# AnomalyCoT: A Multi-Scenario Chain-of-Thought Dataset for Multimodal Large Language Models

**Jiaxi Cheng***, **Yuliang Xu***, **Shoupeng Wang***, **Ma Tao***, **Yuchen He***, **Jinghe Zhang**, **Sihang Cai**, **Jiawei Zhen**, **Jingyi Jia**, **Yao Wan**, **Yan Xia**, **Zhou Zhao**†
{jiaxicheng, zhaozhou}@zju.edu.cn

## Abstract

Industrial Anomaly Detection (IAD) is an indispensable quality control technology in modern production processes. Recently, on account of the outstanding visual comprehension and cross-domain knowledge transfer capabilities of Multimodal Large Language Models (MLLMs), existing studies have explored the application of MLLMs in the IAD domain and established some multimodal IAD datasets. However, although the latest datasets contain various fundamental IAD tasks, they formulate tasks in a general question-and-answer format lacking a rigorous reasoning process, and they are relatively limited in the diversity of scenarios, which restricts their reliability in practical applications. In this paper, we propose AnomalyCoT, a multimodal Chain-of-Thought (CoT) dataset for multi-scenario IAD tasks. It consists of 37,565 IAD samples with the CoT data and is defined by challenging composite IAD tasks. Meanwhile, the CoT data for each sample provides precise coordinates of anomaly regions, thereby improving visual comprehension of defects across different types. AnomalyCoT is constructed through a systematic pipeline and involves multiple manual operations. Based on AnomalyCoT, we conducted a comprehensive evaluation of various mainstream MLLMs and fine-tuned representative models in different ways. The final results show that Gemini-2.0-flash achieved the best performance in the direct evaluation with an accuracy rate of 59.6%, while Llama 3.2-Vision achieves the best performance after LoRA fine-tuning with an accuracy rate of 94.0%. Among all the fine-tuned models, the average accuracy improvement reaches 36.5%, demonstrating the potential of integrating CoT datasets in future applications within the IAD field. The source code and data are available at `https://github.com/Zhaolutuan/AnomalyCoT`.

## 1 Introduction

Industrial Anomaly Detection (IAD) has emerged as a crucial requirement in the modern industrial production processes. IAD facilitates the real-time monitoring and identification of product anomalies using automated techniques, thereby enabling timely intervention to prevent the propagation of errors and maintain product quality. Existing IAD methods exhibit high real-time performance and accuracy, effectively mitigating economic losses in industrial production and facilitating the transition towards unmanned operations [1, 2]. With the rise of Multimodal Large Language Models (MLLMs) , *i.e.* GPT-4 [3] and Gemini [4], which can undertake various human tasks, recent research has also attempted to apply MLLMs to IAD tasks. AnomalyGPT [5], as the first method to use MLLMs to solve IAD tasks, overcomes the limitation of most past IAD methods that require manual threshold setting for anomaly detection and is directly fine-tuned on IAD datasets. Furthermore, some studies utilize the joint modeling capacity of MLLMs and combine instruction fine-tuning methods [6, 7]

*Equal contribution.
†Corresponding Author.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

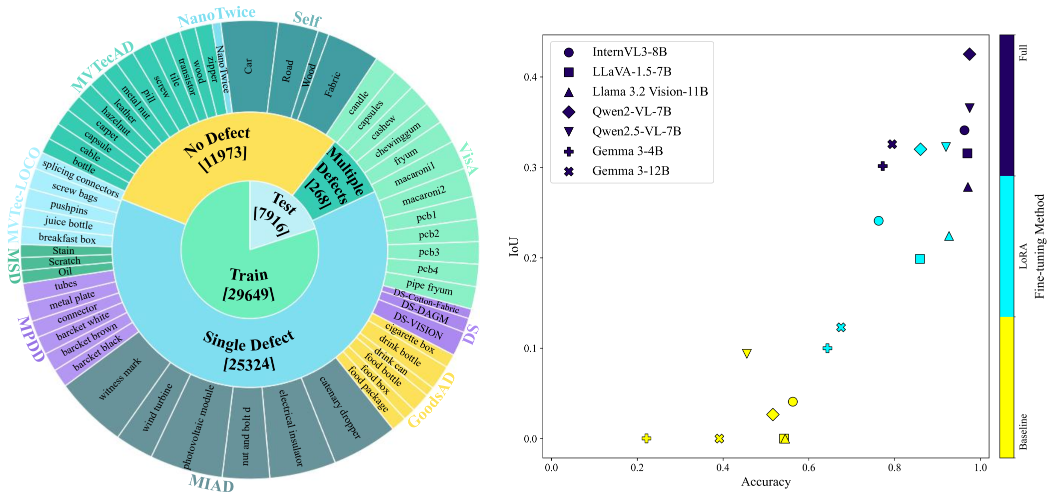



Figure 1: Left: The outermost layer shows the datasets that make up AnomalyCoT and the corresponding objects, the middle layer shows the number of different defects, and the innermost layer shows the division of the training set and the test set. Right: The performances of different MLLMs, including pre-trained, LoRA fine-tuned, and full fine-tuned models, are shown in this figure.

to solve IAD tasks using the visual question-answering paradigm, demonstrating the flexibility of MLLMs in task adaptation [8, 9].

In recent years, significant progress has been achieved in applying MLLMs to Industrial Anomaly Detection (IAD) tasks, concurrent with the development of a series of high-quality IAD datasets. MVTec AD [10] is one of the most influential datasets in the IAD field, not only containing over 5,300 high-resolution images of multiple objects exhibiting diverse defect types, but also providing pixel-precise annotations for all anomalies. This facilitates its application in both binary classification and anomaly detection tasks, thereby significantly advancing the development of IAD methods. MIAD [11] is a brand-new large-scale outdoor IAD dataset, containing over 100K high-resolution images and covering various types of structural and logical anomalies, providing strong support for the application of IAD tasks in **outdoor scenarios** and addressing the previous scarcity of outdoor anomaly samples. Other studies, such as MMAD [12], have proposed the first benchmark for evaluating the comprehensive performance of MLLM in IAD tasks. By designing multiple key subtasks of MLLM in anomaly detection and collecting public IAD datasets, a multimodal benchmark dataset in the form of question-and-answer for IAD was constructed. This benchmark specifically addresses the limited availability of dedicated resources for applying MLLMs within the industrial domain.

While the aforementioned datasets have made significant contributions to IAD research and application, they are subject to certain limitations. Firstly, MVTec AD and MIAD concentrate on manufacturing and maintenance inspection scenarios, respectively. This domain-specific focus and limited data diversity restrict their generalizability and direct applicability to broader real-world environments. Secondly, MMAD provides multi-angle IAD QA and general anomaly localization information, enabling the determination of sample status (normal/anomalous). Nevertheless, it does not offer a systematic analysis explaining the reasons for the observed normal or anomalous states. Additionally, since MMAD only employs a simple question-and-answer format, which lacks interpretability in the decision-making process and logical reasoning in intuitive representation, it becomes difficult to determine its reliability in practical applications. Overall, existing IAD datasets exhibit significant potential for expansion and require further refinement to enhance their applicability to real-world scenarios. Therefore, developing a systematic Chain-of-Thought (CoT) dataset is of particular importance, as it can facilitate the deployment of MLLM in IAD tasks within complex, dynamic environments.

In this paper, we introduce **AnomalyCoT**, the first multimodal CoT dataset for IAD tasks. To address the limitations of the current IAD datasets and facilitate the application of MLLM in real-world IAD tasks, our dataset offers three significant advantages. (i) The first advantage of the dataset lies in its **extensive coverage of IAD scenarios**. As illustrated in the left panel of Figure 1, we achieve substantial growth in both data scale and scenario diversity by systematically integrating and performing unified preprocessing on a wide range of publicly available IAD datasets; (ii) Another advantage of the dataset is that it unifies the two core tasks of IAD, namely anomaly discrimination and defect classification, to define an **end-to-end detection task**. Meanwhile, it includes the analysis process of a large number of normal or anomalous samples. We design the dataset with a question-reasoning-answer structure and provide no defect options in each sample's question options, enabling a complete analysis of normal or anomalous samples based on anomaly judgment; (iii) Finally, the most significant advantage of our dataset is its **interpretable reasoning process**, which provides precise coordinates of anomalous regions for anomalous samples to assist in analysis. To effectively address the ambiguity of decision-making basis in the simple question-and-answer form, we design and generate rigorous reasoning processes by combining accurate visual information of anomalous regions and the rich knowledge contained in large models, ensuring the reliability of the dataset in practical applications.

