Interpretable Logical Anomaly Classification via Constraint Decomposition and Instruction Fine-Tuning

Xufei Zhang*, Xinjiao Zhou, Ziling Deng, Dongdong Geng, Jianxiong Wang
Beijing XingYun Digital Technology Co., Ltd., Beijing, China

***Abstract*—Logical anomalies are violations of predefined constraints on object quantity, spatial layout, and compositional relationships in industrial images. While prior work largely treats anomaly detection as a binary decision, such formulations cannot indicate which logical rule is broken and therefore offer limited value for quality assurance. We introduce Logical Anomaly Classification (LAC), a task that unifies anomaly detection and fine-grained violation classification in a single inference step. To tackle LAC, we propose *LogiCls*, a vision–language framework that decomposes complex logical constraints into a sequence of verifiable subqueries. We further present a data-centric instruction synthesis pipeline that generates chain-of-thought (CoT) supervision for these subqueries, coupling precise grounding annotations with diverse image-text augmentations to adapt vision language models (VLMs) to logic-sensitive reasoning. Training is stabilized by a difficulty-aware resampling strategy that emphasizes challenging subqueries and long tail constraint types. Extensive experiments demonstrate that *LogiCls* delivers robust, interpretable, and accurate industrial logical anomaly classification, providing both the predicted violation categories and their evidence trails.**

***Index Terms*—Logical Anomaly, Anomaly Classification, Data Synthesis**

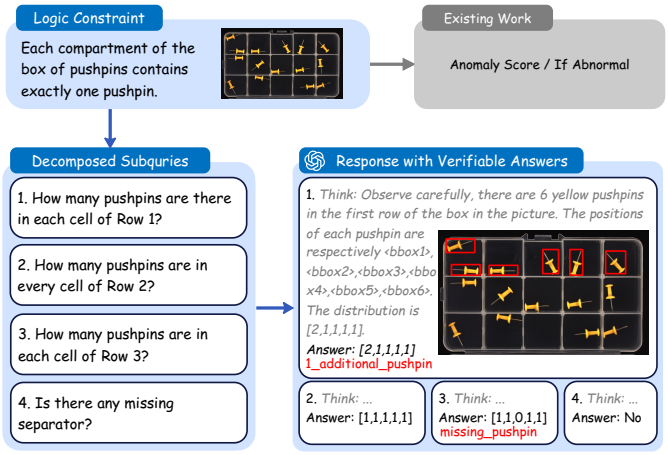

Fig. 1. Comparison of different approaches to logical anomalies. Existing methods typically output an anomaly score or a normal abnormal binary decision, whereas our setting aims to explicitly identify normal samples and classify specific logical violation types.

I. Introduction

Anomaly detection and classification [1]–[4] are indispensable for industrial visual inspection, directly affecting yield, rework cost, and downstream process stability. Industrial anomalies can be broadly categorized into structural and logical. Structural anomalies such as cracks, scratches, or contamination typically manifest as localized appearance defects and have been extensively studied with representation learning and reconstruction-based pipelines. Logical anomalies, in contrast, arise from violations of high level constraints on quantity, spatial arrangement, and composition, where evidence is often subtle, globally distributed, and only meaningful under a rule-based interpretation of the scene.

This constraint-centric view also appears in recent controllable generation and editing research, which increasingly emphasizes explicit constraint satisfaction such as consistent object quantity and layout [5], subject consistent transformations under user intent [6], and fine-grained controllable garment design [7]. Related studies on pose-guided generation further highlight that global structure and spatial configuration must be treated as first class constraints rather than incidental appearance cues [8]. These advances suggest that high level constraints are central to modern visual reasoning and controllability. For industrial inspection, the corresponding requirement is to verify such constraints from observations under data scarcity and subtle visual cues, which makes logical anomaly understanding substantially more challenging than structural defect recognition.

