

LDU-Bench: Multimodal LLM Evaluation for Lithography Defect Understanding under Layout-Varying Circuit Backgrounds

Huanglong Ji Botong Zhao Shujing Lyu Yue Lu[§]

East China Normal University, Shanghai, China

51275904013@stu.ecnu.edu.cn, zhaobotong19@mails.ucas.edu.cn

sjlv@cs.ecnu.edu.cn, ylu@cee.ecnu.edu.cn

[§]Corresponding author.

Abstract

Multimodal large language models have demonstrated strong defect recognition capability in industrial anomaly detection. However, in lithography review, merely determining whether an image contains a defect is insufficient for engineering inspection; models must also understand defect morphology, spatial location, and the potential causes supported by visible evidence. To this end, this paper proposes LDU-Bench, a multi-task multimodal benchmark for lithography defect understanding. Constructed from real lithography and integrated-circuit review images, LDU-Bench decomposes the review workflow into four independent tasks: defect triage, morphology recognition, coarse localization, and image-conditioned cause analysis. It systematically evaluates models using task-level metrics, diagnostic readouts, and the Lithography Closure Score (LCS). Experimental results show that although existing MLLMs can perform defect triage relatively reliably, this ability does not stably transfer to downstream review stages. Morphology alignment, effective localization, and evidence-to-cause mapping remain the major bottlenecks. Further diagnostics indicate that this capability break is not a fluctuation of a single metric, but reflects insufficient structured understanding across semantic levels. Overall, LDU-Bench provides a quantifiable and diagnostic unified platform for evaluating the usability, failure points, and capability boundaries of industrial MLLMs in lithography review chains.

1 Introduction

Existing semiconductor and lithography defect inspection work largely frames the problem as classification, detection, localization, or segmentation (Dehaerne et al., 2025; De Ridder et al., 2023; Shin and Lee, 2016; Yang et al., 2017; Liao et al., 2022; Dey et al., 2022). These methods can output classes, bounding boxes, or segmentation masks,

but they do not directly provide morphology descriptions, spatial explanations, or cause evidence that can be used in downstream human review.

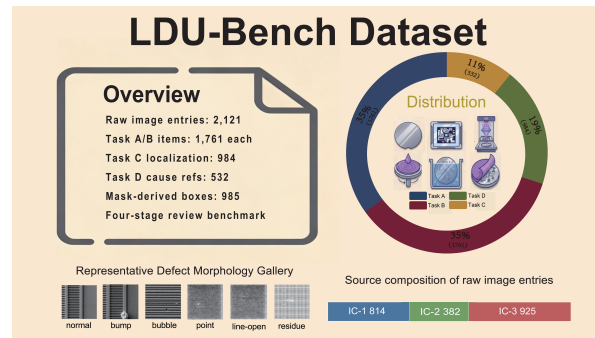


Figure 1: Dataset and task split. LDU-Bench uses lithography IC-SEM review images with image-level labels, mask-derived spatial annotations, and reviewed cause references.

Recent industrial anomaly detection work has introduced vision-language interfaces. Compared with labels, boxes, or masks, natural language can express morphology, location, and cause cues, and is therefore closer to the judgment required in review. MVTec AD, VisA, Real-IAD, and MVTec LOCO standardize anomaly detection and segmentation tasks (Bergmann et al., 2019, 2022; Zou et al., 2022; Wang et al., 2024). Representative industrial anomaly detection methods further study patch-distribution modeling, reconstruction-based discrimination, memory-based retrieval, and efficient student-teacher detection (Defard et al., 2021; Zavrtanik et al., 2021; Roth et al., 2022; Batzner et al., 2024). Building on image-text alignment from CLIP (Radford et al., 2021), WinCLIP, ALFA, AnomalyCLIP, and AnomalyGPT further adapt vision-language models to anomaly recognition, segmentation, and interactive localization (Jeong et al., 2023; Zhu et al., 2024; Zhou et al., 2024; Gu et al., 2024). MMAD organizes industrial anomaly detection as multimodal question answer-

ing and systematically evaluates MLLMs (Jiang et al., 2025). However, existing work mainly targets general industrial objects. Its evaluation objectives remain anomaly discovery, segmentation, or QA accuracy, and do not specifically test whether strong defect triage in lithography implies downstream morphology recognition, coarse localization, and image-conditioned cause analysis.

Lithography review places stronger requirements on this capability chain. The appearance of many industrial inspection datasets is relatively stable, whereas the normal background of circuit images varies with design layer, local routing, and pattern density. Defects are often embedded in dense periodic or semi-periodic structures. Therefore, review understanding requires more than detecting a local defect. A model must organize local visual morphology, spatial location, and possible cause into an auditable information chain. Evaluating only whether a defect is found cannot determine whether the model has downstream review understanding.

Therefore, we propose LDU-Bench, a multi-task multimodal benchmark for lithography defect understanding. The de-identified benchmark is publicly available for non-commercial research evaluation at <https://github.com/0wen1103/LDU-bench>, with release contents and licensing details provided in Appendix B. It is constructed from real lithography and integrated-circuit review images and divides the review process into four independently evaluated tasks: defect triage, morphology recognition, coarse localization, and image-conditioned cause analysis. LDU-Bench uses deterministic scorers, diagnostic readouts, and LCS as an overall review-chain summary. This design directly tests whether defect-triage performance is sufficient to indicate downstream review understanding. The results show that high triage scores do not imply stable performance in later review stages. For the evaluated MLLMs, morphology grounding, localization utility, and evidence-to-cause mapping remain the main bottlenecks.