To fully manifest the aforementioned advantages of the dataset and exert the comprehensive capabilities of MLLM in addressing IAD tasks, we constructed a novel pipeline. Firstly, we collected the public datasets dedicated to studying different IAD tasks, forming a comprehensive IAD scenario. Subsequently, we designed a composite IAD task encompassing no defect options and distractors to achieve the synergetic completion of anomaly discrimination and defect classification. By **manually annotating** a large amount of data and precisely locating anomalous regions in the form of coordinate pairs, we have also designed a general semantic prompt embedded with rich visual information. Combined with QwenVL-Max, we generate a clear and logically rigorous reasoning process to establish a complete CoT data. Additionally, we carried out rule-based filtering and iterative updates on the CoT data. On this basis, all the data were manually verified to ensure the accuracy and rationality of the CoT. Finally, we collected 37,565 samples from 59 types of scenarios in 13 public datasets.

In the experiment, we first conducted a comprehensive evaluation of various mainstream MLLMs using AnomalyCoT, such as the open-source models Qwen2.5-VL [13] and InternVL3 [14], commercial models GPT-4o [15] and Gemini-2.0-flash, and the IAD model AnomalyGPT. We also perform different types of fine-tuning training on representative MLLMs like Qwen2.5-VL and choose Intersection over Union (IoU [16]) as the inference evaluation metric due to the post-training inference processes mostly including the coordinates of anomaly regions. The experimental results indicate that by adopting different fine-tuning methods on the major MLLMs, both the accuracy and IoU of the models are significantly improved compared to the case of direct evaluation, as shown in the right panel of Figure 1. This reflects the effectiveness of our dataset in the IAD task. Additionally, we have conducted thorough ablation experiments under various settings, which help to prove the reliability of the designed structure. Overall, our dataset has played a significant role in challenging major IAD tasks and demonstrates great potential for application in future IAD tasks with high-precision requirements.

Our contributions are summarized as follows:

- We construct AnomalyCoT, a new dataset for testing the comprehensive reasoning capabilities of mainstream MLLMs in the IAD task. To the best of our knowledge, our proposed dataset is **the first multimodal CoT dataset in the IAD task**. This dataset sets new requirements for the application logic of MLLMs in the industrial field.
- We have significantly expanded the scenarios of the IAD task and introduced a new pipeline for generating accurate and rigorous reasoning processes for anomaly detection tasks.
- We adopt structured CoT data and conduct fine-tuning experiments on representative MLLMs to comprehensively evaluate the performance of MLLMs on AnomalyCoT and achieve normative analysis of MLLMs in specific tasks.

## 2 Related Work

### 2.1 Industrial Anomaly Detection

Industrial anomaly detection is crucial for optimizing the overall quality of products and ensuring the healthy development of the industry. Traditional IAD research have primarily focused on locating and classifying defects in novel environments. Common IAD methods include reconstruction-based methods [17, 18], synthesis-based methods [19], and feature embedding-based methods

[20, 21, 22]. Reconstruction-based methods learn the reconstruction capability of normal samples and calculate the reconstruction error to detect anomalies. Synthesis-based methods employ data augmentation techniques to synthesize anomalous samples, transforming the original problem into a binary classification task distinguishing between normal and synthesized anomalous instances. Feature embedding-based methods model the feature distribution of normal samples and quantify the feature deviation of test samples. These methods typically require learning the distribution of a large number of samples of existing categories, making it difficult to learn new category instances in dynamic environments. Recent research has predominantly focused on performing IAD tasks with few-shot learning. Several studies integrated visual language models like CLIP, such as InCTRL [23], which employs a few positive samples as contextual prompts to learn the residual between test samples and prompts based on CLIP. AnomalyCLIP [24] learns object-agnostic textual prompts to capture different features within samples, focusing on anomalous regions. However, these models overemphasize predefined anomaly concepts, resulting in limited generalization in new scenarios. Given the flexibility of MLLMs in handling complex visual and textual inputs, addressing the aforementioned limitations becomes possible.

Some recent studies have focused on applying MLLMs to IAD tasks and yielding promising outcomes. MMAD has established an MLLM benchmark test encompassing seven key sub-tasks of IAD, conducting a comprehensive evaluation of various mainstream MLLMs. Other studies directly fine-tune MLLMs using public IAD datasets, such as AnomalyGPT and FabGPT [25], but the performance of such models is often influenced by their expert models. Moreover, models like AnomalyGPT not only are susceptible to overfitting due to the limited scale of IAD data but also have rough anomaly localization and lack rigorous reasoning processes to substantiate detection results, hindering their applicability in real-world scenarios. Consequently, proposing the first multimodal chain-of-thought dataset based on IAD tasks is of great significance.

### 2.2 Multimodal Large Language Model

MLLM integrates multiple modalities, including vision, into LLMs to form large models endowed with visual understanding capabilities. The cross-modal information input to MLLM is initially mapped to the text modality and subsequently processed by LLM. This modeling paradigm has demonstrated robust performance in a series of visual tasks. Early research such as BLIP2 [26] and Flamingo [27] adopted the frozen visual encoder paradigm, achieving end-to-end visual question answering capabilities through alignment with LLM. Subsequently, the LLaVA series [28] and MiniGPT-4 [29] introduced visual instruction fine-tuning methods, significantly enhancing the ability to follow intricate instructions. Models like Qwen-VL series [30], VisionLLM [31], and KOSMOS2 [32] integrate regional visual features to bolster the visual foundation capabilities of MLLM, enabling it to perform tasks such as regional semantic localization. InternVL [33] expands the visual encoder and optimizes cross-modal attention to align with LLM semantics. Additionally, Gemini exhibits cross-domain general visual understanding capabilities through large-scale multimodal pre-training. Models such as DeepSeek-VL2 [34] and Uni-MoE [35] employ a mixture-of-experts architecture to enhance multimodal understanding capabilities. However, existing MLLMs still struggle to reason as naturally as humans, which limits their effectiveness in real-world complex environments. Therefore, fine-tuning training on a large amount of chain-of-thought data from multiple scenarios on mainstream MLLMs can enable them to acquire good reasoning abilities to handle challenging visual tasks.

## 3 Dataset

### 3.1 Data Collection

Anomaly detection tasks in real-world settings are typically complex and variable, and various types of defects can occur in different products. Therefore, the datasets we construct are required to cover multiple scenarios and defect types of anomaly detection. We initially acquired and sampled from nine publicly available IAD datasets. To further expand the scenarios, we also manually gathered data from four scenarios applicable to IAD tasks to form several independent datasets and eliminated low-quality data through manual examination. MVTec AD is one of the most renowned IAD datasets, containing high-resolution images of various objects and corresponding to different types of defects. The VisA [36] dataset further enhances the complexity of scenarios and the data scale, including different instances of the same type of objects. GoodsAD [37] offers the possibility for intelligent

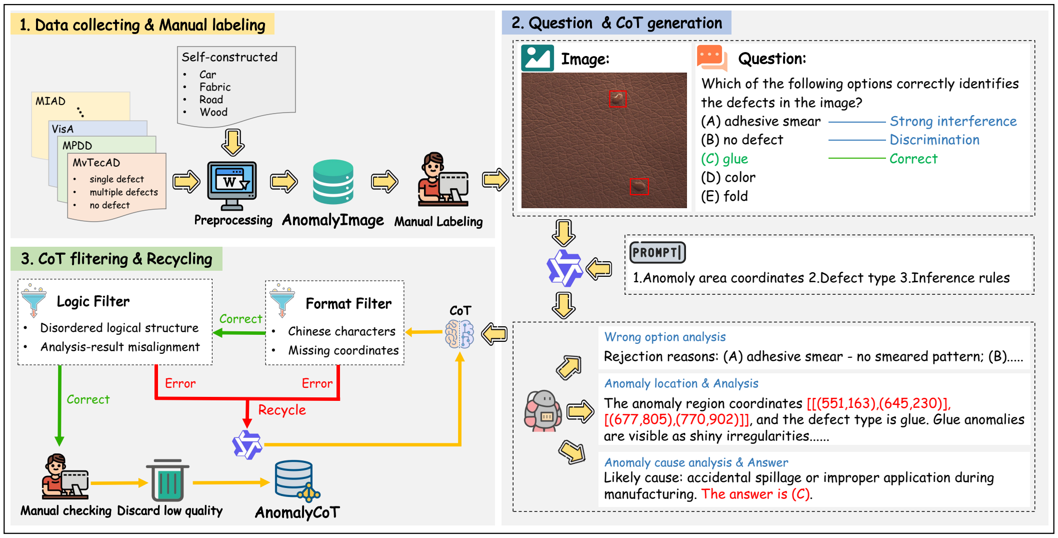



Figure 2: Data generation process of AnomalyCoT. The construction of AnomalyCoT mainly conserns three parts: the collection of image data and manual labeling, the generation of questions and CoT, the filtering and recycling of CoT data.

applications of anomaly detection by establishing the first IAD commodity dataset. MVTec LOCO AD [38] is dedicated to exploring logical anomalies and constructs a novel dataset covering both structural and logical anomalies. We also adopted Defect Spectrum [39], which effectively optimizes based on multiple key IAD datasets.

