is driven in, PEBs create a tight grip that prevents loosening, enhances load-bearing capacity, and ensures durability of the connection.

  The standard assembly must contain three plastic expansion screws, each consisting of two parts: (1) plastic anchors, (2) metal screw.

  Reference 1: Correct Plastic expansion bolt assembly. Reference 2: Correct Plastic expansion bolt assembly. Reference 3: Correct Plastic expansion bolt assembly.

  Based on the provided reference images of correctly assembled Plastic expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly must contain three plastic expansion screws, each consisting of two parts: (1) plastic anchors, (2) metal screw. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.50, 0.53] B. Part at [0.40, 0.40] C. Part at [0.60, 0.52] D. Part at [0.60, 0.43] E. Part at [0.35, 0.63] F. Part at [0.50, 0.64]

  Answer: F

  Based on the provided reference images of correctly assembled Plastic expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly must contain three plastic expansion screws, each consisting of two parts: (1) plastic anchors, (2) metal screw. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.40, 0.52]

  B. Part at [0.43, 0.70]

  C. Part at [0.50, 0.44]

  D. Part at [0.47, 0.39]

  E. Part at [0.65, 0.56]

  F. Part at [0.42, 0.62]

  Answer:

AssyVeri PES scenario ICD

- **Model:** Gemini-2.5-Flash
- **Gt answers**: F
- **Pred answer**: F
- **Raw Response1:**

  The test image shows three plastic anchors and four metal screws. A standard assembly, as indicated by the references, should consist of three complete sets, meaning three plastic anchors and three metal screws.

  The parts in the test image are: 1. Plastic anchor (Type 1, ribbed) at [0.35, 0.63] (E in the previous set, B in the current set) 2. Plastic anchor (Type 1, ribbed) at [0.50, 0.53] (A in the previous set, C in the current set) 3. Plastic anchor (Type 2, stepped/split) at [0.60, 0.52] (C in the previous set, E in the current set) 4. Metal screw at [0.40, 0.40] (B in the previous set, D in the current set) 5. Metal screw at [0.60, 0.43] (D in the previous set, A in the current set) 6. Metal screw at [0.50, 0.64] (F in the previous set, F in the current set) 7. Metal screw at [0.42, 0.62] (F in the current set, this is the same screw as F in the previous set, but the coordinates are slightly different; let's re-evaluate the coordinates for the screws)

  Let's use the coordinates provided in the current prompt:

  A. Part at [0.40, 0.52] (Metal screw) B. Part at [0.43, 0.70] (Plastic anchor, ribbed) C. Part at [0.50, 0.44] (Plastic anchor, ribbed) D. Part at [0.47, 0.39] (Metal screw) E. Part at [0.65, 0.56] (Plastic anchor, stepped/split) F. Part at [0.42, 0.62] (Metal screw)

  Wait, there are four screws visible in the image. Let's identify all parts and their types:

  1. Plastic Anchor (ribbed, bottom left): [0.43, 0.70] (B) 2. Plastic Anchor (ribbed, top center): [0.50, 0.44] (C) 3. Plastic Anchor (stepped/split, right): [0.65, 0.56] (E) 4. Metal Screw (top right): [0.60, 0.43] (Not listed, but close to A) 5. Metal Screw (top center): [0.47, 0.39] (D) 6. Metal Screw (bottom center): [0.40, 0.52] (A) 7. Metal Screw (bottom left): [0.42, 0.62] (F)

  There are 3 plastic anchors (B, C, E) and 4 metal screws (A, D, F, and one unlisted). Since the standard assembly requires 3 plastic anchors and 3 metal screws, one of the metal screws is extra/unnecessary.

AssyVeri PES scenario ICD

- **Model:** Gemini-2.5-Flash
- **Gt answers**: F
- **Pred answer**: F
- **Raw Response2:**

  Let's re-examine the coordinates and the image to confirm the parts listed in the choices:

  A. Part at [0.40, 0.52]: Metal screw (bottom center) B. Part at [0.43, 0.70]: Plastic anchor (bottom left) C. Part at [0.50, 0.44]: Plastic anchor (top center) D. Part at [0.47, 0.39]: Metal screw (top center) E. Part at [0.65, 0.56]: Plastic anchor (right) F. Part at [0.42, 0.62]: Metal screw (bottom left, partially overlapping with B)

  The image contains 3 plastic anchors and 4 metal screws. Since the expected assembly is 3 sets (3 anchors + 3 screws), one of the metal screws is the mismatched part (extra/unnecessary). All options A, D, and F are metal screws. The fourth screw is located around [0.60, 0.43] (top right).

  Since the task is to identify *which part* uses the wrong model number or is extra/unnecessary, and there are 4 screws when only 3 are needed, any of the screws could be considered the extra part if they are all identical. Assuming all screws are the same model, we must choose one of the screw options provided in the choices.

  In the previous attempt, the answer was F, which corresponds to the metal screw at [0.42, 0.62]. This screw is indeed one of the four screws present, making it an extra component if the assembly is meant to be 3 sets.

  Answer: F

ASSYVERI PES SCENARIO Zero-Shot

- **System Prompt:**

  General task:

  - You will be given a single image of an assembly containing multiple industrial parts that match each other well.

  - Your job is to identify which part in the provided test image uses the wrong model number (incorrect size or shape) and thus deviates from the rest.

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (A, B, C, ...), each describing one candidate part location.

  - Choose the letter corresponding to the mismatched part.

  Input format:

  - Query: a single line stating the task.

  - Image: the image to evaluate.

  - Choices: a list of options, each with a normalized coordinate for one part in the image (normalized to [0,1]).

  Output format:

  - Return exactly one letter in the format "Answer: X" at the end of your response.

  - You could include your reasoning before the answer.

ASSYVERI PES SCENARIO Zero-Shot

- **User Prompt:**

  Plastic Expansion Bolts (PEB) stand out as one of the most commonly used variants. Their primary function is to provide a simple yet effective way to anchor screws into walls, panels, or other base materials where direct fastening is not feasible. By expanding inside the drilled hole as the screw is driven in, PEBs create a tight grip that prevents loosening, enhances load-bearing capacity, and ensures durability of the connection.

  The standard assembly must contain three plastic expansion screws, each consisting of two parts: (1) plastic anchors, (2) metal screw.

  Based on the provided reference images of correctly assembled Plastic expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly must contain three plastic expansion screws, each consisting of two parts: (1) plastic anchors, (2) metal screw. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.40, 0.52]

  B. Part at [0.43, 0.70]

  C. Part at [0.50, 0.44]

  D. Part at [0.47, 0.39]

  E. Part at [0.65, 0.56]

  F. Part at [0.42, 0.62]

  Answer:

- **Model:** Gemini-2.5-Flash
- **Gt answers**: F
- **Pred answer**: E

AssyVeri PES scenario Zero-Shot

- **Raw Response:**

  The image shows an assembly of plastic expansion bolts, which should consist of three plastic anchors and three metal screws.