2 Method

2.1 Overall Framework

The core of LDU-Bench is to test whether models can map defect evidence to reviewable morphology, location, and image-supported cause information. Therefore, we organize lithography defect understanding into four independent tasks: defect triage, morphology recognition, coarse localization,

and image-conditioned cause analysis. These tasks follow the order of review information needs, but they are evaluated independently. This design avoids cascade error and directly tests whether defect triage can stably extend to downstream review understanding.

This design allows LDU-Bench to answer two questions. First, how does a model perform at each review stage. Second, whether the model’s review ability can remain continuous across stages, rather than being strong only at one stage. In addition to task-level metrics, we introduce LCS as an overall summary to describe the model’s completion of the four review stages. By combining task-level scores and LCS, LDU-Bench can compare overall performance while exposing where the review chain becomes weak. Figure 2 shows the structure of the LDU-Bench dataset.

2.2 Dataset Construction

We construct LDU-Bench from IC-SEM lithography review images, covering multiple semiconductor process stages, including BEOL, DEP, and DPR.

The dataset is organized with a unified manifest. This manifest integrates image-level annotations, spatial regions, and expert-reviewed cause labels using a consistent schema.

This design supports multi-stage evaluation of defect understanding. Each image is associated within a unified framework with triage labels, morphology categories, spatial annotations, and cause-level explanations.

2.3 Benchmark Tasks and Metrics

The scoring goal of LDU-Bench is not to produce a single overall ranking, but to answer a more specific question: at which review stage does a model begin to become unreliable. Therefore, we first define the main metric for each of the four tasks, then use diagnostic analyses to explain the sources of low scores, and finally use LCS to provide an overall score.

Task A performs defect triage on 1,761 images, requiring the model to make a binary decision between good and defect. This corresponds to the initial screening stage in industrial review workflows. Task B uses the same image set and requires predicting a fine-grained morphology label from normal plus 11 defect categories. Task B can be viewed as a fine-grained extension of Task A, requiring the model to further match the detection

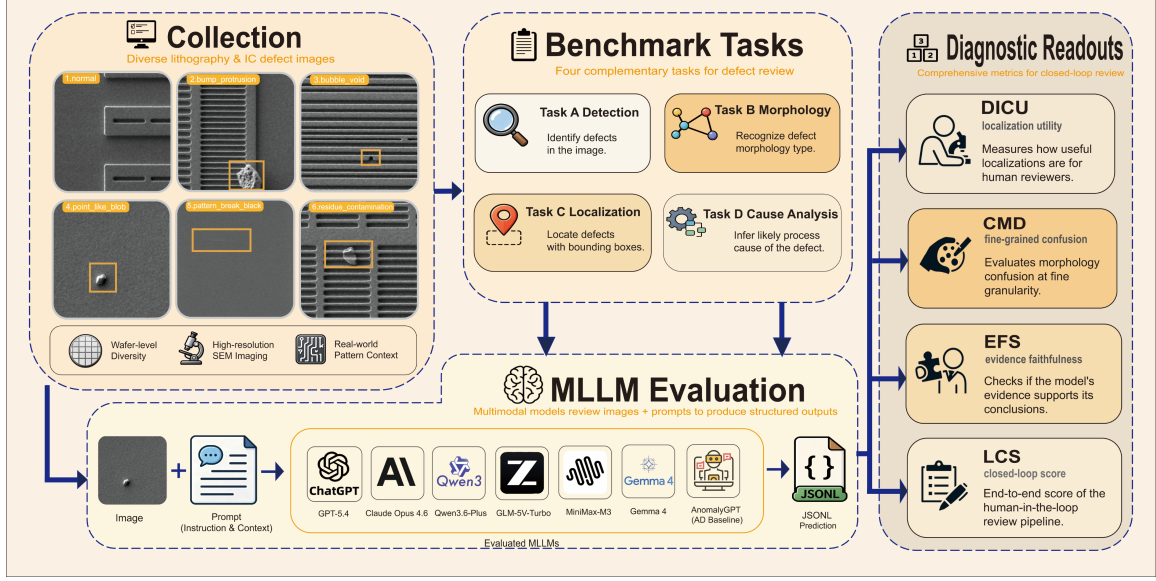


Figure 2: Overall workflow of LDU-Bench. The benchmark decomposes lithography review into defect triage, morphology recognition, coarse localization, and image-conditioned cause analysis, while also evaluating task-level performance and diagnostic readouts.

result with the corresponding morphology label after defect detection.

Task C evaluates spatial grounding on 984 mask-annotated samples, where the model must output an $xyxy$ bounding box. This setting follows the common evaluation paradigm of grounded vision-language tasks (Peng et al., 2023; Chen et al., 2023; You et al., 2023) and simplifies pixel-level masks into bounding box expressions, allowing the evaluation to focus on coarse-grained spatial localization ability.