In industrial fields with less attention, MPDD [40] and NanoTwice [41] are respectively designed for anomaly detection tasks in the production processes of painted metal parts and nanofiber materials. MSD [42] is a potential dataset for surface defect detection of smartphone screens. MIAD is the largest component of the dataset in this paper, containing over 100K high-resolution outdoor IAD images. The large-scale, high-quality data make a significant contribution to the application of IAD tasks in practical environments. Additionally, we collected IAD data related to wood, fabric, car, and road cracks, screened them for high-quality samples, and then manually classified the unlabeled data to build four independent datasets. These independent datasets fill the voids in certain scenarios of the current task. We integrated the aforementioned processed datasets, precisely located all anomalous regions, and classified the defects within them through manual annotation, then verified the correctness of the labels through manual inspection, and finally manually corrected the incorrect labels. A total of 10 individuals participated in this labeling process, which accumulated to 230 hours.

### 3.2 Question Construction

In the actual production process, operators typically focus on whether there are anomalous situations in the products and the possible types of defects. When necessary, they also need to understand the scale and specific location information of the anomalies to analyze the causes and prevent subsequent influences. Thus, we designed a key task integrating anomaly discrimination and defect classification, generating a reasoning process that combines anomalous region information with cause analysis grounded in the framework of this task. We presented the task in the form of single-choice questions to evaluate the output of MLLM. Previous work [43, 44] has also demonstrated the rationality of this approach. To effectively implement this complex task and avoid the inherent biases in MLLM, we first defined the no defect option and manually designed strong distractors with high semantic similarity to the corresponding answer items of the samples. Subsequently, we manually constructed defect type libraries for different scenarios and randomly selected other defect types from them as supplementary options, ultimately forming a difficult task with five options.

### 3.3 Data Generation

The existing public IAD datasets lack the reasoning process from task to result, which makes it difficult for MLLMs to conduct rigorous evaluations of IAD tasks directly. Therefore, we have constructed a systematic pipeline to generate complete CoT for IAD samples. Our process utilizes

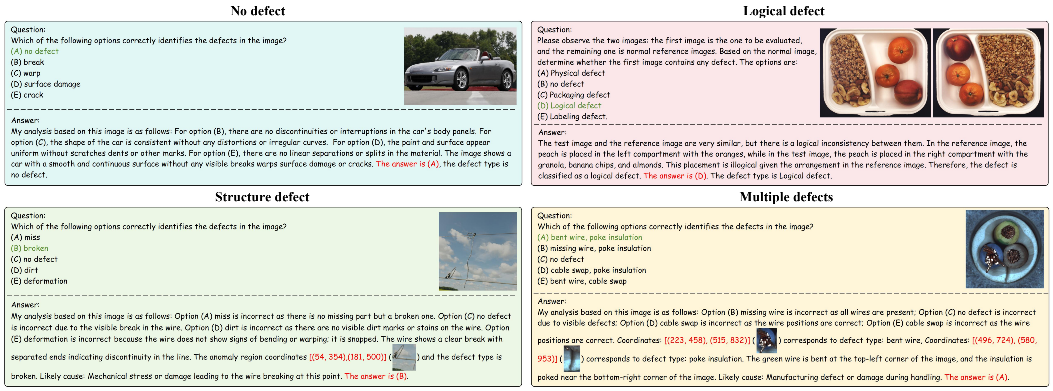



Figure 3: Samples of different defects. In AnomalyCoT, there are samples with different defects including normal image without any defect, single defect, multiple defects and logical defect.

the outstanding visual perception and text reasoning capabilities of QwenVL-Max, combined with manual annotation, semantic prompts, and rule-based filtering mechanisms, as depicted in Figure 2. Firstly, we design challenging questions for IAD samples. Since most MLLMs provide rough anomaly localization in text descriptions when performing IAD tasks and have difficulty forming reasoning processes that conform to the thinking patterns of professionals, we further design comprehensive task prompts. The task prompts for samples not only include precise visual prompts such as the coordinates of the annotated anomalous regions but also semantic prompts like defect type labels and standardized reasoning guidance. By learning the rich prompt content, QwenVL-Max generates a rigorous reasoning process consisting of important components such as the analysis of incorrect options, analysis based on anomaly localization, analysis of the causes of anomalies, and the answer. This reasoning process is then combined with the task to form complete CoT data, effectively simulating the cognitive patterns humans employ when addressing analogous problems.

To ensure the accuracy and rationality of the generated CoT, we have developed a two-step filtering approach. In the first stage, the CoT data undergoes Chinese character detection and missing coordinate detection in the first step initially, separating data with obvious format errors. Subsequently, the data undergoes logical error filtering in the second step. We integrate LLMs with superior natural language comprehension to filter data samples exhibiting both analysis-answer misalignment and disordered logical architectures (such as reversed cause and effect), with representative examples demonstrated in Appendix E. We also re-collected samples corresponding to the erroneous data to minimize data loss. Finally, all the data that have passed through the two-step filtering are manually inspected, and low-quality data are discarded to form our dataset AnomalyCoT. This inspection process involved 5 individuals and took a cumulative 10 hours. Figure 3 illustrates the four CoT data types in AnomalyCoT.

# 4 Experiments

## 4.1 Settings

To understand the effectiveness of AnomalyCoT, we conducted comprehensive evaluation on a series of MLLMs. For each question, the model is required to choose the correct option and identify the coordinates of the anomalous region, which reflect the model's ability to classify and locate anomalies. According to these, we report two aspects of metrics: accuracy and different types of IoU.

**Baselines.** Though many MLLMs have strong capability in vision-text comprehension, they may not generalize well in IAD tasks. Therefore, to understand the basic capabilities of MLLMs in this specific domain, we conduct evaluation on both commerical and open-source models. For commercial models, we tested GPT-4o and Gemini-2.0-flash. For open-source models, we tested and adapted AnomalyGPT, InternVL3, LLaVA-1.5 [45], Llama 3.2-Vision[3], Qwen2-VL [46], Qwen2.5-VL [13] and Gemma3 [47].

[3] https://github.com/worldart/meta-llama_llama-models/tree/main/models/llama3_2

Table 1: Overall evaluation results.

| Model | Scale | Type | Accuracy | IoU | GIoU | DIoU |
|---|---|---|---|---|---|---|
| GPT-4o | - | Pre-trained | 49.19% | 0.0364 | -0.2529 | -0.102 |
| Gemini-2.0-flash | - | Pre-trained | 59.64% | 0.0413 | -0.1923 | -0.057 |
| AnomalyGPT | 7B | Pre-trained | 19.17% | 0.0000 | -0.0414 | -0.0306 |
| InternVL 3 | 8B | Pre-trained | 57.33% | 0.0405 | -0.1159 | -0.0272 |
| | | LoRA | 91.27%(+33.94%) | 0.2667(+0.2262) | 0.0153(+0.1312) | 0.1747(+0.2019) |
| | | Full | 94.48%(+37.15%) | 0.4045(+0.3640) | **0.2154**(+0.3313) | 0.3532(+0.3804) |
| LLaVA-1.5 | 7B | Pre-trained | 28.84% | 0.0000 | 0.0000 | 0.0000 |
| | | LoRA | 59.12%(+30.28%) | 0.2123(+0.2123) | -0.0772(+-0.0772) | 0.0999(+0.0999) |
| | | Full | 96.94%(+68.10%) | 0.3155(+0.3155) | 0.1139(+0.1139) | 0.2540(+0.2540) |
| Llama 3.2-Vision | 11B | Pre-trained | 55.07% | 0.0001 | -0.0110 | -0.0054 |
| | | LoRA | 94.02%(+38.95%) | 0.2483(+0.2482) | 0.0194(+0.0304) | 0.1635(+0.1689) |
| | | Full | 97.07%(+42.00%) | 0.2784(+0.2783) | 0.0566(+0.0676) | 0.2059(+0.2113) |
| Qwen2-VL | 7B | Pre-trained | 51.61% | 0.0268 | -0.0409 | -0.0023 |
| | | LoRA | 86.92%(+35.31%) | 0.3201(+0.2933) | 0.1063(+0.1472) | 0.2496(+0.2519) |
| | | Full | 97.44%(+45.83%) | **0.4252**(+0.3984) | 0.2076(+0.2485) | **0.3642**(+0.3665) |
| Qwen2.5-VL | 7B | Pre-trained | 45.53% | 0.0939 | 0.0751 | 0.0864 |
| | | LoRA | 91.94%(+46.41%) | 0.3225(+0.2286) | 0.1069(+0.0318) | 0.2562(+0.1698) |
| | | Full | **97.46%**(+51.93%) | 0.3655(+0.2716) | 0.1462(+0.0711) | 0.3029(+0.2165) |
| Gemma 3 | 4B | Pre-trained | 22.15% | 0.0003 | -0.0238 | -0.0141 |
| | | LoRA | 64.33%(+42.18%) | 0.1000(+0.0997) | -0.2447(+-0.2209) | -0.0316(+-0.0175) |
| | | Full | 77.23%(+55.08%) | 0.3015(+0.3012) | 0.1329(+0.1567) | 0.2556(+0.2697) |
| Gemma 3 | 12B | Pre-trained | 39.12% | 0.0002 | -0.0149 | -0.0085 |
| | | LoRA | 67.48%(+28.36%) | 0.1232(+0.1230) | -0.1434(+-0.1285) | 0.0142(+0.0227) |
| | | Full | 79.42%(+40.30%) | 0.3257(+0.3255) | 0.1070(+0.1219) | 0.2652(+0.2737) |