Research on structural anomaly detection and classification is well established [9]–[13]. Logical anomalies were introduced as a distinct benchmark direction with MVTec LOCO [14]. While recent work begins to study logical anomaly detection, fine-grained classification of logical violations remains largely underexplored. From a modality perspective, existing approaches fall into two lines. The first relies primarily on the visual modality, which is necessary but often insufficient to capture compositional constraints and long range dependencies. The second leverages multimodal feature matching to model attributes, entity relations, and higher order compositions. More recently, training free systems have emerged that concatenate multiple base models and query powerful vision language models to reach a final decision [15], [16]. Although effective, such multi-component pipelines can be operationally heavy, sometimes depending on online services, which leads to complex workflows and high

* Corresponding author. Email: zhang.xufei@xydigit.com

deployment costs.

To bridge this gap, we introduce Logical Anomaly Classification, which determines whether an image contains a logical anomaly and identifies its specific violation category. To better reflect real world compositions, we reconstruct a dataset such that the training split contains normal samples and single anomaly samples, while the test split includes images with multiple coexisting logical anomalies. Addressing Logical Anomaly Classification with a compact vision language model presents several challenges. Logical structures are governed by multiple interdependent constraints. Many violations are subtle and sparsely represented in the training split. Small-scale models have limited spatial and numerical reasoning ability, and the dataset scale and diversity are constrained.

We propose *LogiCls*, a framework that couples fine-grained synthetic data construction with instruction fine-tuning to endow compact vision language models with verifiable logical reasoning. The key idea is to decompose complex constraints into atomic and verifiable subqueries grounded in the image, each designed to map precisely to a specific anomaly category. We then construct an instruction dataset that integrates object grounding, chain of thought reasoning, and image text augmentation, encouraging the model to produce structured and interpretable outputs. During training, a difficulty-aware resampling strategy adjusts sampling probabilities based on prediction errors and uncertainty, improving performance on hard cases. At inference, the compact model answers the subqueries and aggregates their results into a final anomaly decision along with transparent evidence trails.

In summary, our contributions are:

- We define Logical Anomaly Classification and present *LogiCls*, which decomposes complex logical constraints into atomic subqueries for interpretable and verifiable reasoning.
- We develop a fine-grained data synthesis pipeline with object-grounded reasoning templates and introduce a difficulty-aware resampling strategy that adapts sampling based on model errors and uncertainty.
- *LogiCls* with small-scale vision language models outperforms strong baselines in accuracy, improves inference efficiency, and provides transparent reasoning chains that enhance interpretability and robustness.

II. Task Definition and Dataset

Formulation We formulate the logical anomaly classification (LAC) task as an end-to-end set prediction problem. During training, the model is provided with images labeled as either normal or associated with a single anomaly type from a predefined finite set $\mathcal{C} = \{c_1, c_2, ..., c_K\}$, where each c_k denotes a distinct anomaly category. During inference, the model is expected to predict a subset of the extended label space $\mathcal{C}^+ = \{\text{normal}\} \cup \mathcal{C}$, such that the output is an element of the power set $\mathcal{P}(\mathcal{C}^+)$, i.e., a subset of $\mathcal{C}^+$. Specifically, if no anomalies are detected, the predicted set is $\{\text{normal}\}$. Otherwise, it contains the normal label excluded and one or more anomaly types indicating the detected defects. Let $\mathcal{X}$ denote the space of input images. The goal is then to learn a mapping function $f : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{C}^+)$, which generalizes from training samples with single-label annotations (either normal or a single anomaly type) to real-world scenarios where multiple anomaly types may coexist or none can be present.

TABLE I
Distribution of MVTec LOCO FC dataset.