Task D uses 532 expert-reviewed curated samples to construct an image-conditioned cause reasoning task, requiring the model to predict a cause label while also giving a brief rationale.

In evaluation, Task A and Task B use macro-F1. This metric is more robust to class imbalance in the morphology distribution, and accuracy is used as a secondary metric. Task C uses Defect Intersection-Coverage Utility (DICU), which jointly accounts for overlap quality (IoU) and region coverage (GT coverage):

$$\text{DICU}_i = \frac{2 \cdot \text{IoU}_i \cdot \text{GTcov}_i}{\text{IoU}_i + \text{GTcov}_i}.$$

Task D adopts a structured rubric-based evaluation. Exact matches or alias matches for the cause are scored as 1, semantically related matches within the same supergroup are scored as 0.5, and all other cases are scored as 0. The final score combines semantic correctness and keyword-level

evidence alignment. The weights balance semantic correctness and evidence consistency, with a higher weight on semantic matching to reflect the primary objective of cause attribution:

$$S_D = 0.7 \cdot \text{Sem}_i + 0.3 \cdot \text{KeyF1}_i.$$

To further analyze sources of model failure beyond task-level performance, we introduce two diagnostic metrics. CMD compares label-only and definition-guided prompts on confusable defect pairs to distinguish whether morphology errors arise from visual ambiguity or label understanding difficulty. EFS evaluates the visual faithfulness of explanations in Task D, measuring the consistency between generated rationales and observable visual evidence, and identifying inference bias driven by hallucination or non-visual evidence (Li et al., 2023; Guan et al., 2024).

We use a power-mean form of LCS to summarize overall performance:

$$\text{LCS}_p = \left(\frac{S_A^p + S_B^p + S_C^p + S_D^p}{4} \right)^{1/p}, \quad p > 0,$$

where S_A , S_B , S_C , and S_D denote macro-F1, macro-F1, mean DICU, and mean Task D rubric score, respectively. The main results use $p = 0.5$. The model ranking remains stable under $p \in \{0.25, 0.5, 0.75, 1.0\}$.

3 Experiments

3.1 Model Evaluation Setup

Our evaluation covers three types of models. The first type is commercial general MLLMs: GPT-5.4 (OpenAI, 2026), Claude Opus 4.6 (Anthropic, 2026), Qwen3.6-Plus (Alibaba, 2026), GLM-5V-Turbo (Zhipu AI, 2026), and MiniMax-M3 (MiniMax, 2026). We keep one representative model for each provider and keep model release time as consistent as possible, so that a single provider does not dominate the table with multiple versions. The second type is open-weight general MLLMs. We use locally deployed Gemma 4 31B-it (Google DeepMind, 2026) as a reproducible reference. The third type is a local anomaly-detection baseline. We deploy and obtain AnomalyGPT results on Task A and Task C to compare defect triage and coarse localization.

3.2 LDU-Bench Dataset and Metrics

The evaluation in LDU-Bench contains four tasks. Task A and Task B each contain 1,761 images. Task C contains 984 images with spatial annotations. Task D contains 532 reviewed image-conditioned cause references.

One complete A/B/C/D evaluation contains 5,038 task instances. Expanded to six full-chain models, this corresponds to 30,228 model-task evaluation units. With the additional evaluation of AnomalyGPT on Task A and Task C, the overall evaluation covers 32,973 task-level instances.

Task A and Task B use macro-F1 as performance metrics. Task C uses DICU to evaluate spatial grounding ability. Task D uses a deterministic rubric score to evaluate cause reasoning quality. Meanwhile, the diagnostic metric CMD is used to analyze error sources in Task A/B by comparing label-only and definition-guided prompts on confusable defect pairs, distinguishing whether morphology errors arise from visual ambiguity or label understanding difficulty. The diagnostic metric EFS is sampled from Task D outputs to evaluate the visual faithfulness of image-grounded explanations, namely whether the generated rationale depends on visible visual evidence rather than reasoning bias driven by language priors or hallucinated information.

3.3 Inference and Scoring Protocol

To make model differences mainly come from task ability, we control randomness during inference

and scoring. For all controllable decoding interfaces, we use deterministic settings. Commercial APIs that expose a temperature parameter are set to 0. Output length is capped at 512 or 1024 tokens according to the adapter configuration to prevent overly long outputs.

We evaluate Gemma 4 and AnomalyGPT under a unified batch inference setup, using fixed seeds to ensure reproducibility and fair comparison. Scoring uses the same task prompts, output schemas, and parsing rules. To ensure objective and reproducible scoring, the main task scores never use an LLM judge. Except for the human EFS diagnostic, all A/B/C/D metrics are generated by deterministic scorers. Across the six full-chain models, the overall task-level output validity rate is 99.88%; detailed validity results are reported in Table 10.

3.4 Main Results and Diagnostic Findings

Table 1 summarizes the main results on the four LDU-Bench tasks. Columns A/B report macro-F1, column C reports DICU, column D reports the deterministic rubric score, and the LCS column reports the four-task closure score. AnomalyGPT is included as an anomaly-detection baseline and contains only A/C results.