  Analysis of Components:

  1. **Plastic Anchors (Yellow):** There are three plastic anchors present.

  * Anchor 1 (Top-center): Appears to be a standard ribbed anchor.

  * Anchor 2 (Bottom-left): Appears to be a standard ribbed anchor.

  * Anchor 3 (Right-center, [0.65, 0.56]): This anchor has a distinct, square-wave or zig-zag pattern along its body, which is different in shape/design from the other two anchors.

  2. **Metal Screws (Silver):** There are three metal screws present. All three screws appear to be standard wood/self-tapping screws of similar size and shape.

  Identifying the Mismatched Part:

  The task is to identify the part that deviates from the rest or is incorrectly specified. The plastic anchor at the right-center position has a different design (square-wave pattern) compared to the other two plastic anchors (ribbed pattern). This difference in model/shape makes it the deviating component.

  The candidate options for the plastic anchors are:

  A. Part at [0.40, 0.52] (Metal screw) - Incorrect type

  B. Part at [0.43, 0.70] (Metal screw) - Incorrect type

  C. Part at [0.50, 0.44] (Plastic anchor, top-center) - Matches Anchor 2

  D. Part at [0.47, 0.39] (Metal screw) - Incorrect type

  E. Part at [0.65, 0.56] (Plastic anchor, right-center) - **Mismatched shape/model**

  F. Part at [0.42, 0.62] (Plastic anchor, bottom-left) - Matches Anchor 1

  The plastic anchor at [0.65, 0.56] (Option E) is the one with the different design compared to the other two plastic anchors.

  Answer: E

AssyVeri PES scenario Ref-Cond

- **System prompt:** The same with AssyVeri PCs scenario ICD.

- **User prompt:**

  Plastic Expansion Bolts (PEB) stand out as one of the most commonly used variants. Their primary function is to provide a simple yet effective way to anchor screws into walls, panels, or other base materials where direct fastening is not feasible. By expanding inside the drilled hole as the screw is driven in, PEBs create a tight grip that prevents loosening, enhances load-bearing capacity, and ensures durability of the connection.

  The standard assembly must contain three plastic expansion screws, each consisting of two parts:

  (1) plastic anchors,

  (2) metal screw

  Reference 1: Correct plastic expansion bolt assembly.

  Reference 2: Correct plastic expansion bolt assembly.

  Reference 3: Correct plastic expansion bolt assembly.

  Based on the provided reference images of correctly assembled Plastic expansion bolt, check the test image for any extra, unnecessary, or incorrectly specified parts. A standard assembly must contain three plastic expansion screws, each consisting of two parts:

  (1) plastic anchors,

  (2) metal screw.

  Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick one choice corresponding to the mismatched part.

  A. Part at [0.40, 0.52]

  B. Part at [0.43, 0.70]

  C. Part at [0.50, 0.44]

  D. Part at [0.47, 0.39]

  E. Part at [0.65, 0.56]

  F. Part at [0.42, 0.62]

AssyVeri PES scenario Ref-Cond

- **Model:** Gemini-2.5-Flash
- **Gt answers**: F
- **Pred answer**: F
- **Raw Response:**

  The test image contains three metal screws that look consistent with the reference (same general length and countersunk head style).

  However, among the three plastic anchors, the one at **[0.42, 0.62]** differs in shape/size (it has a noticeably different body profile and proportions compared to the other two anchors that match the reference style), indicating it's the wrong model.

  Answer: F

Examples of CNC scenario

AssyVeri CNC scenario ICD

- **System Prompt:**

  General task:

  - You will be given one or several reference images of correct assemblies, followed by a query and a test image expected to be the same assembly.

  - Your job is to identify which part in the provided test image uses the wrong model number (incorrect size or shape) and thus deviates from the rest.

  - Instead of returning a coordinate directly, you will be presented with a list of multiple-choice options (like A, B, C), each describing one candidate part location.

  - Choose the letter(s) corresponding to the mismatched part(s).

  Input format:- References: one or more images showing correct assemblies.

  - Query: a single line stating the task.

  - Test Image: the image to evaluate.

  - Choices: a list of options, each with a normalized coordinate for one part in the test image (normalized to [0,1]).

  Output format:

  - Return exactly the letter of the mismatched part at the end of the answer in a format start with "Answer:" and then letter choice like Answer: C.

  - Also you can return your reasoning based on the provided reference images but be sure to give the final answer only based on the test image.

AssyVeri CNC scenario ICD

- **User Prompt:**

  In CNC machining, the primary role of fixtures is to ensure that the workpiece is accurately positioned and firmly secured during various cutting, milling, or drilling operations.

  A typical CNC fixture is composed of several fundamental elements that work together to deliver reliable clamping performance. These include: (1) the flange nut, (2) the T-bolt half-thread screw, (3) the pressure plate, and (4) the triangular wedge.

  Reference 1: Correct Computer Numerical Control (CNC) fixture assembly.

  Reference 2: Correct Computer Numerical Control (CNC) fixture assembly.

  Reference 3: Correct Computer Numerical Control (CNC) fixture assembly.

  Based on the provided reference images of correctly assembled Computer Numerical Control (CNC) fixture, check the test image for any extra, unnecessary, or incorrectly specified parts. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick the letter corresponding to the mismatched part.

  A. Part at [0.24, 0.50]

  B. Part at [0.59, 0.60]

  C. Part at [0.47, 0.83]

  D. Part at [0.33, 0.36]

  Answer: C

  Based on the provided reference images of correctly assembled Computer Numerical Control (CNC) fixture, check the test image for any extra, unnecessary, or incorrectly specified parts. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick one choice corresponding to the mismatched part.

  A. Part at [0.70, 0.44]

  B. Part at [0.69, 0.81]

  C. Part at [0.45, 0.40]

  D. Part at [0.29, 0.58]

AssyVeri CNC scenario ICD

- **Model:** Gemini-2.5-Flash
- **Gt answers**: D
- **Pred answer**: B
- **Raw Response1:**

  The reference images show a CNC fixture assembly typically consisting of a T-slot bolt, a clamping strap (with a slot), a serrated block (or similar work support/grip), and a nut.

  Let's examine the parts in the test image and compare them to the references:

  1. **Part A (Clamping Strap):** Located at [0.70, 0.44]. This is a metal strap with a central slot. This part appears consistent in shape and function with the clamping strap seen in all reference images.

  2. **Part C (T-slot Bolt):** Located at [0.45, 0.40]. This is a bolt with a T-shaped head and threaded shaft. This part appears consistent with the bolt seen in all reference images.

  3. **Part D (Serrated Block/Work Support):** Located at [0.29, 0.58]. This is a rectangular block with serrations on one face and a hole in the center. This part is consistent with the serrated block/work support seen in Reference 1 and Reference 3 (though Reference 2 and 4 show a different, triangular serrated part, the rectangular serrated block is a valid component in this type of kit, as shown in References 1 and 3).