**Fine-tuning.** We conduct both LoRA and full fine-tuning for selected models on AnomalyCoT. The tuning experiments are relied on LLaMA-Factory [48] and the samples are organized by standardized sharegpt format as shown in Appendix C. We train both models with batch size 32 by 3 epochs and more detailed configuration is recorded in Appendix D. In addition, the system prompt used for questioning is shown in Appendix B.

### 4.2 Experimental results

As shown in Table 1, we compare the performance over a series of MLLMs by metrics including accuracy, IoU [16], GIoU [49] and DIoU [50]. The results indicate a general performance bottleneck for MLLMs on IAD tasks, characterized by accuracy rates typically ranging from 20% to 60% and significantly low IoU scores. Specifically, the commercial MLLMs GPT-4o and Gemini-2.0-flash achieve accuracy rates of only 49.19% and 59.64% respectively, while the average accuracy of pre-trained open-source models is even lower, at only 45.55%. Critically, the IoU scores for these MLLMs are near zero. These deficiencies suggest that MLLMs, when relying solely on their general visual capabilities, exhibit limitations in accurately identifying the defect types and the precise coordinates of anomalous regions.

However, there are significant performance improvement when trained on AnomalyCoT. Following LoRA fine-tuning, Llama 3.2-Vision achieves the highest accuracy of 94.02%, while Qwen2.5-VL attains the highest IoU score of 0.3225. Compared to baseline models, the average accuracy rate and IoU score increase by 36.49% and 0.20 respectively. Furthermore, full fine-tuning results in even greater enhancements, with the highest accuracy reaching 97.46% and the highest IoU score reaching 0.4252. These results suggest that MLLMs can effectively learn the diverse scenarios represented in AnomalyCoT, leading to great performance of anomaly detection and location. Moreover, compared with LoRA, the further performance improvement observed through full fine-tuning indicates that the diversity of our dataset scenarios do not lead to overfitting issues. The comprehension ability of the visual model is the basis of multimodal tasks, and the addition of reasoning ability of the language model has made a great contribution to the outstanding performance in IAD tasks.

### 4.3 Ablation Study

The responses in AnomalyCoT consist of CoT reasoning and the coordinates of defects. To comprehensively understand the contributions to the great performance of MLLMs, we conduct ablation

Table 2: Ablation experiment results.

| Model | Scale | Accuracy w.o. coordinate | Accuracy w.o. cot | IoU w.o. cot | GIoU w.o. cot | DIoU w.o. cot |
|---|---|---|---|---|---|---|
| InternVL 3 | 8B | 90.42%(+33.09%) | 62.25%(+4.92%) | 0.2653(+0.2248) | 0.0329(+0.1488) | 0.1778(+0.2050) |
| LLaVA-1.5 | 7B | 59.35%(+30.51%) | 64.45%(+35.61%) | 0.2631(+0.2631) | 0.027(+0.0270) | 0.1792(+0.1792) |
| Llama 3.2-Vision | 11B | **93.24%**(+37.17%) | **86.27%**(+30.20%) | 0.2528(+0.2527) | 0.0833(+0.0943) | 0.1982(+0.2036) |
| Qwen2-VL | 7B | 86.66%(+35.05%) | 80.99%(+29.38%) | **0.3652**(+0.3384) | **0.1779**(+0.2188) | **0.3102**(+0.3125) |
| Qwen2.5-VL | 7B | 89.70%(+44.17%) | 82.71%(+37.18%) | 0.3164(+0.2225) | 0.1701(+0.0950) | 0.2929(+0.2065) |
| Gemma 3 | 4B | 64.42%(+42.27%) | 53.15%(+31.00%) | 0.1147(+0.1144) | -0.1228(-0.0990) | 0.0417(+0.0558) |
| Gemma 3 | 12B | 53.51%(+14.39%) | 62.12%(+22.00%) | 0.2088(+0.2086) | -0.0178(-0.0029) | 0.1274(+0.1359) |

**Notations:** *w.o.* stands for without.

experiments for these two factors. Based on AnomalyCoT, we construct two ablation datasets. One has no coordinate in analysis process and the other has no specific reasoning but only answers and coordinates. We conduct LoRA fine-tuning for open-sources MLLMs in these two datasets and the results are recorded in Table 2.

**Fine-tuning without CoT.** In this study, we investigate the effect of natural language reasoning in combination with visual features in anomaly detection. When removing specific reasoning process, MLLMs can only learn from the direct mapping from the original image and question to the correct option and coordinates. The performances of MLLMs are all greatly improved compared to baselines, but they are still much lower than the results of models fine-tuned on AnomalyCoT. This suggusts that through the correct option and the coordinates of anomaly regions, MLLMs can learn effective knowledge of anomaly detection tasks while the CoT reasoning can provide further comprehension of the question and image.

**Fine-tuning without coordinate.** In the ablation of coordinates, we reconstruct the related sentences or provide general orientation to ensure there is only a difference in location information compared with AnomalyCoT. This dataset can still guide MLLMs to reason based on the question and the image. The performance of each model have declined on the whole compared to AnomalyCoT, but the results vary according to the different models. For MLLMs including LLaVA-1.5-7B and Gemma3-4B, the performances of the ablation experiment are a little higher than AnomalyCoT. This indicates that different models have different sensitivity to textual interpretation and numerical coordinates and learning and fitting capabilities.

We innovatively introduced CoT reasoning in the multimodal anomaly detection task and provided the accurate coordinates of the anomaly region in CoT, which is of great benefit to the performance of the anomaly detection task.

### 4.4 Multi-scenario Analysis

The unbalanced number of samples has always been an important problem in the training of deep learning models [51], and it may lead to overfitting and low confidence. Since AnomalyCoT consists of images with different sample numbers from different industrial scenarios, we verify whether it has this problem. Table 3 shows the performance of Llama 3.2-Vision in 5 main scenarios, and the results of all scenarios are recorded in Appendix G.

Firstly, the trend in fine-tuning efficacy across individual scenarios aligns consistently with the overall results, with full fine-tuning outperforming LoRA fine-tuning and in turn outperforming the pre-trained model performance. This consistency demonstrates that the results of our experiment are universal in each scenario, and resolves concerns about potential biases caused by the imbalance of samples in the sub-datasets. Notably, the accuracy rates of the pre-trained model on MIAD and MVTec-LOCO are 49.03% and 73.08% respectively, with a gap of 24.05%. However, after fine-tuning, the accuracy rates on both sets reach over 90%, with a gap of less than 4%. Besides, Figure 4 visualizes the responses of Llama 3.2-Vision in two scenarios, showing that the model can extract different features for different objects and conduct different analyses.