The most important finding in Table 1 is not the model ranking, but that defect-triage ability does not stably transfer to downstream review understanding tasks. The five commercial MLLMs achieve high F1 on Task A, showing that current models already have some defect-triage ability. However, this advantage does not carry over to Task B/C/D. In other words, models can usually judge whether an image contains a defect, but they cannot stably answer what morphology the defect has, where it is located, and which possible cause is supported by visible evidence.

This gap first appears in Task B. GPT-5.4 obtains the highest scores on Task A, Task C, and Task D, but drops to 0.201 on morphology recognition. GLM-5V-Turbo obtains the highest Task B score of 0.408, but does not maintain the same advantage in other review stages. This indicates that Task B is not a natural extension of Task A. A model that can complete defect triage does not necessarily map local visual morphology stably to the controlled morphology labels in LDU-Bench. Therefore, morphology naming is the first bottleneck between defect detection and review understanding.

The results of Task C and Task D further show

Model	Task A macro-F1	Task B macro-F1	Task C DICU	Task D score	LCS _{0.5}	Output validity
GPT-5.4	0.932	0.201	0.492	0.409	0.474	100.00%
GLM-5V-Turbo	0.884	0.408	0.353	0.297	0.462	100.00%
MiniMax-M3	0.909	0.239	0.338	0.395	0.439	100.00%
Qwen3.6-Plus	0.856	0.249	0.269	0.376	0.408	99.96%
Claude Opus 4.6	0.876	0.266	0.254	0.255	0.378	100.00%
Gemma 4 31B-it	0.504	0.060	0.046	0.080	0.132	99.34%
AnomalyGPT	0.759	–	0.384	–	–	100.00% on A/C

Table 1: Main results on LDU-Bench. Bold values indicate the best performance in each column. AnomalyGPT is an AD baseline and is not evaluated on the full review chain.

that this performance gap is not limited to the level of label naming, but extends across two deeper capability dimensions: localization and cause reasoning. On spatial grounding and cause explanation tasks, the upper bound of current commercial MLLMs remains limited. The highest DICU in Task C is only 0.492, and the highest rubric score in Task D is only 0.409. This indicates that even the strongest model struggles to simultaneously provide stable coarse localization and reliable image-evidence-based attribution.

In the comparison across model types, AnomalyGPT outperforms all commercial MLLMs except GPT-5.4 on Task C, indicating that specialized industrial anomaly detection models still have advantages on coarse localization. However, this model does not support morphology recognition or cause analysis, and therefore cannot form a complete review reasoning chain. By contrast, Gemma 4 31B-it, as the open-weight baseline model, lags clearly behind closed-source models on all four tasks, showing that current general open-weight foundation MLLMs still have a substantial capability gap in this structured lithography review task.

To explain the failure causes behind these scores, Figure 3 reports diagnostics from three angles: morphology grounding, localization utility, and evidence-to-cause mapping. Figure 3A shows that some morphology recognition errors can be mitigated by short definitions. GPT-5.4 improves from 0.530 accuracy / 0.480 macro-F1 under label-only prompting to 0.630 / 0.604 under definition-guided prompting. Other commercial models also reach 0.583 to 0.661 macro-F1 in definition-guided CMD. This shows that lower Task B scores cannot be fully attributed to visual discrimination failure. Because fab-specific morphology labels differ across fabs in annotation standards and defect definitions, label semantics are inconsistent across sources, which affects the model’s ability to maintain stable semantic

alignment.

Figure 3B shows that effective localization cannot be judged only by defect coverage. Large geometric boxes can easily obtain high GT coverage. For example, the full-image box has GT coverage of 1.000, but its mIoU is only 0.0387 and its DICU is only 0.0663. The center-half box still has 0.7257 GT coverage, but its DICU is only 0.1352. This means that overly large prediction boxes cover the defect while also covering a large amount of irrelevant background. Real model localizations are better than these simple geometric methods, but still do not reach reliable review quality. GPT-5.4 obtains the highest DICU, 0.4917, while other commercial models mainly fall between 0.25 and 0.35. Therefore, the bottleneck of Task C is not whether a model can roughly point to an abnormal region, but whether it can provide a compact and useful box for review.

Figure 3C further analyzes evidence faithfulness in Task D outputs. Among 250 manually scored responses, mean relevance is 0.992, mean consistency is 0.760, hallucination rate is 1.2%, and mean EFS is about 0.755. This means that models usually write visible anomaly cues and rarely explicitly cite image-external information. However, the Task D score on the same subset remains low. The best model reaches only 0.216. This shows that the presence of visible evidence is not equivalent to correct cause attribution. A model may capture local anomaly cues, but still fail to map them to the controlled cause reference. Therefore, this paper mainly uses EFS to audit whether Task D explanations have image evidence, rather than as a model-ranking metric.