  4. **Part B (Nut):** Located at [0.69, 0.81]. This is a small hexagonal nut.

AssyVeri CNC scenario ICD

- **Raw Response2:**

  Now, let's look closely at the nut (Part B) in the test image and compare it to the nuts in the reference images.

  * In Reference 1, 2, 3, and 4, the nut is a standard hexagonal nut, often a flange nut or a standard hex nut, designed to thread onto the bolt.

  * In the Test Image, the part at [0.69, 0.81] (Part B) is a small, low-profile hexagonal piece. It appears to be a **T-nut** or a similar specialized nut, but its size and shape relative to the bolt (Part C) suggest it might be too small or the wrong type of nut for this specific assembly, especially when compared to the larger, more robust nuts shown in the references (e.g., Reference 2 and 4 show a clearly larger nut).

  However, upon closer inspection of the test image, the part at [0.69, 0.81] (Part B) is a **T-nut** (or a similar specialized nut) which is typically used to slide into a T-slot on a machine table. This type of nut is not typically used as the clamping nut *on top* of the clamping strap in this assembly; a standard hex nut or flange nut is used there (as seen in the references).

  Let's re-examine the parts in the first set of images provided in the prompt (which are the same as the second set, just different crops/views).

  In the test image:

  A. Part at [0.70, 0.44] - Clamping Strap (Correct)

  B. Part at [0.69, 0.81] - Nut (Mismatched type/size compared to references)

  C. Part at [0.45, 0.40] - T-slot Bolt (Correct)

  D. Part at [0.29, 0.58] - Serrated Block (Correct type, seen in References 1 and 3)

  The nut (Part B) is significantly smaller and appears to be a different type (possibly a T-nut or a very small hex nut) than the nuts shown in the reference images, which are larger and more standard hex nuts for clamping. Therefore, Part B is the mismatched component.

  Answer: B

ASSYVERI CNC SCENARIO Ref-Cond

- **System prompt:** The same with ASSYVERI Computer Numerical Control (CNC) fixture Image In-Context Demonstration.
- **User prompt:**

  In CNC machining, the primary role of fixtures is to ensure that the workpiece is accurately positioned and firmly secured during various cutting, milling, or drilling operations.

  A typical CNC fixture is composed of several fundamental elements that work together to deliver reliable clamping performance. These include:

  (1) the flange nut,

  (2) the T-bolt half-thread screw,

  (3) the pressure plate, and

  (4) the triangular wedge.

  Reference 1: correct Computer Numerical Control (CNC) fixture assembly.

  Reference 2: correct Computer Numerical Control (CNC) fixture assembly.

  Reference 3: correct Computer Numerical Control (CNC) fixture assembly.

  Based on the provided reference images of correctly assembled Computer Numerical Control (CNC) fixture, check the test image for any extra, unnecessary, or incorrectly specified parts. Determine whether the test image contains any components that should not be present or have incorrect specifications. Identify any such problematic parts.

  Below are candidate parts with their normalized coordinates in [0,1].

  Pick one choice corresponding to the mismatched part.

  A. Part at [0.70, 0.44]

  B. Part at [0.69, 0.81]

  C. Part at [0.45, 0.40]

  D. Part at [0.29, 0.58]

ASSYVERI CNC SCENARIO Ref-Cond

- **Raw Response:**

  To determine the mismatched part, we compare the test image with the reference images of correct CNC fixture assemblies. The references consistently show four key components: a rectangular slotted piece, a bolt, a circular nut with a hexagonal interior, and a rectangular serrated block.

  - **Option A** ([0.70, 0.44]): This corresponds to the rectangular slotted piece, which matches the shape and size seen in all references.

  - **Option C** ([0.45, 0.40]): This is the bolt, which aligns with the standard bolt shape and size in the references.

  - **Option D** ([0.29, 0.58]): This is the rectangular serrated block, which matches the serrated edge and overall form in references 1.

  - **Option B** ([0.69, 0.81]): This is the small hex nut. In the references, the nut is consistently larger and has a circular shape with a hexagonal interior (not a small hex nut). The size and shape of this part deviate from the correct specification.

  Thus, the mismatched part is **B**, as it does not match the expected size or shape of the nut in the reference assemblies.

  Answer: B

Examples of SWN SCENARIO

ASSYVERI SWN SCENARIO ICD

- **System Prompt:**

  You are an expert in industrial assembly quality inspection.

  Manufacturing plays a crucial role in modern society, supporting everything from everyday consumer products to advanced industrial machinery. None of these achievements would be possible without precise manufacturing processes, where a wide range of workpieces form the backbone of production. These workpieces are fundamental not only to the structural integrity of equipment but also to ensuring safe, stable, and long-term functionality. From household appliances to large-scale mechanical systems, such components are essential in holding structures together and enabling smooth operation across diverse applications.

  Among the most widely used workpieces are bolts and nuts. A bolt engages with a nut to create fastening, which provides strong connections and stability between components. However, bolts and nuts alone are often insufficient for reliable assembly. Complementary parts such as washers are also required. Flat washers help distribute pressure evenly and protect contact surfaces, while spring washers are designed to prevent loosening during operation, thereby improving durability and safety. By combining these elements, mechanical assemblies achieve higher precision and resilience.

  In practical manufacturing and assembly, bolts, nuts, and washers are commonly paired to form standardized fastening sets. The most typical combination includes: (1) bolt, (2) nut, (3) flat washer, (4) spring washer. This combination ensures that fastening systems remain secure and consistent across a wide range of industrial and everyday applications.

  Based on the provided 3V rendered reference images (showing front, side, and top perspectives) of a correctly assembled common manufacturing workpiece combination, check the test image for any mismatches in component configuration. A standard assembly must contain four M18 components: (1) M18 bolt, (2) M18 nut, (3) M18 flat washer, (4) M18 spring washer. In the test image, each part is labeled with a letter (A, B, C, ..., E, etc.) and rendered in a distinct color for identification. Your task is to determine whether the components are complete and correctly matched, and identify any issues such as missing specifications, incorrect components, or extra parts. If a problem is found, return only the letter corresponding to the faulty part (e.g., "B" or "E").

ASSYVERI SWN SCENARIO ICD

- **User Prompt:**

  REFERENCE 1 (CORRECT EXAMPLE): This is a correctly assembled product. All parts meet specifications and no extra parts are present. Use this as baseline for comparison.

  REFERENCE 2 (CORRECT EXAMPLE): This is a correctly assembled product. All parts meet specifications and no extra parts are present. Use this as baseline for comparison.

  REFERENCE 3 (CORRECT EXAMPLE): This is a correctly assembled product. All parts meet specifications and no extra parts are present. Use this as baseline for comparison.