Furthermore, from the result data of the ablation experiments in various scenarios, we found that after removing the thought chain, the accuracy of some scenarios decreased significantly (such as VisA),

Table 3: Results of Llama 3.2-Vision in 5 main scenarios.

| Scenario | Type | Accuracy | IoU | GIoU | DIoU |
|---|---|---|---|---|---|
| GoodsAD 2962 | Pre-trained | 54.21% | 0.0000 | 0.0000 | 0.0000 |
| | LoRA | 90.07%(+35.86%) | 0.0184(+0.0184) | 0.0174(+0.0174) | 0.0178(+0.0178) |
| | Full | 97.31%(+43.10%) | 0.0339(+0.0339) | 0.0318(+0.0318) | 0.0321(+0.0321) |
| MIAD 16725 | Pre-trained | 49.03% | 0.0000 | -0.8973 | -0.4190 |
| | LoRA | 96.23%(+47.20%) | 0.3473(+0.3473) | 0.0030(+0.9003) | 0.2190(+0.6380) |
| | Full | **97.73%**(+48.70%) | **0.3907**(+0.3907) | 0.0682(+0.9655) | **0.2834**(+0.7024) |
| MVTec-LOCO 1427 | Pre-trained | 73.08% | 0.0000 | 0.0000 | 0.0000 |
| | LoRA | 92.66%(+19.58%) | 0.0393(+0.0393) | -0.0031(-0.0031) | 0.0046(+0.0046) |
| | Full | 94.06%(+20.98%) | 0.0497(+0.0497) | 0.0171(+0.0171) | 0.0235(+0.0235) |
| MVTecAD 2906 | Pre-trained | 68.95% | 0.0000 | -0.2702 | -0.1222 |
| | LoRA | 87.65%(+18.70%) | 0.2068(+0.2068) | 0.1455(+0.4157) | 0.1717(+0.2939) |
| | Full | 97.26%(+28.31%) | 0.2223(+0.2223) | **0.1661**(+0.4363) | 0.1931(+0.3153) |
| VisA 2381 | Pre-trained | 64.50% | 0.0000 | 0.0000 | 0.0000 |
| | LoRA | 89.50%(+25.00%) | 0.0110(+0.0110) | -0.0353(-0.0353) | -0.0230(-0.0230) |
| | Full | 94.33%(+29.83%) | 0.0216(+0.0216) | -0.0175(-0.0175) | -0.0033(-0.0033) |

and after removing the regional coordinates, the accuracy of some scenarios decreased significantly (such as Nanotwice), and most scenarios belonged to the former. This indicates that in multiple scenarios of actual industrial anomaly detection, it is very necessary to provide both language and text explanations as well as the coordinates of the anomaly regions simultaneously in the training corpus.

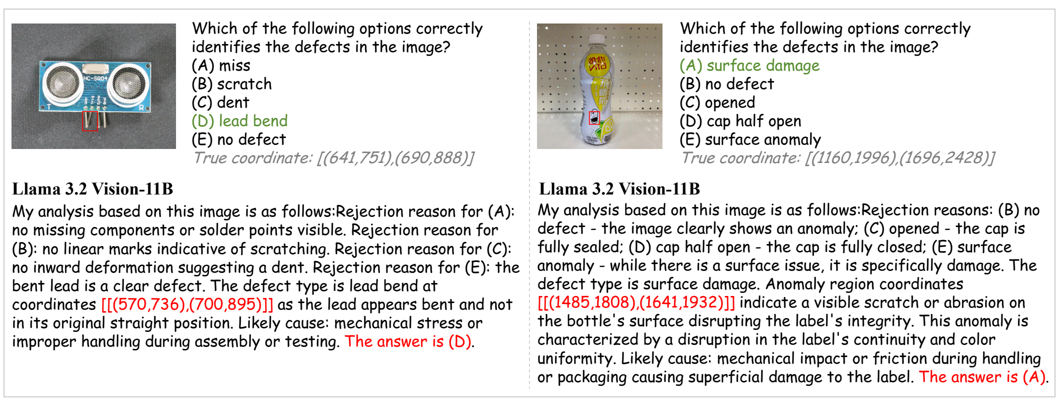



Figure 4: Comparison of responses in different scenarios. MLLMs can make corresponding analyses for different image scenarios, and we present the responses from Llama 3.2-Vision (LoRA fine-tuned) in sub-datasets VisA and GoodsAD.

## 5 Conclusion

In this paper, we briefly analyze the various limitations of previous anomaly detection datasets, particularly the lack of rigorous reasoning processes with precise visual information. To address these issues, we introduce AnomalyCoT, the first multimodal CoT dataset for the IAD task, which contains diverse scenarios and challenging tasks, providing strong support for the application of MLLMs in real-world environments to perform IAD tasks. In addition, we conduct different types of fine-tuning training on representative MLLMs, leveraging the CoT data to effectively learn professional thinking patterns across different scenarios. The experimental results demonstrate the reliability of the dataset, and all trained models achieve significant improvements in key metrics. Our contributions significantly advance the application of MLLMs in real-scene anomaly detection, establishing a solid foundation for future related studies.

## References


[1] Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. A comprehensive library for benchmarking multi-class visual anomaly detection. *arXiv preprint arXiv:2406.03262*, 2024.

[2] Xi Jiang, Guoyang Xie, Jinbao Wang, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. A survey of visual sensory anomaly detection. *arXiv preprint arXiv:2202.07006*, 2022.

[3] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[5] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[7] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc., 2023.

[8] Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.

[9] Jiangning Zhang, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, and Yong Liu. Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection. *arXiv preprint arXiv:2311.02612*, 1, 2023.

[10] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 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, 2019.

[11] Tianpeng Bao, Jiadong Chen, Wei Li, Xiang Wang, Jingjing Fei, Liwei Wu, Rui Zhao, and Ye Zheng. Miad: A maintenance inspection dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1002, 2023.

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

[13] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

[14] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[16] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, MM '16, page 516–520. ACM, October 2016.

[17] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. *CoRR*, abs/2201.10703, 2022.

[18] Xi Jiang, Ying Chen, Qiang Nie, Jianlin Liu, Yong Liu, Chengjie Wang, and Feng Zheng. Toward multi-class anomaly detection: Exploring class-aware unified model against inter-class interference. *CoRR*, abs/2403.14213, 2024.

[19] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8310–8319. IEEE, 2021.

[20] Jihun Yi and Sungroh Yoon. Patch SVDD: patch-level SVDD for anomaly detection and segmentation. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part VI*, volume 12627 of *Lecture Notes in Computer Science*, pages 375–390. Springer, 2020.

[21] Denis A. Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1819–1828. IEEE, 2022.

[22] Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14143–14152. IEEE, 2023.

[23] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 17826–17836. IEEE, 2024.

[24] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[25] Yuqi Jiang, Xudong Lu, Qian Jin, Qi Sun, Hanming Wu, and Cheng Zhuo. Fabgpt: An efficient large multimodal model for complex wafer defect knowledge queries. In Jinjun Xiong and Robert Wille, editors, *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, ICCAD 2024, Newark Liberty International Airport Marriott, NJ, USA, October 27-31, 2024*, pages 35:1–35:8. ACM, 2024.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[27] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[28] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[29] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[30] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2024.

[31] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[32] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824, 2023.

[33] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238, 2023.

[34] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *CoRR*, abs/2412.10302, 2024.

[35] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439, 2025.

[36] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.

[37] Jian Zhang, Runwei Ding, Miaoju Ban, and Linhui Dai. Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation. *IEEE Robotics and Automation Letters*, 9(3):2008–2015, 2024.

[38] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022.

[39] Shuai Yang, Zhifei Chen, Pengguang Chen, Xi Fang, Yixun Liang, Shu Liu, and Yingcong Chen. Defect spectrum: a granular look of large-scale defect datasets with rich semantics. In *European Conference on Computer Vision*, pages 187–203. Springer, 2024.

[40] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021.

[41] Diego Carrera, Fabio Manganini, Giacomo Boracchi, and Ettore Lanzarone. Defect detection in sem images of nanofibrous materials. *IEEE Transactions on Industrial Informatics*, 13(2):551–561, 2016.

[42] Jian Zhang, Runwei Ding, Miaoju Ban, and Tianyu Guo. Fdsnet: An accurate real-time surface defect segmentation network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3803–3807. IEEE, 2022.

[43] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13299–13308. IEEE, 2024.

[44] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer, 2024.

[45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[47] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[48] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

[49] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019.

[50] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression, 2019.

[51] Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901, 2024.

# A AnomalyImage

Table 4 presents the information of AnomalyCoT in more detail, including the objects contained in each sub-dataset and the corresponding defect types. It also records the specific number of samples corresponding to different objects and different numbers of defects.