In addition to individual diagnostics, LCS reveals systematic differences in model capability at the overall review-chain level. Table 1 shows that the commercial models occupy a relatively narrow range on Task A, while their observed LCS values

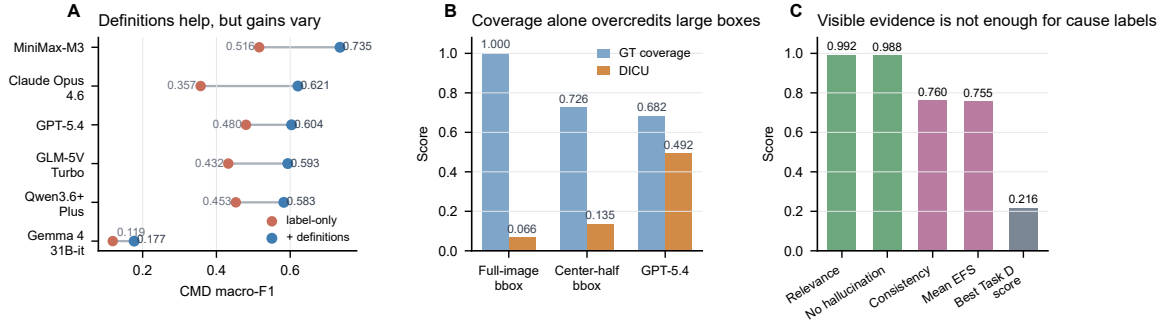


Figure 3: Diagnostic readouts for morphology grounding, localization utility, and evidence-to-cause mapping. Panel (A) reports CMD macro-F1 under label-only and definition-guided settings. Panel (B) compares GT coverage and DICU for geometric boxes and GPT-5.4. Panel (C) reports EFS pilot scores and the best Task D score on the same subset.

are more separated (GPT-5.4: 0.474; GLM-5V-Turbo: 0.462). These descriptive values indicate that the overall differences arise mainly from downstream review tasks rather than triage alone.

GPT-5.4’s observed LCS is supported by its Task C and Task D scores, while GLM-5V-Turbo narrows the gap through its relative advantage on Task B. Other models weaken at multiple downstream stages. Therefore, LCS is not intended as a linear end-to-end success measure, but as an aggregate diagnostic of capability balance across the independently evaluated tasks.

Combining the main results and diagnostic analyses, the major limitation of current MLLMs is not the absence of defect perception, but the difficulty of maintaining stable ability as task semantic level increases. Although models perform relatively stably on the basic defect-triage task (Task A), performance shows a systematic downward trend when tasks extend to higher-level structured understanding, such as morphology grounding, spatial localization (Task C), and evidence-to-cause reasoning (Task D). This phenomenon indicates that existing MLLMs are more biased toward low-level visual discrimination and lack consistent generalization in structured understanding and causal reasoning across semantic levels.

3.5 Reliability Analysis

This section further tests whether the evaluation protocol provides stable and interpretable signals. We mainly analyze the annotation reliability of morphology labels.

For Task B, we conduct an independent blind-labeling experiment on 100 samples to test whether the frozen morphology taxonomy in LDU-Bench is sufficient to support the main evaluation. The

result reaches 79.0% exact agreement and Cohen’s $\kappa = 0.727$ (Cohen, 1960), indicating high annotator consistency. Further inspection shows that disagreements mainly occur between visually adjacent categories, such as bump_protrusion and buried_anomaly. This shows that Task B is not an arbitrary subjective classification task, but has an auditable label basis.

The stage-wise results support the same central finding. The drop from Task A to Task B remains clear across BEOL, DEP, and DPR, rather than being driven by one process stage. The clearest stage effect appears in Task C, where DEP is more challenging. Task D covers BEOL and DEP, so we do not extend its stage-level interpretation to DPR.

3.6 From Diagnostic Readouts to Deployment Decisions

To make the diagnostic readouts easier to use in practice, Table 2 gives representative mappings from an observed signal to a deployment action.

These mappings turn LDU-Bench into a pre-deployment capability gate. They indicate where a model can be used, where human review remains necessary, and which bottleneck should be addressed next.

Because the four tasks are independently scored and use partly different subsets, LCS is an aggregate diagnostic of capability balance. It is not evidence that a model completes a sample-level end-to-end review chain. Commercial-model rankings are likewise descriptive results for the frozen benchmark and the reported access windows.

In practice, the results support a hybrid workflow. Specialized CV systems can handle established detection and localization stages, while MLLMs can be assessed for terminology grounding and struc-

Evaluation signal	What it indicates	Suggested action
Strong Task A, weak Tasks B/C/D, or low LCS ($p = 0.5$)	Capability is uneven across downstream review tasks	Use only for prescreening; retain human review downstream
CMD macro-F1 rises from 0.480 to 0.604	Terminology grounding is a major bottleneck, not a purely visual error	Add definitions or domain adaptation first
Best Task C DICU is only 0.492	Localization is not yet reliable enough for review	Add a dedicated localizer or require human confirmation
On the same 250-response subset, mean EFS is 0.755 but the best Task D score is only 0.216	Visible evidence does not guarantee correct attribution	Keep cause outputs as candidates, not automated decisions

Table 2: Representative mappings from diagnostic signals to deployment actions.

tured language outputs. LDU-Bench determines whether the latter are reliable enough and where they should enter the workflow.