  Examine the provided 3V rendered image. Each part is labeled with a letter (A, B, C, ..., E, etc.) and rendered in a distinct color. Identify which part is extra (unnecessary) or has wrong specifications. Return only the letter.

  Answer: C

  Examine the provided 3V rendered image. Each part is labeled with a letter (A, B, C, ..., E, etc.) and rendered in a distinct color. Identify which part is extra (unnecessary) or has wrong specifications. Return only the letter.

  Answer:

- **Model:** Qwen3-VL-235B
- **Gt answers**: C
- **Pred answer**: C

- **Raw Response:** So, let's analyze this. The standard assembly needs four M18 components: bolt, nut, flat washer, spring washer. Let's check each part.

  First, identify each part by color and shape. In the reference examples, the bolt is the long threaded part (blue in some references), nut is the hexagonal part (red), flat washer is a simple ring (light blue), spring washer is a split ring (pink).

  Looking at the current test image:

  - A: In Front View, A is on the blue bolt? Wait, no. Wait, in Front View, the blue long part is the bolt (A? Wait, no. Wait, the bolt should be the main threaded part. Let's check the references. In Reference 1, the bolt is blue (A in some views?), but maybe better to check the counts.

  Standard has 4 parts: bolt, nut, flat washer, spring washer. So 4 components. Let's count the labeled parts here: A, B, C, D. Wait, but maybe one is extra or wrong.

  Wait, in the Front View, there's a blue bolt (A?), red nut (B), light blue flat washer (D), pink spring washer (C). But wait, standard is bolt, nut, flat washer, spring washer. So 4 parts. Let's check if all are present.

  Wait, the problem is if there's an extra part or wrong spec. Wait, in the current image, are there more than 4 parts? No, A, B, C, D are four. Wait, but maybe one is the wrong type.

  Wait, in the references, the spring washer is pink (C in references), flat washer is light blue (D), nut is red (B), bolt is blue (A). So standard: A=bolt, B=nut, C=spring washer, D=flat washer.

  Now check the current image. Front View: blue bolt (A) is the long part, red nut (B) is hexagonal, light blue D is flat washer, pink C is spring washer. Wait, but in the Front View, the pink C is a flat ring, not a spring washer. Spring washers are split (like a C-shape), but here C is a full ring (flat washer), and D is also a ring. Wait, no—wait, spring washers are typically split, so they look like a "C" shape. Flat washers are solid rings.

  Looking at the current image:

  - C: pink, in Front View is a full ring (flat washer shape), but spring washer should be split. So maybe C is supposed to be spring washer but is flat, or D is wrong.

  Wait, in the reference examples, let's check. For example, Reference 1: C is pink, and in Side View, it's a split washer (spring washer). So C should be spring washer (split), D is flat washer (solid ring).

ASSYVERI SWN SCENARIO ICD

- In the current test image, let's check the shapes:

  - C: pink. In Front View, it's a full ring (solid), which would be a flat washer. But spring washer should be split. So C is supposed to be spring washer but is flat, so it's the wrong type. Alternatively, maybe C is an extra part, but there are four parts. Wait, no—standard is four parts. Wait, maybe the issue is that C is a flat washer but should be spring, and D is... Wait, D is light blue. In Front View, D is a solid ring (flat washer), but maybe there are two flat washers, which is wrong.

  Wait, standard assembly is bolt, nut, flat washer, spring washer. So two washers: flat (D) and spring (C). So C should be spring (split), D flat (solid).

  In the current image, let's check the shapes:

  - C (pink): In Front View, it's a solid ring (so flat washer), Side View shows it as a split? Wait, Side View: C is pink, looks like a split (C-shape), so maybe it is a spring washer. Wait, Front View: C is pink, solid ring? No, maybe the rendering. Wait, in the Front View, C is a pink ring, but if it's a spring washer, it should have a split. Wait, maybe in the image, C is a solid ring (flat washer), and D is also a solid ring (flat washer), so there are two flat washers, which is wrong. The standard requires one flat and one spring.

  So if C and D are both flat washers (solid rings), then one of them is wrong. Let's check the references. In Reference 1, D is light blue (flat washer, solid), C is pink (spring washer, split). So in the test image, if C is a solid ring (flat), then C is the wrong type (should be spring, but is flat), so C is the faulty part.

  Alternatively, maybe the bolt is labeled wrong. Wait, A is blue, which in references is the bolt. B is red, which is the nut. So A=bolt, B=nut, D=flat washer, C should be spring washer. If C is not a spring washer (i.e., it's a flat washer), then C is the wrong component.

  Yes, so the issue is that C is supposed to be a spring washer but is a flat washer (or vice versa), but in the standard, we need one of each. If both C and D are flat washers, then C is the extra or wrong one. Wait, no—there should be two washers: flat and spring. So if C is a flat washer (same as D), then there's an extra flat washer, so C is the faulty part (should be spring, but is flat).

  Therefore, the faulty part is C.

ASSYVERI SWN SCENARIO Zero-Shot

- **System Prompt:** Remains the same.
- **User Prompt:**

  Examine the provided 3V rendered image (front, side, and top views merged). Each part is labeled with a letter (A, B, C, ..., E, etc.) and rendered in a distinct color. Identify which part, if any, is extra (unnecessary) or has wrong specifications. Return only the letter corresponding to the problematic part.

  Answer:

- **Raw Response 1:**

  The standard assembly requires four M18 components: (1) M18 bolt, (2) M18 nut, (3) M18 flat washer, and (4) M18 spring washer.

  Let's identify the components in the image:

  * **A (Dark Blue):** This is the M18 bolt (visible in all three views).

  * **B (Red):** This is the M18 nut (hexagonal shape, visible in all three views).

  * **C (Pink):** This is a flat, ring-shaped component, which is the M18 flat washer.

  * **D (Light Blue):** This is a split, ring-shaped component, which is the M18 spring washer.

  The image shows four components: a bolt (A), a nut (B), a flat washer (C), and a spring washer (D). This matches the required set of four components.

  However, the assembly configuration is incorrect. In a standard assembly, the components are typically stacked in the order: Bolt Head $\rightarrow$ Workpiece $\rightarrow$ Flat Washer $\rightarrow$ Spring Washer $\rightarrow$ Nut.

  In the provided image, the components are not assembled but are shown as separate, floating parts in the three views. The prompt asks to check the test image for any mismatches in component configuration, and to identify any issues such as missing specifications, incorrect components, or **extra parts**.

ASSYVERI SWN SCENARIO Zero-Shot

- **Raw Response 2:**

  Upon closer inspection of the labels in the Front View: * The dark blue component is the bolt (A). * The red component is the nut (B). * The pink component is the flat washer (C). * The light blue component is the spring washer (D).

  The labels A and B are placed on the nut (B) in the Front View, suggesting that the nut (B) might be labeled twice or that one of the labels refers to a feature or an internal component. However, looking at the Side View and Top View, A clearly labels the bolt, and B clearly labels the nut.