Scenario	Training Set		Test Set		
	Normal	Single Anomaly	Normal	Single Anomaly	Multi Anomaly
Breakfast Box	351	66	102	17	0
Juice Bottle	335	104	94	26	12
Screw Bag	360	90	122	23	24
Pushpins	372	32	138	48	11
Splicing Connectors	360	86	119	22	0

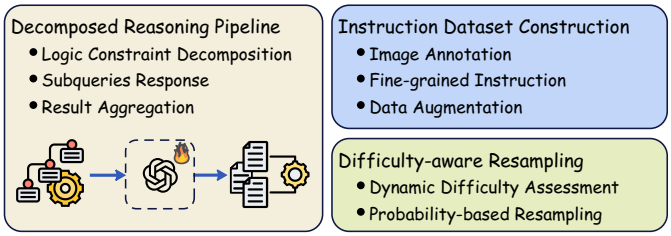

Fig. 2. Overview of the proposed *LogiCls* framework. The method first decomposes complex logical constraints into atomic subqueries. Then, an instruction dataset is constructed through image annotation, fine-grained reasoning templates, and data augmentation. During training, a difficulty-aware resampling strategy dynamically adjusts sampling probabilities. Finally, a fine-tuned model performs inference by aggregating subquery outputs for accurate logical anomaly classification.

MVTec LOCO FC Dataset. The MVTec LOCO [14] dataset is the most popular industrial logic anomaly dataset and contains five scene categories. The training set originally included only normal images, whereas the testing set included both normal and abnormal images. This dataset is widely used in LAD research, and to our knowledge, the classification task has never been studied on it. Following the methodology of MVREC [11], we utilize the provided masks to categorize anomalies and resplit the data into *MVTec LOCO FC* (Tab. I). For anomalous samples that are only in the original test set, 80% are randomly sampled into the new training set and 20% into the new test set. All samples with multiple logical anomalies are manually placed in the new test set. Samples containing multiple co-occurring logical anomalies are placed exclusively in the test set to evaluate the model's ability to handle complex, real-world scenarios. For normal samples, the new dataset is divided in the same way as the original dataset.

III. Method

As illustrated in Fig. 2, we propose *LogiCls*, a framework that decomposes complex industrial logical constraints into tractable subqueries and performs anomaly classification by aggregating verifiable subquery answers. The framework follows a data-driven strategy: we synthesize instruction data to enhance spatial and counting ability in small-scale VLMs, and we apply difficulty-aware resampling during fine-tuning for robust classification.

A. Logic Constraint Decomposition

In industrial logical anomaly scenarios, classification requires object recognition together with reasoning over constraints on spatial relations, quantity, and attribute consistency across multiple objects. Direct end-to-end inference to get anomaly types $\hat{y} \in \mathcal{P}(\mathcal{C}^+)$ with a VLM g_θ on an image $I \in \mathcal{X}$ with corresponding logic constraint text L often yields suboptimal accuracy and limited interpretability:

$$\hat{y} = g_\theta(I, L). \tag{1}$$

We therefore decompose L into atomic subqueries $\{q_t\}_{t=1}^T$, each targeting a specific feature or violation pattern. For each subquery, we obtain

$$z_t = g_\theta(I, q_t), \quad t = 1, \ldots, T, \tag{2}$$

and then map the set of answers $\{z_t\}_{t=1}^T$ to final anomaly categories with a scenario-specific aggregator h

$$\hat{y} = h(z_1, \ldots, z_T). \tag{3}$$

Each subquery returns a verifiable output such as a numeric value $z_t \in \mathbb{R}$, a boolean value in $\{0, 1\}$, or a categorical element from a finite set. The aggregation in Eq. 3 provides a reasoning trail and a precise mapping to categories $\hat{y} \in \mathcal{P}(\mathcal{C}^+)$.