4 Conclusion

Task-level scores, diagnostic experiments, and LCS together give a consistent conclusion: current MLLMs show a systematic capability-structure gap in lithography review tasks. This is precisely where the value of LDU-Bench lies. Its value is not in building a more complex defect recognition benchmark, but in abstracting the review process into a quantifiable and diagnostic multi-level capability evaluation framework, thereby characterizing model behavior differences across semantic levels. The experimental results show that although existing commercial MLLMs already have some anomaly-triage ability, this ability does not naturally extend to stable structured review understanding. This phenomenon indicates that current models still have a systematic break in the continuous modeling process from visual perception to semantic attribution, rather than a lack of ability in a single module.

Therefore, the essence of the problem is not limited to a bottleneck in a specific task, but is an overall alignment problem across visual perception, spatial representation, and causal reasoning.

Limitations

LDU-Bench has several controlled scope boundaries that arise from building a publicly releasable lithography-review benchmark. First, its morphology labels follow a fixed benchmark taxonomy designed to support consistent evaluation across all tested models. Independent blind labeling shows that this taxonomy provides an auditable basis for Task B. Nevertheless, lithography review practices may vary across fabs, equipment conditions, and

annotation protocols. We therefore use Task B to evaluate whether models can align visual defect morphology with a consistent set of review labels, while leaving broader cross-fab taxonomy harmonization to future work. Second, Task D focuses on image-conditioned cause attribution. Since process logs, equipment states, recipe parameters, and historical inspection records are often sensitive and difficult to release in a public benchmark, we currently evaluate whether visible defect cues can be matched to reviewed cause references, instead of attempting full process-level causal diagnosis. Future extensions can incorporate fab-specific taxonomies, process metadata, temporal records, and human-in-the-loop feedback under industrial collaboration, enabling evaluation closer to closed-loop lithography diagnosis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62476092 and the Science and Technology Commission of Shanghai Municipality under Grant No. 24JD1401300.

References

- Kilian Batzner, Lars Heckler, and Rebecca König. 2024. [Efficientad: Accurate visual anomaly detection at millisecond-level latencies](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 127–137.
- Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. 2022. [Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization](#). *International Journal of Computer Vision*, 130:947–969.
- 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.
- Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Vic De Ridder, Bappaditya Dey, Enrique Dehaerne, Sandip Halder, Stefan De Gendt, and Bartel Van Waeyenberge. 2023. [Semi-centernet: A machine learning facilitated approach for semiconductor defect inspection](#). In *38th European Mask and Lithography Conference (EMLC 2023)*, volume 12802, page 128020M. SPIE.
- Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. [Padim: A patch distribution modeling framework for anomaly detection and localization](#). In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 475–489. Springer.
- Enrique Dehaerne, Bappaditya Dey, Victor Blanco, and Jesse Davis. 2025. [Scanning electron microscopy-based automatic defect inspection for semiconductor manufacturing: A systematic review](#). *Journal of Micro/Nanopatterning, Materials, and Metrology*, 24(2):020901.
- Bappaditya Dey, Enrique Dehaerne, Sandip Halder, Philippe Leray, and Magdy A. Bayoumi. 2022. [Deep learning based defect classification and detection in sem images: A mask r-cnn approach](#). In *Metrolgy, Inspection, and Process Control XXXVI*, volume PC12053, page PC120530K. SPIE.
- Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024. [Anomalygpt: Detecting industrial anomalies using large vision-language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1932–1940.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.
- Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616.
- Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. 2025. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *International Conference on Learning Representations*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.
- Lufeng Liao, Sikun Li, Yongqiang Che, Weijie Shi, and Xiangzhao Wang. 2022. [Lithography hotspot detection method based on transfer learning using pre-trained deep convolutional neural network](#). *Applied Sciences*, 12(4):2192.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR.
- 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.
- Moojoon Shin and Jee-Hyong Lee. 2016. Cnn based lithography hotspot detection. *International Journal of Fuzzy Logic and Intelligent Systems*, 16(3):208–215.
- Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. 2024. [Real-ia-d: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892.
- Haoyu Yang, Yajun Lin, Bei Yu, and Evangeline F. Y. Young. 2017. [Lithography hotspot detection: From shallow to deep learning](#). In *2017 30th IEEE International System-on-Chip Conference (SOCC)*, pages 233–238. IEEE.
- Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. 2021. *Draem: A discriminatively trained reconstruction embedding for surface anomaly detection*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. 2024. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*.

Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Junran Wu. 2024. *Do llms understand visual anomalies? uncovering llm’s capabilities in zero-shot anomaly detection*. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 48–57.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. *arXiv preprint arXiv:2207.14315*.

A Reproducibility and Model Identification

This appendix reports model identities, access windows, inference controls, and diagnostic seeds for the frozen evaluation.

A.1 Model Identifiers and Access Windows

Commercial-model results correspond to the access windows in Table 3. Temperature was set to 0 when exposed by the interface, and all per-sample predictions and scorer outputs were archived.

A.2 Inference Controls and Scoring

The following tables report the model-specific decoding and runtime settings, followed by the shared prompting, parsing, and deterministic scoring controls.

A.3 Diagnostic Sampling

Fixed sampling seeds are 20260602 for CMD, 20260604 for the EFS Task D pool, 20260608 for the random geometric localization baseline, and 20260613 for the final Task B blind-relabeling sheet.