  Re-examining the Front View: * The large dark blue object is the bolt (A). * The large red object is the nut (B). * The pink ring is the flat washer (C). * The light blue ring is the spring washer (D).

  The labels A and B are placed inside the red nut (B) in the Front View. This is highly ambiguous.

  Let's assume the labels A, B, C, ... refer to the four distinct physical components:

  A = Bolt

  B = Nut

  C = Flat Washer

  D = Spring Washer

  The set is complete and contains the correct types of components.

  If we interpret the prompt as checking for *extra* parts based on the labels provided in the image, we must look for components that are not part of the standard set. Since the standard set is (Bolt, Nut, Flat Washer, Spring Washer), and the image shows exactly these four distinct components (A, B, C, ...), there are no extra components.

  However, if we strictly follow the labels as they appear in the image, especially in the Front View, the label 'A' is placed on the nut (B) in the Front View, while 'A' labels the bolt in the Side and Top Views. This suggests a labeling error or an ambiguity.

  Let's re-examine the components based on their visual representation and standard function: 1. **Bolt (A):** Present. 2. **Nut (B):** Present. 3. **Flat Washer (C):** Present. 4. **Spring Washer (D):** Present.

  The set is complete. There are no extra components.

ASSYVERI SWN SCENARIO Zero-Shot

- **Raw Response 3:**

  If the issue is a mismatch in specifications (e.g., size), all components are visually represented as being of the same nominal size (M18, as specified in the prompt).

  If the issue is an incorrect component type: * A is a bolt. * B is a nut. * C is a flat washer. * D is a spring washer. All types are correct.

  The only remaining possibility is that one of the components is redundant or incorrectly labeled/represented, leading to an "extra part" interpretation. Since the standard set requires four parts, and four distinct parts are shown, none are technically "extra."

  However, in quality inspection tasks involving visual representations, sometimes a component is included that is not required for the *specific* assembly type, or a component is duplicated.

  Let's assume the image is intended to show a standard assembly set, and we must identify a component that is redundant if the assembly only required three parts, or if one of the washers is redundant.

  The standard combination is (1) bolt, (2) nut, (3) flat washer, (4) spring washer.

  If we look at the components C (Flat Washer) and D (Spring Washer), they are often used together, but sometimes only one is required. If the requirement was only for a flat washer, the spring washer (D) would be extra. If the requirement was only for a spring washer, the flat washer (C) would be extra. Since the prompt explicitly states the standard combination includes *both* flat and spring washers, neither C nor D is extra based on the definition provided.

  Given the constraints of the problem (identifying a faulty part, which includes extra parts), and the fact that the set is complete and correct according to the definition, the only potential "fault" is the ambiguity or redundancy in the labeling itself, or a subtle visual error.

ASSYVERI SWN SCENARIO Zero-Shot

- **Raw Response 3:**

  Let's assume the visual representation is flawed and one component is redundant. In many industrial applications, the spring washer (D) is used to prevent loosening, and the flat washer (C) is used for load distribution. If the application only required load distribution, the spring washer (D) would be considered extra. If the application only required anti-loosening, the flat washer (C) would be considered extra.

  Since the prompt defines the standard set as including all four, we must look for an error in the components themselves.

  Component D (Spring Washer) is visually represented as a split ring, which is correct for a spring washer. Component C (Flat Washer) is visually represented as a solid ring, which is correct for a flat washer.

  If we assume the standard assembly only requires one type of washer, and the flat washer (C) is the primary component for load distribution, then the spring washer (D) is the most likely candidate for being "extra" if the assembly was simplified.

  If we strictly adhere to the provided definition: "The most typical combination includes: (1) bolt, (2) nut, (3) flat washer, (4) spring washer," then all four components (A, B, C, ...) are required.

  If the image is a test image that must be checked against the standard, and the standard is the four-part set, then the image is correct in terms of component count and type.

  However, in typical quality inspection scenarios where a component is flagged as "faulty," it often means it is redundant or incorrect. Given the ambiguity, and the common practice of sometimes omitting one washer type, let's consider the possibility that one washer is redundant.

  In many contexts, the spring washer (D) is considered an optional anti-loosening feature, while the flat washer (C) is more fundamental for surface protection and load distribution. If one must be flagged as extra, it is often the spring washer.

  Let's assume the intended error is the inclusion of the spring washer (D) as an extra component beyond a basic bolt/nut/flat washer assembly.

  D

ASSYVERI SWN SCENARIO Ref-Cond

- **System Prompt:** remains the same
- **User Prompt:**

  REFERENCE 1 (CORRECT EXAMPLE): This is a correctly assembled product. All parts meet specifications and no extra parts are present. Use this as baseline for comparison.

  REFERENCE 2 (CORRECT EXAMPLE): This is a correctly assembled product. All parts meet specifications and no extra parts are present. Use this as baseline for comparison.

  REFERENCE 3 (CORRECT EXAMPLE): This is a correctly assembled product. All parts meet specifications and no extra parts are present. Use this as baseline for comparison.

  Based on the reference images showing correct assemblies, examine the TEST IMAGE above. Each part is labeled with a letter (A, B, C, ..., E, etc.) and rendered in a distinct color. Identify which part, if any, is extra (unnecessary) or has wrong specifications compared to the standard assembly. Return only the letter corresponding to the problematic part.

  Answer:

- **Model:** Qwen3-VL-235B
- **Gt answers**: C
- **Pred answer**: A

ASSYVERI SWN SCENARIO Ref-Cond

- **Raw Response 1:**

  So, let's try to figure out which part is problematic. First, I need to recall the standard assembly: it should have four M18 components: bolt (B), nut (A), flat washer (D), spring washer (C). Wait, let me check the references.

  Looking at the reference images. In each correct example, the parts should be: bolt (the long threaded part), nut (hexagonal), flat washer (circular, flat), spring washer (circular but with a split, like a spring). Let's identify each label.

  In the references, let's see:

  - Bolt is the long part with threads. In references, B is often the bolt? Wait, no. Wait, in the first reference image (top left), the red long part is labeled B. So B is the bolt.

  - Nut is the hexagonal part. In the first reference, the blue hexagonal part is A. So A is the nut.

  - Flat washer (D): in the first reference, D is the light blue circular washer.

  - Spring washer (C): pink circular with a split, so C is spring washer.

  Now, the test image is the fourth set (the last one with front, side, top views). Let's check each part.

  Test image:

  Front View: blue bolt (A?), red nut (B), light blue washer (D), pink spring washer (C). Wait, but let's check the labels.

  Wait, in the test image's front view: the blue long threaded part is labeled A. Wait, but in the references, the bolt is usually the red or blue long part. Wait, maybe the color coding is consistent.