B. Fine-grained CoT Instruction Data

Since logical anomalies hinge on spatial and counting constraints and small-scale VLMs underperform in such settings, we build a three-step instruction synthesis pipeline. **(1) Image annotation.** Following LogSAD [15], we apply CLIP [17] and SAM [18] for open-vocabulary segmentation [19] on MVTec LOCO FC to obtain masks and location coordinates. **(2) Fine-grained instruction.** For each image I and subquery q_t, we define $g_\theta(I, q_t) \to (c_t, z_t)$, where c_t is a CoT enclosed by `<think>` and `</think>` with grounding to bounding boxes, and z_t is the final answer enclosed by `<answer>` and `</answer>`. We refine reasoning into direction, distance, size, counting, and other types. Examples are provided in Fig. 1 and Fig. 4, and the collection forms D_{cot}. **(3) Data augmentation.** To enlarge D_{cot}, we build D_{aug} with modality specific strategies. For direction and counting, we cut and paste masked objects with Poisson image editing [20] and update the labels. For other types, we add noise and apply style transfer. For text, large language models produce between 10 and 20 paraphrases per subquery to increase linguistic diversity.

C. Difficulty-aware Resampling

We introduce a dynamic resampling scheme that focuses training on challenging subqueries by adjusting sampling probabilities according to an online estimate of difficulty.

Dynamic difficulty assessment. At the end of each epoch, the model evaluates the difficulty of every sample in D_{cot}. For sample i, we compute a score

$$d_i \;=\; \alpha \cdot \mathbf{1}[\hat{y}_i \neq y_i] \;+\; \beta \cdot \text{Perplexity}(\hat{y}_i), \tag{4}$$

```
System Prompt:
- Task: You need to answer questions
based on the image uploaded.
- Format: Your response should include
a concise reasoning process and an
accurate answer, with the reasoning and
answer enclosed separately using
<think> and </think>, and <answer> and
</answer> tags, for example:
<think>This is the reasoning
process</think><answer>This is the
answer</answer>.
User Prompt: [SUBQUERY]
Case:
- Input Image: [Example Image]
- Output: [Reasoning phase and answer]
On the test image, provide a concise
reasoning and the corresponding answer.
Test:
- Input: [Test Image]
- Output:
```

Fig. 3. Illustration of ICL. The content inside the brackets will be replaced with real samples during the experiments.

where $\mathbf{1}[\cdot]$ is the indicator function and $\alpha = 1$, $\beta = 0.2$ in all experiments.

Probability-based resampling. Let $\mathcal{Q}_t$ be the set of augmented samples in $\mathcal{D}_{\text{aug}}$ derived from base subquery q_t. We assign a sampling probability

$$P_s^{(t)} = \frac{\left(\sum_{i \in \mathcal{Q}_t} d_i\right)^\gamma}{\sum_{l=1}^T \left(\sum_{i \in \mathcal{Q}_l} d_i\right)^\gamma}, \tag{5}$$

where $\gamma > 0$ controls how strongly difficulty influences sampling. Minibatches are drawn from $\mathcal{D}_{\text{aug}}$ according to $\{P_s^{(t)}\}_{t=1}^T$, which increases exposure to hard subqueries while retaining coverage of easier ones.

IV. Experiment

A. Settings

Setup. As the first study of LAC, we evaluate on MVTec LOCO FC. All inputs use native resolution without resize or crop. Following prior work [21], we adopt a unified in-context learning (ICL) format that integrates decomposed subqueries, a single positive training image, the test image, and explicit instructions for both reasoning and output. This setup provides a consistent scaffold for the model to infer step-by-step from the exemplars and apply the same reasoning to the test case, as shown in Fig. 3. Baselines are Gemini-2.5-pro [22], GPT4.1-2025-0414[1], GPT-4o [23], GLM-4.5V [24], InternVL3-78B-Instruct [25] and QwenVL2.5-72B, 7B, 3B-Instruct [26]. We also fine-tune 3B and 7B models with *LogiCls*.

Metrics. Images may be normal or carry multiple anomaly types, so we report two measures. *Binary F1* collapses all anomaly types into a single anomaly class. *Macro F1* averages the F1 score over all anomaly classes to reflect class-balanced performance.

[1]https://openai.com/index/gpt-4-1/

TABLE II
THE RESULTS OF MVTEC LOCO FC INCLUDE *binary F1* AND MULTI-CLASSIFIED *macro F1* FOR LOGIC ANOMALIES.