B Dataset, Annotation, and Release

B.1 Annotation and Diagnostic Audits

Task B labels were audited through independent blind relabeling by a professional with semiconductor defect-review experience. A 100-sample BEOL subset stratified by morphology class (seed

20260613) covered six classes and yielded 79.0 percent exact agreement with Cohen’s kappa of 0.727. Task D uses professionally reviewed references and a frozen rubric: canonical or alias matches score 1, same-supergroup matches score 0.5, and other matches score 0; the final score combines semantic matching and keyword F1 with weights 0.7 and 0.3. Task D is scoped to benchmark-local image-conditioned attribution, with independent cause-label agreement treated as a separate taxonomy-validation target.

EFS audits whether a Task D rationale stays grounded in visible image evidence. For response i , the diagnostic is

$$\text{EFS}_i = \text{Rel}_i \text{Con}_i (1 - \text{Hall}_i),$$

where relevance and consistency take values in $\{0, 0.5, 1\}$ and hallucination takes values in $\{0, 1\}$. The pilot contains 50 BEOL Task D samples across four cause classes and five commercial models, giving 250 responses scored by one annotator. A second annotator scored a model-blind, difficulty-enriched subset of 100 responses. After harmonizing the binary hallucination coding, exact agreement was 93 percent for relevance, 48 percent for consistency, and 84 percent for hallucination. EFS is reported as an exploratory audit, separate from the leaderboard and LCS.

B.2 Authorization, De-identification, and Release

The source images are proprietary industrial inspection data. The data owner authorized their research use and public release as a de-identified benchmark. Source identifiers are replaced by anonymized benchmark IDs, and the de-identification protocol removes fab, equipment, batch, and sample identifiers, recipes, process logs, equipment states, and other sensitive fields. The benchmark contains no personal data.

The de-identified benchmark is publicly available at <https://github.com/0wen1103/LDU-bench> (release v1.0). The release archive contains images, task manifests and labels, and one pixel-level mask annotation for each of the 984 Task C instances. It also includes the frozen Task D rubric and documentation, while the repository provides prediction schemas and a deterministic scorer. Dataset use is limited to non-commercial research evaluation under the repository’s Research-Only Data License; the evaluation code is MIT-licensed.

Evaluation label	Identifier / implementation	Model developer	Reported access window	Evaluated tasks
GPT-5.4	gpt-5.4	OpenAI	2026-04-28 to 2026-06-08	A/B/C/D
Claude Opus 4.6	claude-opus-4-6	Anthropic	2026-05-06 to 2026-06-08	A/B/C/D
Qwen3.6-Plus	qwen3.6-plus	Alibaba	2026-04-23 to 2026-06-08	A/B/C/D
GLM-5V-Turbo	GLM-5V-Turbo	Zhipu AI	2026-06-07 to 2026-06-08	A/B/C/D
MiniMax-M3	MiniMax-M3	MiniMax	2026-06-05 to 2026-06-08	A/B/C/D
Gemma 4 31B-it	google/gemma-4-31B-it	Google DeepMind	2026-05-27 to 2026-06-08	A/B/C/D
AnomalyGPT	AnomalyGPT public implementation	AnomalyGPT authors	2026-06-01 to 2026-06-08	A/C

Table 3: Reported model identifiers and access windows for the evaluated snapshot.

Model	Recorded decoding and runtime controls
GPT-5.4	Interface-default decoding.
Claude Opus 4.6	Temperature 0; 512-token cap.
Qwen3.6-Plus	Temperature 0; 1,024-token cap.
GLM-5V-Turbo	Temperature 0; 1,024-token cap; coordinates returned in a normalized 1024×1024 space are rescaled to the native image size before scoring.
MiniMax-M3	Temperature 0; 512-token cap.
Gemma 4 31B-it	do_sample=false; 512 new-token cap; fixed local checkpoint; 4-bit NF4 quantization with bfloat16 compute; device_map=auto.
AnomalyGPT	Fixed seed 20260607 for the local A/C evaluation; the mask-level AD-Seg summary uses seed 20260608 and a 224×224 map.

Table 4: Model-specific decoding and runtime controls retained in the evaluation records.

Item	Frozen setting
Prompt and schema	Each task uses a fixed prompt template and task-specific output schema. Structured answers are required whenever deterministic parsing is part of the metric.
Image input	The original task image files are used. Model-specific input adapters encode the hosted-model inputs; the local Gemma adapter uses the checkpoint processor defaults.
Invalid outputs	Invalid JSON, illegal labels, missing answers, and unparsable coordinates remain in the denominator and receive the task-specific zero or incorrect outcome.
Task A/B	Deterministic label normalization; macro-F1 is the main metric and accuracy is auxiliary.
Task C	Coordinates are parsed as xyxy boxes. Invalid boxes receive zero DICU; valid boxes are scored from IoU and ground-truth coverage.
Task D	A frozen rubric applies deterministic label, alias, supergroup, and keyword matching. All main A/B/C/D metrics are produced by deterministic scorers.
EFS	Human-scored exploratory diagnostic for Task D explanations, reported separately from model ranking and LCS.

Table 5: Shared inference and scoring controls.

C Additional Results and Evaluation Scope

C.1 Sibling Models and Baseline Coverage

The main table reports one current representative per commercial model developer, and Table 6 adds four sibling models on Tasks A and B. Each sibling scores lower on Task B than on Task A, supporting the same cross-task pattern across the evaluated checkpoints.