  Wait, in the correct references, the bolt is the long threaded component. Let's check the references:

  Reference 1 (first correct example):

  - Front view: red bolt (B), blue nut (A), pink spring washer (C), light blue flat washer (D).

  Yes, so B is bolt, A is nut, C is spring washer, D is flat washer.

  So standard parts: B (bolt), A (nut), C (spring washer), D (flat washer).

  Now, test image:

  Check each part. Let's list all parts in test image.

  Front View:

  - Blue long threaded part: labeled A. But in references, A is the nut (hexagonal), so this A is the bolt? Wait, no. Wait, in references, A is the nut (hexagonal). So if the test image has a blue long threaded part labeled A, that's wrong because A should be the nut (hexagonal), not the bolt.

  Wait, let's check the test image's parts:

ASSYVERI SWN SCENARIO Ref-Cond

- **Raw Response 2:**

  Test image front view:

  - A is on the blue long threaded part (bolt). But in references, A is the nut (hexagonal). So A is mislabeled here. Wait, no. Wait, maybe the labels are consistent across all images. The labels A, B, C, D are fixed per part type.

  Wait, the problem says each part is labeled with a letter (A, B, C, D, E, etc.) and rendered in a distinct color. So A is a specific part type, B another, etc.

  In the correct references, the four parts are A, B, C, D. Let's confirm:

  In reference 1 (first correct example):

  - A is the nut (hexagonal, blue in front view).

  - B is the bolt (long, red).

  - C is spring washer (pink, circular with split).

  - D is flat washer (light blue, circular).

  Yes, that's consistent.

  Now, test image:

  We need to check if all four parts are present (A, B, C, D) and no extras.

  Test image:

  Front View:

  - A is on the blue bolt (long threaded part). Wait, in references, A is the nut, so this A is in the bolt's position. That's a problem.

  Wait, let's list all parts in test image:

  - A: appears as a blue long threaded part (bolt) in front view, and as a blue hexagonal part in side view? Wait no.

  Wait, test image front view:

  - Leftmost is blue long threaded (A), then red hexagonal (B), then light blue circular (D), then pink circular (C).

  Side view:

  - Top: D (light blue, flat washer), then C (pink, spring washer), B (red hexagonal, nut), A (blue, which looks like a bolt head? Wait, A in side view is a blue part that's the bolt head? Wait, no.

  Wait, in the test image's side view:

  - The blue part labeled A is the bolt head (since it's the large end of the bolt). So A is the bolt? But in references, A is the nut.

  Wait, this is key. In the correct assembly:

  - A should be the nut (hexagonal, fits onto the bolt).

  - B should be the bolt (long threaded).

  - C: spring washer.

  - D: flat washer.

  But in the test image, let's check the parts:

  In test image, let's see which parts are present.

ASSYVERI SWN SCENARIO Ref-Cond

- **Raw Response 3:**

  Test image:

  Front View:

  - A: blue, long threaded (so A is the bolt here)

  - B: red, hexagonal (so B is the nut)

  - C: pink, spring washer

  - D: light blue, flat washer

  Wait, but in the references, the bolt is B, nut is A. So here, A is the bolt, B is the nut. That's swapped.

  But the problem is about wrong specifications or extra parts. Let's check if all four parts are present.

  Wait, in test image, do we have all four: A, B, C, D?

  Yes: A, B, C, D are all present. But the issue is that in the correct assembly, the bolt should be a specific type. Wait, no: the problem says the standard assembly must contain four M18 components: (1) M18 bolt, (2) M18 nut, (3) M18 flat washer, (4) M18 spring washer. The labels A, B, C, D are assigned to these parts in the references.

  In the references, A is the nut, B is the bolt, C is spring washer, D is flat washer.

  In the test image, let's check:

  - Is there a bolt? Yes, but it's labeled A.

  - Is there a nut? Yes, labeled B.

  - C is spring washer.

  - D is flat washer.

  But the problem is about "wrong specifications" or "incorrect components". Wait, no—the labels are part of the identification. Each letter corresponds to a part type. So if in the test image, the part labeled A is not the nut (which it should be), then A is wrong.

  Wait, but maybe the issue is an extra part or missing part.

  Wait, let's count the parts. In correct assembly: 4 parts (A, B, C, D). Test image: are there 4 parts?

  Test image front view: A, B, C, D are all there. Side view: A, B, C, D. Top view: A, B, C, D. So no extra parts.

  But wait, in the test image, let's check the configuration. In the correct assembly, the bolt (B) should have the nut (A) at the end, with washers in between. Wait, the standard combination is bolt, then flat washer, spring washer, then nut? Or the order may vary but the parts must be present.

ASSYVERI SWN SCENARIO Ref-Cond

- **Raw Response 4:**

  Wait, the key is: the test image may have a part that is not supposed to be there, or a part that's missing, or a wrong type.

  Wait, looking at the test image, let's see the side view. In the side view, there's a part D (flat washer), C (spring washer), B (nut), and A (bolt). But in the correct references, when assembled, the bolt (B) is the main shaft, nut (A) is at the end. In the test image, the bolt is labeled A, which is supposed to be the nut. So A is misused as the bolt.

  Wait, but the problem says "identify any issues such as missing specifications, incorrect components, or extra parts".

  Wait, another angle: in the correct references, each part is one of the four. In the test image, do we have all four? Let's see:

  - Bolt: should be present. In test image, A is the bolt (blue long).

  - Nut: B is the nut (red hexagonal).

  - Flat washer: D (light blue).

  - Spring washer: C (pink).

  So all four parts are present. But the problem is that the labels are assigned to wrong parts. Wait, no—the labels are fixed. A is always supposed to be the nut. So if in the test image, A is the bolt, then A is the wrong component.

  Wait, let's check the reference images again.

  Reference 1 (first correct example):

  - Front view: A is blue hexagon (nut), B is red bolt, C is pink spring washer, D is light blue flat washer.

  Yes, so A = nut, B = bolt, C = spring washer, D = flat washer.

  In test image:

  - A is the blue bolt (long threaded), so A is not the nut. Therefore, A is the wrong component—it's a bolt but should be a nut.

  Wait, but maybe the bolt is supposed to be B. So B should be the bolt. In test image, B is the red hexagon (nut), so B is the nut. Therefore, B is correct as the nut. A is the bolt, but A should be the nut. So A is a bolt where a nut should be. Therefore, A is the incorrect part.

  Alternatively, maybe there's an extra part. Wait, no, all four parts are present.

  Wait, let's look at the test image's side view. In the side view, there's a blue part labeled A which is the bolt (since it's the long shaft), and a red part labeled B which is the nut. But in the correct assembly, the bolt should be B, so A should not be the bolt. Therefore, A is the wrong part because it's a bolt but labeled as A (which should be nut).

  Yes, that's the issue. So the problematic part is A.