Table 4: Construction of AnomalyImage.

| Dataset | Object | Number | Sample | | | Defect types |
|---|---|---|---|---|---|---|
| | | | single | multiple | no | |
| MVTecAD 2906 | bottle | 174 | 63 | - | 111 | broken, contamination |
| | cable | 231 | 80 | 10 | 141 | bent wire, cable swap, cut, missing, poke insulation |
| | capsule | 229 | 109 | - | 120 | crack, faulty imprint, poke, scratch, squeeze |
| | carpet | 231 | 80 | - | 151 | color, cut, metal contamination, thread |
| | hazelnut | 212 | 58 | - | 154 | crack, cut, hole, print |
| | leather | 226 | 85 | - | 141 | color, cut, fold, glue, poke |
| | metal nut | 211 | 90 | - | 121 | bent, color, flip, scratch |
| | pill | 284 | 123 | 16 | 145 | color, contamination, crack, faulty imprint, type, scratch |
| | screw | 296 | 115 | - | 181 | manipulated front, scratch, thread |
| | tile | 213 | 83 | - | 130 | crack, glue strip, gray stroke, oil, rough |
| | transistor | 179 | 38 | - | 141 | bent lead, cut lead, damaged case, misplaced |
| | wood | 185 | 46 | 9 | 130 | color, hole, liquid, scratch |
| | zipper | 235 | 102 | 13 | 120 | broken, fabric, rough, split, squeezed |
| MVTec-LOCO 1427 | breakfast box | 260 | 158 | - | 102 | logic, structure |
| | juice bottle | 283 | 189 | - | 94 | logic, structure |
| | pushpins | 308 | 170 | - | 138 | logic, structure |
| | screw bag | 340 | 218 | - | 122 | logic, structure |
| | splicing connectors | 236 | 117 | - | 119 | logic, structure |
| MSD | phone | 1220 | 1200 | - | 20 | oil, scratch, stain |
| MPDD 434 | bracket black | 73 | 47 | - | 26 | hole, scratch |
| | bracket brown | 82 | 50 | - | 32 | parts mismatch |
| | bracket white | 60 | 30 | - | 30 | defective printing, scratch |
| | connector | 22 | 13 | - | 9 | parts mismatch |
| | metal plate | 96 | 70 | - | 26 | rust, scratch |
| | tubes | 101 | 69 | - | 32 | color, crush-bend cutting, cut, deformation, flat crush, |
| MIAD 16725 | catenary dropper | 3054 | 2426 | - | 628 | broken, looseness, miss |
| | electrical insulator | 2868 | 2025 | - | 843 | broken |
| | nut and bolt d | 2272 | 1592 | - | 680 | missnut, looseness |
| | photovoltaic module | 3473 | 2051 | - | 1422 | broken, foreign body, miss |
| | wind turbine | 1994 | 1501 | - | 493 | crack |
| | witness mark | 3064 | 1993 | - | 1071 | looseness |
| GoodsAD 2962 | cigarette box | 426 | 243 | - | 183 | opened |
| | drink bottle | 772 | 417 | - | 355 | cup half open, cup open, surface damage |
| | drink can | 292 | 146 | - | 146 | deformation, strew missing, surface damage |
| | food bottle | 599 | 356 | - | 243 | deformation, opened, surface damage |
| | food box | 391 | 247 | - | 144 | deformation, opened, surface damage |
| | food package | 482 | 229 | - | 253 | broken, surface anomaly |
| Defect Spectrum 1197 | DS-Cotton-Fabric | 102 | 86 | - | 16 | bubble, texture |
| | DS-DAGM | 330 | 266 | - | 64 | color, crush, dirty, scratch, texture |
| | DS-VISION | 765 | 589 | 20 | 156 | scratch, crack, crush, bump, dirty, gap, friction, texture, color, fiber, point |
| VisA 2381 | candle | 197 | 94 | 3 | 100 | bump, dent, deposit, scratch, uneven edge, wick |
| | capsules | 199 | 99 | - | 100 | blister, flattening, inclusion, leakage, scratch |
| | cashew | 187 | 92 | - | 95 | adhesion, coating residue, discoloration, edge chip, hole, inclusion, pinhole, scratch, surface damage |
| | chewinggum | 199 | 99 | - | 100 | contamination, corner chip, scratch, surface damage |
| | fryum | 200 | 80 | 20 | 100 | break, deposit, scratch |
| | macaroni1 | 200 | 100 | - | 100 | chip, crack, discoloration, hole, scratch, stain |
| | macaroni2 | 200 | 100 | - | 100 | chip, crack, discoloration, hole, scratch, stain |
| | pcb1 | 200 | 100 | - | 100 | lead bend, miss, scratch, solder residue |
| | pcb2 | 205 | 100 | - | 105 | lead bend, miss, scratch, solder residue |
| | pcb3 | 198 | 98 | - | 100 | lead bend, miss, scratch, solder residue |
| | pcb4 | 197 | 97 | - | 100 | component damage, foreign material, miss, oxidation, scratch |
| | pipe_fryum | 199 | 99 | - | 100 | break, deposit, overlap, scratchm, stain |
| NanoTwice | NanoTwice | 49 | 35 | - | 14 | dust |
| Self 8466 | Car | 2979 | 2008 | 8 | 963 | break, crack, dent, surface damage |
| | Road | 2247 | 2162 | - | 85 | crack |
| | Wood | 194 | 194 | - | - | crack, scratch |
| | Fabric | 3046 | 2272 | 280 | 494 | color deviation, crease, hole, seam mark, stain, watermark, weaving defect |

## B Prompt

**Prompt for CoT generation**

# role definition
You are an AI model for anomaly detection. You should provide a reasoning process for the following question as instructed.
# correct answer
The correct answer is <option> and the anomaly region coordinates <boxes>.
# reasoning
Please analyze the entire image to support the correct choice. For each incorrect option, provide a rejection reason citing specific missing features. If the answer is \"no defect\", please provide a brief analysis; If not,immediately specify the anomaly region coordinates and analyse this region in detail. Then state \"The defect type is <type>\", appending \"Likely cause:\" with a explanation. Finally, conclude your response with \"The answer is <option>.\". The coordinates are given in pixels, the anomaly region is defined by its top-left and bottom-right corners.

Figure 5: Prompt for CoT generation. In the generation of CoT, we provide Qwen-VL-MAX with the correct choice and the anomaly region coordinates if existing. For different questions, these information will be replaced with the corresponding ones, that is, `<\option>` and `<\boxes>` in the figure.

**Prompt for training and evaluation**

# role definition
You are an AI model for anomaly detection. You should answer the following question as instructed.
# reasoning
Please analyze the entire image to support your judgment. For each incorrect option, provide a rejection reason citing specific missing features. If your answer is \"no defect\", please provide a brief analysis; If not, immediately specify the anomaly region coordinates in the format of [[(x1, y1),(x2, y2)], ... ] and analyse this region in detail to identify the defect type. Finally, conclude your response with \"The answer is <option>.\".

Figure 6: Prompt for training and evaluation. This prompt is used in the fine-tuning on AnomalyCoT and the evaluation of MLLMs, and it guides MLLMs to reason out the defect types and coordinates.

**Prompt for ablation of coordinate**

# role definition
You are an AI model for anomaly detection. You should answer the following question as instructed.
# reasoning
Please analyze the entire image to support your judgment. For each incorrect option, provide a rejection reason citing specific missing features. If the answer is \"no defect\", please provide a brief analysis; If not, analyse the anomaly region in detail to identify the defect type. Finally, conclude your response with \"The answer is <option>.\".

Figure 7: Prompt for ablation of coordinate. This prompt is used to fine-tune MLLMs in the ablation study for coordinate.

**Prompt for ablation of CoT**

# role definition
You are an AI model for anomaly detection. You should answer the following question and identify the anomaly region coordinates if existing.

Figure 8: Prompt for ablation of CoT. This prompt is used to fine-tune MLLMs in the ablation study for reasoning process.

# C Data Structure

Each sample in AnomalyCoT is organized in the structure shown Figure 9. It contains two main part, messages and images. In messages there are system prompt, user query and model response. In images there are input images.


```
{
  "messages": [
    {
      "content": "You are an AI model for anomaly detection. You should......",
      "role": "system"
    },
    {
      "content": "<image>Which of the following options correctly identifies the defects in the image? ......",
      "role": "user"
    },
    {
      "content": "My analysis based on this image is as follows: ......",
      "role": "assistant"
    }
  ],
  "images": [
    "AnomalyCOT/MIAD/wind_turbine/crack/0000.jpg"
  ]
}
```


Figure 9: Data structure of AnomalyCoT.

# D Fine-Tuning Configuration

We conducted LoRA and full-parameter fine-tuning on diverse MLLMs using 8 NVIDIA A100 GPUs, and the hyperparameter settings used during the fine-tuning process are detailed in Table 5.

Table 5: Hyperparameter Settings for Fine-tuning.

| **Hyperparameter** | **Value / Strategy** |
|---|---|
| Batch size | 32 |
| Cutoff length | 8192 |
| Optimizer | AdamW |
| Initial learning rate | 1e-4 |
| Learning rate scheduler | Cosine decay |
| Warmup ratio | 0.1 |
| Precision | BF16 |
| Number of epochs | 3 |

## E Sample of logical error

The filtering of CoT consists of two steps. The first is the filtering of the format and the second is the filtering of the logic. Logic error includes various forms such as causal transposition and semantic contradictions, Fig 10 is a specific sample of causal transposition.

Which of the following options correctly identifies the defects in the image?
(A) bulge
(B) no defect
(C) break
(D) surface damage
(E) dent

**Causal transposition**

The answer is (B). The anomaly region coordinates [[(131,244),(219,307)],[(235,237),(376,285)]] The defect type is dent. The front bumper shows an inward deformation consistent with a dent impact damage not linear or sharp to be a crack no elevation for bulge continuous surface no separation for break. Likely cause: low-speed collision or object impact.