Sibling model	Task A	Task B
GPT-5.4-mini	0.824	0.179
Claude Sonnet 4.6	0.783	0.195
Qwen3-VL-Plus	0.816	0.273
GLM-5.1	0.580	0.077

Table 6: Macro-F1 for sibling-model checks on Tasks A and B.

AnomalyGPT provides the specialized Task C

baseline, while the geometric boxes test whether DICU penalizes high defect coverage obtained through excessive background. DICU balances compact overlap and defect coverage to provide an interpretable localization readout.

C.2 Metric Sensitivity

Task D gives semantic label matching the primary weight and uses keyword F1 as supporting evidence. We recomputed the frozen outputs with semantic weight $\alpha \in \{0.5, 0.7, 0.9\}$, where the score is $\alpha \text{ Sem} + (1 - \alpha) \text{ KeyF1}$. The ordering is unchanged across all three settings.

We also hold the four task scores fixed and vary the generalized-mean parameter $p \in \{0.25, 0.5, 0.75, 1.0\}$. Smaller values place more emphasis on weak tasks, while $p = 1$ is the arithmetic mean. The ordering is unchanged across all four settings.

Evaluated model / baseline	A	B	C	D	Role in the evaluation
Five commercial MLLMs	Yes	Yes	Yes	Yes	General models evaluated on all four tasks.
Gemma 4 31B-it	Yes	Yes	Yes	Yes	Open-weight model evaluated on all four tasks.
AnomalyGPT	Yes	–	Yes	–	Specialized anomaly-detection baseline for applicable tasks.
Full-image and center-half boxes	–	–	Yes	–	Geometric sanity checks for large-box coverage effects.
Random geometric box	–	–	Yes	–	Seeded geometric sanity check in the appendix protocol.

Table 7: Task coverage of the evaluated models and baselines.

Model	$\alpha = .5$	$\alpha = .7$	$\alpha = .9$
GPT-5.4	.386	.409	.432
MiniMax-M3	.366	.395	.424
Qwen3.6-Plus	.344	.376	.407
GLM-5V-Turbo	.288	.297	.306
Claude Opus 4.6	.244	.255	.266
Gemma 4 31B-it	.073	.080	.087

(a) Task D semantic weight α .

Model	$p = .25$	$p = .50$	$p = .75$	$p = 1$
GPT-5.4	.4573	.4743	.4914	.5085
GLM-5V-Turbo	.4511	.4619	.4735	.4857
MiniMax-M3	.4254	.4394	.4543	.4700
Qwen3.6-Plus	.3950	.4082	.4223	.4373
Claude Opus 4.6	.3635	.3783	.3947	.4126
Gemma 4 31B-it	.1156	.1318	.1510	.1724

(b) LCS generalized-mean parameter p .

Table 8: Sensitivity analyses over frozen task scores.

C.3 Uncertainty over Frozen Outputs

We report 95 percent percentile bootstrap confidence intervals from 2,000 nonparametric resamples of the frozen per-sample outputs (seed 20260826). Each resample recomputes macro-F1 for Tasks A and B and the mean score for Tasks C and D. These intervals quantify finite-sample uncertainty in the frozen outputs.

C.4 Output Validity

Scoring retains invalid JSON, illegal labels, missing answers, and unparsable coordinates in the denominator and assigns the task-specific zero or incorrect score. Task-level reliability therefore incorporates format compliance.

Model	Task A	Task B	Task C	Task D
GPT-5.4	.932 [.920, .943]	.201 [.181, .220]	.492 [.470, .512]	.409 [.373, .442]
GLM-5V-Turbo	.884 [.869, .899]	.408 [.385, .430]	.353 [.331, .376]	.297 [.265, .329]
MiniMax-M3	.909 [.895, .922]	.239 [.217, .259]	.338 [.320, .356]	.395 [.357, .431]
Qwen3.6-Plus	.856 [.839, .872]	.249 [.227, .268]	.269 [.252, .286]	.376 [.338, .409]
Claude Opus 4.6	.876 [.860, .891]	.266 [.242, .288]	.254 [.237, .270]	.255 [.225, .285]
Gemma 4 31B-it	.504 [.480, .526]	.060 [.056, .064]	.046 [.040, .053]	.080 [.061, .102]

Table 9: Observed task scores and bootstrap 95% confidence intervals over frozen outputs.

Model	Task A valid	Task B valid	Task C valid	Task D valid	Overall valid
GPT-5.4	1761/1761	1761/1761	984/984	532/532	100.00%
Claude Opus 4.6	1761/1761	1761/1761	984/984	532/532	100.00%
Qwen3.6-Plus	1761/1761	1761/1761	982/984	532/532	99.96%
MiniMax-M3	1761/1761	1761/1761	984/984	532/532	100.00%
GLM-5V-Turbo	1761/1761	1761/1761	984/984	532/532	100.00%
Gemma 4 31B-it	1761/1761	1761/1761	951/984	532/532	99.34%
AnomalyGPT	1761/1761	–	984/984	–	100.00% on A/C

Table 10: Output validity and parse rates under the frozen scorer.