E.4 Examples of Bottleneck Analysis A

Below are examples from the cross-image visual grounding analysis (Bottleneck Analysis A in Section 4.5 of the main text). Each example shows a pair of Set-of-Mark annotated images where the model must match a part identified in the first image to its visual counterpart in the second image. We present both coordinate-to-coordinate (C→C) and letter-to-letter (L→L) cases from Mistral-3-8B, illustrating typical failure modes.

Grounding example of PCS SCENARIO 1

- **Setting:** Zero-shot; Coordinates to Coordinates (**C→C**)
- **Gt answer**: A
- **Pred answer:** C
- **Raw response**: In Image 1, the part at coordinates [0.54, 0.59] is labeled **C**. This part is a cylindrical component with a hexagonal section near the top, which is distinct from the other parts.

  In Image 2, we need to identify the same type of component (cylindrical with a hexagonal section near the top). The part labeled **C** in Image 2 matches this description.

  Thus, the correct answer is:

  Answer: C

Grounding example of PCS SCENARIO 2

- **Setting:** Zero-shot; letters to letters (**L→L**)
- **Gt answer**: B
- **Pred answer:** C
- **Raw response**: In Image 1, Part D is a cylindrical component with a smooth outer surface and a threaded inner section that appears to be a barrel or sleeve type with a slightly tapered end.

  Comparing this to Image 2: - Part A is cylindrical but has a different texture and shape, resembling a threaded connector with a wider end. - Part B is a long cylindrical component with a smooth surface and a wider middle section, not matching Part D. - Part C is a cylindrical component with a threaded outer section and a smooth inner section, similar to Part D in Image 1. - Part D in Image 2 is a cylindrical component with a smooth surface and a wider middle section, similar to Part B in Image 1.

  The correct match for Part D in Image 1 is Part C in Image 2.

  Answer: C

E.5 Examples of Bottleneck Analysis B

Below is an example from the missing-part detection analysis (Bottleneck Analysis B in Section 4.5 of the main text). The model is given an explicit assembly specification listing all required components, and must identify which part is absent from the three-view rendering. This example shows a Qwen3-VL-235B response on a SWN SCENARIO case.

Missing part of SWN SCENARIO

- **System Prompt:** You are an expert manufacturing quality inspector. Your task is to examine three-view rendered images (front, side, top) of an assembly and determine whether all required components are present. Each visible part is labeled with a letter and colored distinctly. You will be given a description of what a complete assembly should contain, and you must identify which part (if any) is missing from the test image. Answer with ONLY the letter of the correct option (e.g., 'Answer: C').
- **User Prompt:**

  You are inspecting a three-view rendering of an assembly that should be a complete Metal Expansion Bolt.

  Assembly Description: A standard fastening assembly consists of four M18 components: (1) M18 Bolt – the main fastening element; (2) M18 Nut – engages with the bolt to create a secure connection; (3) M18 Flat Washer – distributes pressure evenly and protects contact surfaces; (4) M18 Spring Washer – prevents loosening during operation. A complete assembly must contain all four components.

  The test image shows front, side, and top views of the assembly. Each part is labeled with a letter and rendered in a distinct color. Compare the visible parts against the expected components listed above. Count the parts and determine if any required component is missing.

  Which of the following parts is missing from this assembly? A. No part is missing

  B. Nut C. Flat Washer D. Spring Washer E. Bolt

  Answer with ONLY the letter of your choice (e.g., 'Answer: C').
- **Setting:** Zero-shot;
- **Model** Qwen3-VL-235B;
- **Gt answer**: A
- **Pred answer**: A

Missing part of SWN SCENARIO

- **Raw Response 1:** Got it, let's analyze this. The assembly should have four components: Bolt, Nut, Flat Washer, Spring Washer. The parts in the image are labeled A, B, C, D. Let's map each label to the components.

  First, let's recall the standard parts. A bolt is the main fastening element (probably the long red one, B). A nut engages with the bolt (maybe the red rounded part in side view? Wait, side view has B overlapping, but maybe B is the bolt. Then washers: flat washer is a flat ring, spring washer is a split ring. Wait, but the problem says four M18 components: Bolt, Nut, Flat Washer, Spring Washer.

  Looking at the labels:

  - A: Blue, in front view is a circle, side view is a ring (so washer), top view is a small blue shape. So A is likely a washer (flat or spring).

  - B: Red, long in front and top views, so that's the bolt (main fastening element).

  - C: Pink, in all views, maybe the nut? Wait, nut is a hexagonal part that engages with the bolt. In side view, C is a pink shape next to B.

  - D: Light blue, in front, side, top views as a small shape. Maybe the other washer.

  Wait, the assembly needs four parts: Bolt (B?), Nut (maybe C?), Flat Washer (A?), Spring Washer (D?).

  But wait, the problem states four components: Bolt, Nut, Flat Washer, Spring Washer. So all four must be present.

  Let's count the labels: A, B, C, D. Four labels. But wait, are all four parts actually present? Wait, maybe one of the labels is a duplicate or one is missing.

  Wait, no—the labels are A, B, C, D. So four parts. But the question is which is missing. Wait, the options include "No part is missing" (A), but let's check.

  Wait, the assembly description says four components: Bolt, Nut, Flat Washer, Spring Washer. So each label should correspond to one of these. Let's map:

  - Bolt: B (red, long, main fastening element)

  - Nut: C (pink, maybe the hex nut)

  - Flat Washer: A (blue, ring)

  - Spring Washer: D (light blue, maybe the spring washer)

  But wait, the problem is to check if all four are present. The labels are A, B, C, D—so four parts. But wait, maybe one of the components is not present. Wait, no—the labels are distinct, so four parts. But the options include "No part is missing" as option A. But wait, the problem says "which of the following parts is missing". Wait, maybe the labels correspond to different parts.

Missing part of SWN SCENARIO

- **Raw Response 2:**

  Wait, let's think again. The four required components are Bolt, Nut, Flat Washer, Spring Washer. Let's see which is missing.

  If the assembly has all four, then option A. But maybe one of the labels is not a required component. Wait, no—the labels are the parts in the assembly. Wait, the problem says "each visible part is labeled with a letter". So if there are four labels (A,B,C,D), but maybe one of the required components is not present, meaning that one of the labels is not a required part, but no—wait, the required parts are four, so the assembly should have four parts. If there are four parts labeled, but one of the required components is missing, that would mean that one of the labels is a different part, but the problem states that the assembly should consist of those four. Wait, maybe the labels correspond to the four, but one of the required components is not present.

  Wait, let's list the required components:

  1. M18 Bolt – main fastening element (B, since it's the long red part)

  2. M18 Nut – engages with bolt (maybe C, pink)

  3. M18 Flat Washer – distributes pressure (A, blue ring)

  4. M18 Spring Washer – prevents loosening (D, light blue)

  So if all four are present, then A (no part missing). But wait, the options are A. No part missing, B. Spring Washer, C. Flat Washer, D. Bolt, E. Nut.