Model	Method	Breakfast Box		Juice Bottle		Screw Bag		Pushpins		Splicing Connector		Average	
		binary F1	macro F1	binary F1	macro F1	binary F1	macro F1	binary F1	macro F1	binary F1	macro F1	binary F1	macro F1
Gemini-2.5-pro [22]	ICL	93.45	89.74	85.56	39.86	72.45	50.87	40.87	32.58	42.39	48.64	64.23	49.89
GPT4.1-2025-0414	ICL	90.75	83.46	81.25	35.48	63.50	42.78	43.87	30.65	43.16	51.45	61.98	45.91
GPT-4o [23]	ICL	86.67	61.82	81.25	35.48	43.48	28.44	39.71	27.67	35.90	42.86	54.45	37.39
GLM-4.5V$_{no\ think}$ [24]	ICL	55.46	65.05	57.36	49.74	44.67	41.24	35.85	38.26	53.33	44.05	47.89	46.68
QwenVL2.5-72B-Instruct [26]	ICL	52.00	68.03	50.35	35.54	41.44	8.19	32.39	19.14	45.16	40.74	42.99	31.25
InternVL3-78B-Instruct [25]	ICL	45.16	30.30	81.25	33.37	39.66	11.16	34.59	35.27	33.96	40.42	45.39	29.74
QwenVL2.5-7B-Instruct [26]	ICL	34.67	20.15	56.60	10.48	40.35	10.38	38.02	23.80	26.09	21.30	39.03	17.45
	LogiCls	**100.0**	**97.39**	**90.96**	**86.98**	**92.16**	**88.74**	**87.80**	**98.00**	**95.24**	**98.41**	**92.62**	**94.00**
QwenVL2.5-3B-Instruct [26]	ICL	21.92	3.51	13.79	0.00	40.72	8.53	34.20	13.89	27.16	3.18	28.86	6.65
	LogiCls	**100.0**	**92.16**	**81.45**	**75.44**	**89.52**	**85.03**	**79.01**	**76.89**	**95.24**	**94.07**	**88.09**	**84.05**

TABLE III
ABLATION STUDY ON COMPONENTS OF OUR FRAMEWORK.

Decomp.	Reason.	Bbox.	Resample.	SFT	Avg.binary F1	Avg.macro F1
✓	✗	✗	✗	✗	40.18	27.24
✓	✗	✗	✗	✓	45.34	24.20
✓	✓	✗	✗	✓	55.77	39.31
✓	✓	✓	✗	✓	72.37	51.62
✓	✓	✓	✓	✓	**92.62**	**94.00**

B. Main Results

We present *binary F1* and *macro F1* for LAC in Tab. II. First, focusing on the "Average" columns reveals that, without any training, large-scale VLMs generally outperform small-scale VLMs, and closed-source models perform better than open-source models. Our model performed the best, demonstrating that small-scale VLMs can be enhanced via our framework. Comparing the Breakfast Box and Screw Bag scenarios, we find that every method yields higher metrics for the former than for the latter. In the Breakfast Box, the types of logical anomalies mainly involve the combination of common objects (such as cereal, starfruit, oranges, etc.), while in the Screw Bag, the anomalies are all related to counting similar objects (such as screws, nuts, and washers of different lengths). This indicates that different types of logical anomalies present varying levels of difficulty for VLMs. Counting similar objects is more challenging than identifying the combination anomalies of common objects.

C. Ablation Study

We conduct ablation experiments on the MVTec LOCO FC dataset using the QwenVL2.5-7B-Instruct model [26]. The results, summarized in Tab. III, reveal the contribution of each component in our proposed method.

Constraint Decomposition. We first decompose the original logical constraints into subqueries and directly prompt the model. This experiment establishes the baseline for subsequent enhancements.