**Correct**

My analysis based on this image is as follows: Not linear or sharp to be a crack no elevation for bulge continuous surface no separation for break. The front bumper shows an inward deformation consistent with a dent impact damage. The anomaly region coordinates [[(131,244),(219,307)],[(235,237),(376,285)]]. Likely cause: low-speed collision or object impact. The defect type is dent. The answer is (B).

Figure 10: Sample of causal transposition. The reversal and effect is a typical error of the generated CoT data. To guarantee the quality of AnomalyCoT, we filter and recycle these erroneous data. This figure shows the original CoT with logic error and the correct CoT after recycling.

## F Case Study

We demonstrate the responses of different models in this section. Figure 11,12,13 report the responses of different fine-tuned models based on Qwen2.5-VL-7B of questions of different difficulties. Figure 14 reports the responses of different MLLMs after LoRA fine-tuning on AnomalyCoT.

Which of the following options correctly identifies the defects in the image?
(A) color
(B) fold
(C) adhesive smear
(D) glue
(E)no defect

*True coordinate: [(571,512),(708,621)]*

**Full**

My analysis based on this image is as follows:Color option (A) is incorrect as the color variation is due to the reflective property of the glue and not a color defect; fold (B) is incorrect as there are no creases or disruptions in the texture pattern; adhesive smear (C) is incorrect because smears typically show spread marks which are absent here; no defect (E) is incorrect as the presence of the glue spot clearly indicates a defect. The anomaly region coordinates [[(659,574),(838,697)]] and the defect type is glue. The defect appears as a glossy, raised spot that disrupts the uniform texture of the material indicating an adhesive substance. Likely cause: accidental spill or application of adhesive during manufacturing or handling. The answer is (D).

**LoRA**

My analysis based on this image is as follows:Rejection reasons: (A) color - no significant color deviation; (B) fold - no creases or folds present; (C) adhesive smear - while similar, the shape and texture suggest a more defined substance; (E) no defect - the anomaly clearly deviates from the uniform texture. The anomaly region coordinates [[(618,535),(787,663)]] and the defect type is glue. The defect appears as an irregularly shaped, glossy spot inconsistent with the surrounding texture indicating adhesive residue. Likely cause: accidental spill or application of adhesive during manufacturing or handling. The answer is (D).

**Without detection boxes**

My analysis based on this image is as follows:Rejection reason for (A): No significant color variation observed. Rejection reason for (B): No evidence of folding or creasing. Rejection reason for (C): Absence of adhesive smear characteristics such as smeared texture or color change. Rejection reason for (E): Presence of glue indicates a defect. The anomaly region coordinates are identified, and the defect type is glue. The defect appears as a shiny, irregularly shaped substance inconsistent with the texture of the surrounding material indicating adhesive properties. Likely cause: Excess adhesive applied during manufacturing or handling. The answer is (D).

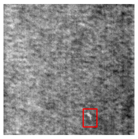
**Without CoT**

The answer is (D) glue. The anomaly region coordinates are [[(908, 785), (1169, 963)]].

**Pre-train**

The anomaly region coordinates are [[(0.45, 0.37), (0.55, 0.43)]]. This image shows a small, circular or oval-shaped mark on the surface of a brown leather material. The mark is shiny and reflective, indicating that it may be made of a smooth, glossy material. The shape and size of the mark suggest that it could be a smear of glue, which is often used to bond leather together. The location of the mark in the center of the image, surrounded by the textured pattern of the leather, further supports this conclusion. Overall, the presence of this shiny, circular mark on the leather surface suggests that the correct answer is (D) glue.

Figure 11: Comparison between different fine-tuned models based on Qwen2.5-VL-7B. The photo shows a leather surface with glue. All the fine-tuned models can select the correct defect type and identify the coordinates of the defect.



Which of the following options correctly identifies the defects in the image?
(A) no defect
(B) bump
(C) crush
(D) shade
(E) color True coordinate: [(310,417),(351,458)]

**Full**
My analysis based on this image is as follows:For option (A), no defect is incorrect as there is a visible anomaly. Option (B), bump, is rejected because there is no evidence of a raised surface; the texture remains consistent except for the color variation. Option (C), crush, is incorrect as there are no signs of compression or distortion in the texture. Option (D), shade, is rejected because the anomaly is not a shadow but a change in color. The defect type is color. The anomaly region coordinates [[(317,415),(349,444)]] show a noticeable variation in color compared to the surrounding area indicating a color defect. Likely cause: The color variation could be due to a material impurity or a manufacturing defect affecting the dyeing process. The answer is (E).

**LoRA**
My analysis based on this image is as follows:Option (A) no defect rejected because there is visible irregularity. Option (B) bump rejected because there is no raised area. Option (C) crush rejected because there is no indentation or compression. Option (D) shade rejected because lighting variations are consistent across the image. The anomaly region coordinates [[(329,416),(378,459)]] show a lighter spot compared to the surrounding area indicating an inconsistency in color. Likely cause: uneven dye application or material inconsistency. The answer is (E). The defect type is color.

**Without detection boxes**
My analysis based on this image is as follows:Option (A) rejected as there are visible defects. Option (B) rejected as there's no raised surface indicative of a bump. Option (C) rejected as there's no compression or deformation suggesting a crush. Option (E) rejected as the color variation is consistent with the texture and doesn't indicate a defect. The anomaly region coordinates are identified in the image. The defect type is shade. The specified area shows a darker patch compared to the surrounding uniform texture indicating a shading issue. Likely cause: uneven lighting or shadowing during capture. The answer is (D).

**Without CoT**
The answer is (C)crush. The anomaly region coordinates are [[(345, 416), (378, 452)]].

**Pre-train**
Analysis: The image appears to be uniformly gray without any visible defects such as bumps, crushes, shadows, or color variations. It seems to be a consistent background pattern with no anomalies detected. The answer is (A) no defect.

Figure 12: Comparison between different fine-tuned models based on Qwen2.5-VL-7B. The photo shows a textile surface with color inconsistence. Only the models fine-tuned on complete CoT data can select the correct defect type and identify the coordinates of the defect.

Which of the following options correctly identifies the defects in the image?
(A) break
(B) dent True coordinate: [[(347,178),(284,301)],[(284,257),(233,369)]]
(C) bulge
(D) no defect
(E) surface damage

**Full**
My analysis based on this image is as follows:Rejection reasons: (A) break - no visible cracks or separation of material; (C) bulge - the deformation is inward not outward; (D) no defect - the dent is clearly visible; (E) surface damage - while present, it's secondary to the dent and not the primary defect. Anomaly region coordinates [[(194,295),(268,372)],[(276,216),(345,303)]] indicate a noticeable inward deformation on the car's front left side consistent with a dent. Likely cause: impact from another object causing the metal to deform inward without breaking. The answer is (B). The defect type is dent.

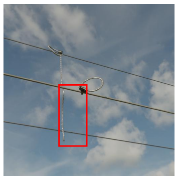
**LoRA**
My analysis based on this image is as follows:For option (A), there's no visible break as the surface continuity is intact without any cracks or separations. Option (C) is incorrect because a bulge would show an outward deformation which isn't observed here. Option (D) is rejected as the presence of a dent indicates a defect. Option (E) is not accurate since surface damage typically refers to scratches or paint issues rather than structural deformations. The anomaly region coordinates [[(345,165),(257,303)]] and the defect type is dent. The front left side of the car shows a noticeable inward deformation consistent with a dent which disrupts the smooth contour of the vehicle's bodywork. Likely cause: impact from another object causing the metal to deform inwardly. The answer is (B).

**Without detection boxes**
My analysis based on this image is as follows:Rejection reasons: (A) break - no visible cracks or separation; (C) bulge - no outward protrusion; (D) no defect - clear presence of deformation; (E) surface damage - no scratches or paint disruption. The anomaly region coordinates correspond to the identified area and the defect type is dent. The front left side of the car shows an inward deformation consistent with a dent likely caused by a low-speed impact or object strike. Likely cause: accidental collision or contact with another object. The answer is (B).

**Without CoT**
The answer is (B)dent. The anomaly region coordinates are [[(270, 235), (318, 348)]].

**Pre-train**
The answer is (D) no defect. There are no visible signs of break, dent, bulge, or surface damage in the image of the vehicle. The vehicle appears to be in good condition without any noticeable anomalies.

Figure 13: Comparison between different fine-tuned models based on Qwen2.5-VL-7B. The photo shows a car with dents. All the fine-tuned models can select the correct defect type, but only full-parameters fine-tuned model identify the coordinates of the defect in an expected degree of detail.