  Wait, but let's check the side view. In side view, there's A (blue ring), B (red, overlapping), C (pink), D (light blue at top). So all four labels are present. So why would a part be missing? Wait, maybe one of the components is not present. Wait, but the labels are A, B, C, D. So four parts. The required is four. So maybe all are present. But wait, the problem says "a complete assembly must contain all four components". So if all four are present, answer A.

  But wait, let's check the assembly description again. The four components are Bolt, Nut, Flat Washer, Spring Washer. Let's see which part is which.

  Bolt is the main fastening element (B), Nut is the one that threads onto the bolt (probably C), Flat Washer (A), Spring Washer (D). So all four are present. But wait, the options include A. No part missing. So why would the answer not be A?

Missing part of SWN SCENARIO

- **Raw Response 3:**

  Wait, maybe I made a mistake. Let's count the parts. The labels are A, B, C, D. Four parts. The required is four. So all are present. But maybe one of the components is not present. Wait, perhaps the Spring Washer is missing. Wait, in the side view, D is a light blue shape at the top. Is D the Spring Washer? Let's think: Spring Washers are typically helical or split, but in a rendering, maybe it's a different shape. Flat Washer is a flat ring (A), which matches the blue ring in side view. Bolt is B, Nut is C. Then D would be the Spring Washer. So all four are present. So answer A.

  But wait, the problem says "determine whether all required components are present". If all four labels correspond to the four required components, then no part is missing. But let's check the options. Option A is "No part is missing".

  Wait, but maybe the assembly is supposed to have four parts, but in reality, one of the labels is not a required component. No, the problem states that the assembly should consist of those four, so each required component must be present. Since there are four labeled parts, and four required, they must map one-to-one. Therefore, no part is missing. So answer A.

  But wait, let's re-examine. The user might have a trick. Let's see: the assembly is a Metal Expansion Bolt. Wait, the description says "a standard fastening assembly consists of four M18 components: (1) M18 Bolt... (2) M18 Nut... (3) M18 Flat Washer... (4) M18 Spring Washer".

  Wait, but a Metal Expansion Bolt might not require a nut? Wait, no, the problem states that the assembly should have all four. So regardless of real-world, per problem statement, four parts are required.

  Looking at the image: front view has A, B, C, D. Side view has A, B, C, D. Top view has A, B, C, D. So all four parts are present. So answer A.

  But wait, the options list A as "No part is missing". So the answer should be A.

  Wait, but maybe I'm wrong. Let's check again. Let's list each component:

  - Bolt (B): present (red, long)

  - Nut (C): present (pink)

  - Flat Washer (A): present (blue ring)

  - Spring Washer (D): present (light blue)

  All four are present. So no part missing. Answer A

E.6 Examples of Bottleneck Analysis C

Below is an example from the raw point cloud text input analysis (Bottleneck Analysis C in Section 4.5 of the main text). The model receives 3D coordinates serialized as integer-scaled text tables and must perform the standard ASSYVERI task with-

out any visual rendering. This example shows a Qwen3-235B response on a SWN SCENARIO case.

Pure point input example of SWN SCENARIO

- **System Prompt:** You are an expert at analyzing 3D point cloud data for industrial quality control.

  Your task:

  - Identify which part (if any) deviates from the expected specification based on 3D point-cloud data.

  - At most one part deviates from the others.

  Input format:

  - "Description of context:" (optional) - Background information about the parts and their function.

  - "Task:" - The specific task instruction.

  - "Query:" - Indicates that the 3D point cloud data follows.

  - "Data:" - 3D coordinates of each part's point cloud, formatted as:

  Part A:

  x, y, z

  1234, 5678, 9012

  1235, 5679, 9011

  ...

  Part B:

  x, y, z

  ...

  Output format

  - Return the Part ID (e.g., "Part A") of the deviating part.

  - If all parts match specifications, state that all parts are correct.

Pure point input example of SWN SCENARIO

- **User Prompt:**

  Task: There are 4 parts in total. Identify the only part whose 3D point-cloud data deviates from the specified model's design compared to the others.

  Query: The 3D point cloud data for all parts is provided below.

  Data:

  Part A:

  x, y, z

  3930, -3952, 42338

  4664, -4289, 42390

  3381, -3782, 42378

  ...

  Part B:

  ...

  5180, -698, 32471

  4244, -1049, 32322

  Please answer the query by thinking about the shape, size or other features of the part that represented by the point cloud data.

  You may use any reasoning approach you feel is needed to make an accurate determination.

  Output format:

  - Return the Part ID of the part in question like: "Part X". Be sure to give the result at the end of your response.",

- **Brief info:** Zero-shot;
- **Model:** Qwen3-235B
- **Gt answer**: C
- **Pred answer:** C
- **Raw response1**:

  To identify the deviating part from the 3D point cloud data, I analyzed the dimensional characteristics, aspect ratios, and spatial distribution of points for each part.

  First, I examined the coordinate ranges for each dimension (x, y, z) across all parts:

  Part A:

  - x: ~2500-5800 (span ~3300)

  - y: ~-4300 to -1400 (span ~2900)

  - z: ~40700-42600 (span ~1900)

Pure point input example of SWN SCENARIO

- **Raw response2**: Part B:

  - x: ~9400-12100 (span ~2700)

  - y: ~1000-3600 (span ~2600)

  - z: ~37000-38300 (span ~1300)

  Part C:

  - x: ~6300-16400 (span ~10100)

  - y: ~-6400-600 (span ~7000)

  - z: ~43700-49900 (span ~6200)

  Part D:

  - x: ~4200-7100 (span ~2900)

  - y: ~-2900-300 (span ~3200)

  - z: ~31300-33900 (span ~2600)

  The most striking difference is in the z dimension span:

  - Part C's z span is ~6200, which is 3.26 larger than Part A's (1900), 4.77 larger than Part B's (1300), and 2.38 larger than Part D's (2600)

  When examining aspect ratios (normalized by the smallest dimension):

  - y:z ratio: Part A (1.53:1), Part B (2.00:1), Part C (1.13:1), Part D (1.23:1)

  - Part C has a significantly smaller y:z ratio compared to the others, indicating it's much "taller" relative to its width

  Part C also shows an unusual pattern in point distribution:

  - Contains numerous points with extreme z values (up to 49900) that are well beyond the z ranges of other parts

  - Has points with both highly negative y values (down to -6400) and positive/near-zero values, creating a wider y distribution

  Most critically, Part C's absolute size in all dimensions is substantially larger than the other parts:

  - x span is ~3.7 larger than the average of other parts

  - y span is ~2.4 large

  - z span is ~3.3 larger

  This disproportionate scaling across all dimensions, combined with the abnormal y:z aspect ratio, indicates that Part C has a fundamentally different shape than the other three parts. In industrial quality control, such significant dimensional deviation from the expected model would classify this as a defective part.

  Part C