Supervised Fine-tuning. Next, we fine-tune the model using the subqueries, corresponding images, and their annotated answers. This step evaluates the effectiveness of fine-tuning beyond simple prompting. While it yields modest improvements, the gains are limited.

Distilled CoT. Building on fine-tuning, we incorporate reasoning CoTs into the model outputs. The CoTs are distilled from QwenVL2.5-72B-Instruct [26] but exclude bounding box information. This addition leads to clear improvements, particularly in *macro F1* $(39.31, +15.11)$, highlighting the benefit of explicit intermediate reasoning.

Fine-grained CoT. We further enhance the reasoning chain by embedding the spatial coordinates of the target objects. This spatial grounding produces significant performance gains $(72.37/51.62, +16.60/+12.31)$, confirming the necessity of aligning relational constraints with explicit object locations to ensure logical consistency.

Difficulty-aware Resampling. Finally, based on the fine-grained CoT dataset, we further improve the two indicators to $92.62/94.00$ $(+20.25/+42.38)$ by using image augmentation and difficulty-aware resampling (§III-C), reaching the optimum. This indicates that the long tail and the bias in the distribution of sample difficulty are effectively mitigated and that the model benefits from being balanced across categories.

In conclusion, the combination of spatial-aware reasoning and data balancing strategies constitutes the primary driver of the model's robust logical reasoning capabilities.

D. Case Study

We further conduct a fine-grained qualitative assessment to understand the specific benefits and limitations of our coordinate-aware reasoning approach. By analyzing the intermediate reasoning traces of our model alongside GPT4.1-2025-0414 and Gemini-2.5-pro, we highlight differences in spatial grounding capabilities. The comparison covers four distinct visual reasoning tasks.

Spatial Reasoning. In Fig. 4a, the cable connection task reveals significant disparities in spatial logic between the models. GPT4.1-2025-0414 fails the task by incorrectly claiming the red cable connects the first slot on both sides, leading to a false "Yes" answer. While Gemini-2.5-pro correctly identifies that the positions do not correspond, it misidentifies the specific slots. In contrast, our model accurately grounds the connection through precise coordinates, and correctly identifies the cable as being connected to the 3rd terminal on one

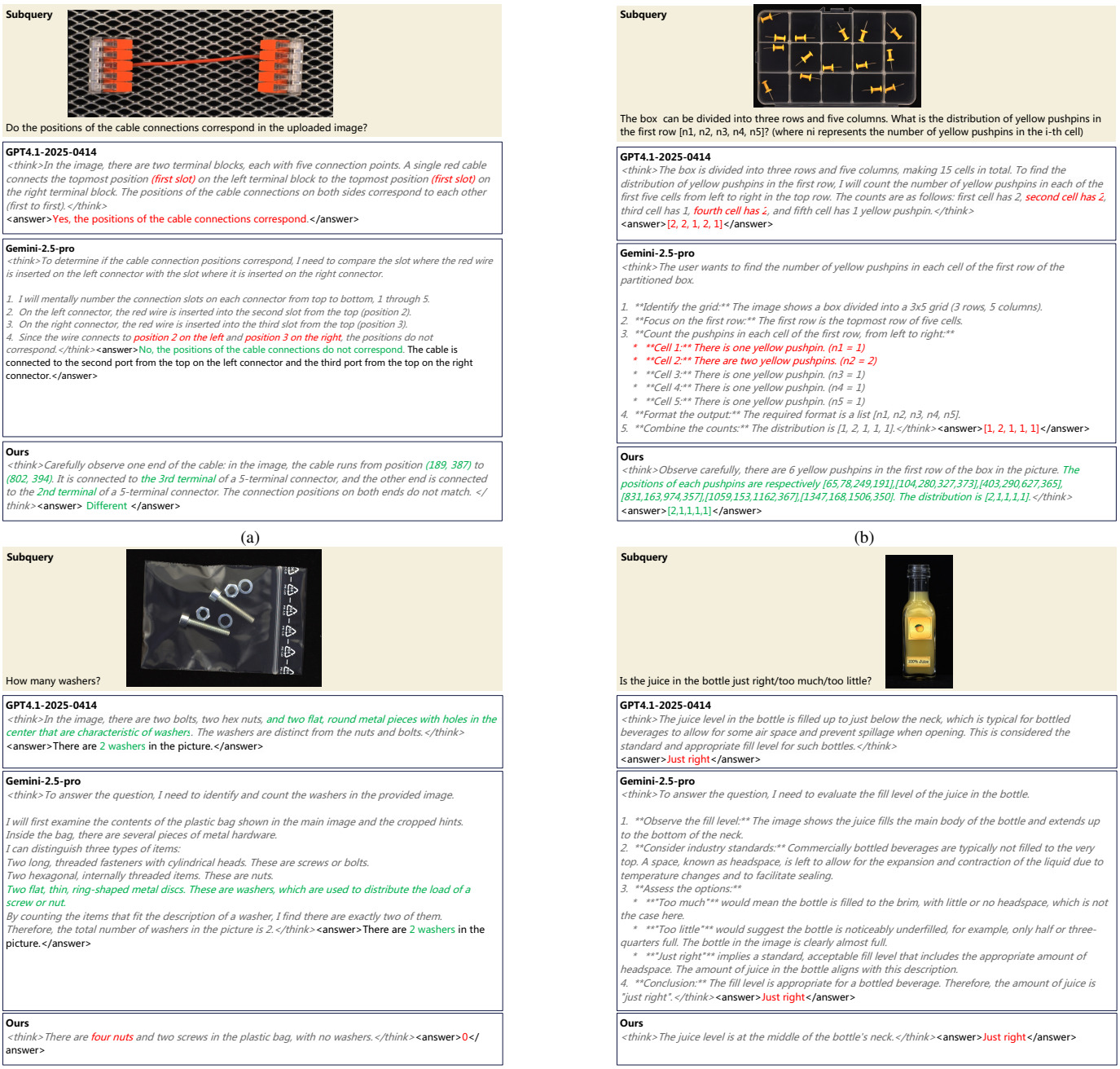

Fig. 4. A comparison with VLMs on the decomposed subqueries. Correct answers are in green and incorrect answers are in red.

end and the 2nd terminal on the other. This demonstrates that coordinate-level fine-grained CoT is essential for resolving complex spatial relationships that generic reasoning traces often misinterpret.

Dense Grid Counting. As illustrated in Fig. 4b, the task requires counting within specific grid partitions. The reasoning traces of baseline models, such as GPT4.1-2025-0414 and Gemini-2.5-pro, reveal counting discrepancies within individual cells. In contrast, our model explicitly generates spatial coordinates for each pushpin during the reasoning phase, allowing it to derive accurate cell-wise counts and the exact distribution. This mechanism results in correct cell-level counts and the exact final distribution, suggesting that spatial-aware fine-grained CoT is critical for mitigating counting hallucinations.

Object Identification. As illustrated in Fig. 4c, both GPT4.1-2025-0414 and Gemini-2.5-pro predict the correct answer, successfully distinguishing between visually similar objects (washers and nuts) within their reasoning traces. In contrast, our model failed to identify the washers, misclassifying them as nuts. This suggests that our current method does not significantly improve fine-grained discrimination for confusingly

similar targets, resulting in performance that falls short of the baseline models in this specific scenario.

Relative Scale Estimation. As shown in Fig. 4d, the fill-level assessment task, which requires precise relative scale estimation, reveals a collective limitation in current models. Both proprietary baselines and our model incorrectly conclude that the liquid level is "Just right", failing to perceive the subtle anomaly. This failure suggests that subtle relative concepts, as opposed to absolute ones, are inherently ambiguous. It highlights the limitations of current ICL and fine-tuning methodologies in addressing such subtleties.

Qualitative analysis demonstrates that *LogiCls* significantly enhances spatial grounding and dense counting accuracy compared to baselines. It effectively reduces hallucinations through precise coordinate-level CoT. However, the approach shows limited improvement in fine-grained object discrimination and subtle relative scale estimation, where it shares common pitfalls with existing models.

V. Conclusion

In this paper, we introduce the LAC task and propose a multi-model framework *LogiCls* to fine-tune small-scale VLMs. Our approach first takes a data-centric approach by using logic composition, constructing fine-grained CoT data, and performing data augmentation. Subsequently, we employ a difficulty-aware resampling strategy to improve the model's understanding of industrial scenarios. However, our approach struggles with visually similar categories and subtle scale anomalies. Future work will integrate hierarchical geometric priors and global-to-local attention to improve structural consistency and fine-grained scale sensitivity.

References

[1] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "Anomalygpt: Detecting industrial anomalies using large vision-language models," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, vol. 38, pp. 1932–1940.

[2] H. Li, R. Zhang, Y. Pan, J. Ren, and F. Shen, "Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[3] S. Zhang and J. Liu, "Feature-constrained and attention-conditioned distillation learning for visual anomaly detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 2945–2949.

[4] J. Zhu, Y.-S. Ong, C. Shen, and G. Pang, "Fine-grained abnormality prompt learning for zero-shot anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22241–22251.

[5] F. Shen, X. Du, Y. Gao, J. Yu, Y. Cao, X. Lei, and J. Tang, "Imagharmony: Controllable image editing with consistent object quantity and layout," *arXiv preprint arXiv:2506.01949*, 2025.

[6] F. Shen, W. Xu, R. Yan, D. Zhang, X. Shu, and J. Tang, "Imagedit: Let any subject transform," *arXiv preprint arXiv:2510.01186*, 2025.

[7] F. Shen, J. Yu, C. Wang, X. Jiang, X. Du, and J. Tang, "Imaggarment-1: Fine-grained garment generation for controllable fashion design," *arXiv preprint arXiv:2504.13176*, 2025.

[8] F. Shen and J. Tang, "Imagpose: A unified conditional framework for pose-guided person generation," *Advances in neural information processing systems*, vol. 37, pp. 6246–6266, 2024.

[9] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 128–138.

[10] Z. Zuo, Z. Wu, B. Chen, and X. Zhong, "A reconstruction-based feature adaptation for anomaly detection with self-supervised multi-scale aggregation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5840–5844.

[11] S. Lyu, R. Zhang, Z. Ma, F. Liao, D. Mo, and W. Wong, "Mvrec: A general few-shot defect classification model using multi-view region-context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 5937–5945.

[12] Z. Huang, X. Li, H. Liu, F. Xue, Y. Wang, and Y. Zhou, "Anomalyncd: Towards novel anomaly class discovery in industrial scenarios," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4755–4765.

[13] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, "Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.

[14] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *International Journal of Computer Vision*, vol. 130, no. 4, pp. 947–969, 2022.

[15] J. Zhang, G. Wang, Y. Jin, and D. Huang, "Towards training-free anomaly detection with vision and language foundation models," *arXiv preprint arXiv:2503.18325*, 2025.

[16] E. Jin, Q. Feng, Y. Mou, G. Lakemeyer, S. Decker, O. Simons, and J. Stegmaier, "Logicad: Explainable anomaly detection via vlm-based text feature extraction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 4129–4137.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[19] H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, and H. Pouransari, "Sam-clip: Merging vision foundation models towards semantic and spatial understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3635–3647.

[20] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.

[21] Y. Zhang, K. Zhou, and Z. Liu, "What makes good examples for visual in-context learning?," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17773–17794, 2023.

[22] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[23] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[24] W. Hong, W. Yu, X. Gu, G. Wang, G. Gan, H. Tang, J. Cheng, J. Qi, J. Ji, L. Pan, et al., "Glm-4.5v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning," *arXiv preprint arXiv:2507.01006*, 2025.

[25] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al., "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[26] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al., "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.