Which of the following options correctly identifies the defects in the image?
(A) paint defect
(B) no defect
(C) broken
(D) deformation
(E) miss

*True coordinate: [(128,119),(201,416)]*

**InternVL3-8B**

My analysis based on this image is as follows:Rejection reasons: (A) paint defect - no paint present on the wire; (B) no defect - the break is evident; (D) deformation - the wire is not bent or deformed but broken; (E) miss - the defect is clearly visible. The anomaly region coordinates [[(129,114),(198,419)]] and the defect type is broken. The wire shows a clear break with separated ends indicating discontinuity in the cable structure. Likely cause: mechanical stress or material fatigue leading to failure. The answer is (C).

**LLaVA-1.5-7B**

My analysis based on this image is as follows:Rejection reasons: (A) paint defect - no visible paint on wires; (B) no defect - clear visual evidence of breakage; (D) deformation - the wire is not bent or deformed but severed; (E) miss - the defect is present and identifiable. The anomaly region coordinates [[(121,105),(216,424)]] and the defect type is broken. The wire shows a clear break with separated ends indicating discontinuity in the line. Likely cause: mechanical stress or damage leading to snapping of the wire. The answer is (C).

**Llama 3.2 Vision-11B**

My analysis based on this image is as follows:Option (A) paint defect rejected as there are no signs of paint on the wire. Option (B) no defect rejected because the break is evident. Option (D) deformation rejected as the wire shows a clean break rather than a bent or twisted shape. Option (E) miss rejected since the defect is present and identifiable. The anomaly region coordinates [[(119,123),(210,453)]] and the defect type is broken. The wire shows a clear break with separated ends indicating discontinuity in the line. Likely cause: mechanical stress or damage leading to snapping of the wire. The answer is (C).

**Qwen2-VL-7B**

My analysis based on this image is as follows:Rejection reasons: (A) paint defect - no visible paint on the wire; (B) no defect - the break is evident; (D) deformation - the wire is not bent or deformed but broken; (E) miss - the defect is present and identifiable. The anomaly region coordinates [[(131,113),(191,416)]] and the defect type is broken. The wire shows a clear break with separated ends indicating discontinuity in the structure. Likely cause: mechanical stress or material fatigue leading to failure. The answer is (C).

**Qwen2.5-VL-7B**

My analysis based on this image is as follows:Rejection reasons: (A) paint defect - no paint visible on wires; (B) no defect - the break is evident; (D) deformation - the wire is not bent or twisted but severed; (E) miss - there is a clear defect present. The anomaly region coordinates [[(132,108),(196,419)]] and the defect type is broken. The wire shows a clear break with separated ends indicating discontinuity in the line. Likely cause: mechanical stress or material fatigue leading to failure at this point. The answer is (C).

**Gemma 3-12B**

The wire shows a clear break with separated ends indicating discontinuity in the line. Likely cause: mechanical stress or material fatigue leading to fracture. The answer is (C)broken. Anomaly region coordinates [(175,200),(38,465)].

**Gemma 3-12B**

(A) paint defect - no visible paint on the cable; (B) no defect - the break is evident; (D) deformation - the shape of the cable is not altered except at the break; (E) miss - the defect is present and identifiable. The anomaly region coordinates [[(113,124),(193,424)]] show a clear break in the cable with frayed edges indicating separation. Likely cause: mechanical stress or material fatigue leading to failure. The answer is (C).The defect type is broken.

Figure 14: Comparison between different MLLMs. The phote shows a broken wire. Almost all the fine-tuned MLLMs can select the correct defect type and identify the coordinates of the defect.

## G Muti-scenario Results

Table 6 shows the complete performance of Llama 3.2-Vision, containing all the fine-tuning methods, in all the scenarios (or sub-datasets). After fine-tuned on AnomalyCoT, the MLLM achieves great performance on every scenario regardless of the number of corresponding samples in training set, which strongly proving the generalization ability of AnomalyCoT.

Table 6: Results of Llama 3.2-Vision in all scenarios

| Scene | Type | Accuracy | IoU | GIoU | DIoU |
|---|---|---|---|---|---|
| | Pre-trained | 42.62% | 0.0000 | -0.8873 | -0.4237 |
| Defect Spectrum | LoRA | 85.65%(+43.03%) | 0.1166(+0.1166) | -0.0913(+0.7960) | 0.0383(+0.4620) |
| | Full | 92.83%(+50.21%) | 0.1851(+0.1851) | -0.0168(+0.8705) | 0.1270(+0.5507) |
| samples: 237 | w.o. CoT | 71.78%(+29.16%) | 0.1536(+0.1536) | -0.0544(+0.8329) | 0.0781(+0.5018) |
| | w.o. coordinate | 86.50%(+43.88%) | - | - | - |
| | Pre-trained | 54.21% | 0.0000 | 0.0000 | 0.0000 |
| GoodsAD | LoRA | 90.07%(+35.86%) | 0.0184(+0.0184) | 0.0174(+0.0174) | 0.0178(+0.0178) |
| | Full | 97.31%(+43.10%) | 0.0339(+0.0339) | 0.0318(+0.0318) | 0.0321(+0.0321) |
| samples: 594 | w.o. CoT | 85.32%(+31.11%) | 0.0000(+0.0000) | 0.0000(+0.0000) | 0.0000(+0.0000) |
| | w.o. coordinate | 90.40%(+36.19%) | - | - | - |
| | Pre-trained | 49.03% | 0.0000 | -0.8973 | -0.4190 |
| MIAD | LoRA | 96.23%(+47.20%) | 0.3473(+0.3473) | 0.0030(+0.9003) | 0.2190(+0.6380) |
| | Full | 97.73%(+48.70%) | 0.3907(+0.3907) | 0.0682(+0.9655) | 0.2834(+0.7024) |
| samples: 3345 | w.o. CoT | 89.57%(+40.54%) | **0.3997**(+0.3997) | 0.1083(+1.0056) | **0.3050**(+0.7240) |
| | w.o. coordinate | 95.87%(+46.84%) | - | - | - |
| | Pre-trained | 59.65% | 0.0000 | -0.4992 | -0.3656 |
| MPDD | LoRA | 78.39%(+18.74%) | 0.1137(+0.1137) | -0.0201(+0.4791) | 0.0395(+0.4051) |
| | Full | 87.90%(+28.25%) | 0.1340(+0.1340) | 0.0208(+0.5200) | 0.0642(+0.4298) |
| samples: 347 | w.o. CoT | 70.89%(+11.24%) | 0.1064(+0.1064) | -0.0425(+0.4567) | 0.0296(+0.3952) |
| | w.o. coordinate | 75.22%(+15.57%) | - | - | - |
| | Pre-trained | 41.80% | 0.0000 | 0.0000 | 0.0000 |
| MSD | LoRA | **100.00%**(+58.20%) | 0.0388(+0.0388) | -0.1101(-0.1101) | -0.0337(-0.0337) |
| | Full | 99.59%(+57.79%) | 0.0562(+0.0562) | -0.0707(-0.0707) | 0.0064(+0.0064) |
| samples: 244 | w.o. CoT | 99.18%(+57.38%) | 0.0817(+0.0817) | -0.0121(-0.0121) | 0.0518(+0.0518) |
| | w.o. coordinate | 98.77%(+56.97%) | - | - | - |
| | Pre-trained | 73.08% | 0.0000 | 0.0000 | 0.0000 |
| MVTec-LOCO | LoRA | 92.66%(+19.58%) | 0.0393(+0.0393) | -0.0031(-0.0031) | 0.0046(+0.0046) |
| | Full | 94.06%(+20.98%) | 0.0497(+0.0497) | 0.0171(+0.0171) | 0.0235(+0.0235) |
| samples: 286 | w.o. CoT | 86.71%(+13.63%) | 0.0656(+0.0656) | 0.0398(+0.0398) | 0.0465(+0.0465) |
| | w.o. coordinate | 92.31%(+19.23%) | - | - | - |
| | Pre-trained | 68.95% | 0.0000 | -0.2702 | -0.1222 |
| MVTecAD | LoRA | 87.65%(+18.70%) | 0.2068(+0.2068) | 0.1455(+0.4157) | 0.1717(+0.2939) |
| | Full | 97.26%(+28.31%) | 0.2223(+0.2223) | 0.1661(+0.4363) | 0.1931(+0.3153) |
| samples: 583 | w.o. CoT | 85.47%(+16.52%) | 0.3554(+0.3554) | **0.1683**(+0.4385) | 0.3040(+0.4262) |
| | w.o. coordinate | 89.54%(+20.59%) | - | - | - |
| | Pre-trained | 33.33% | 0.0097 | -0.4297 | -0.1082 |
| NanoTwice | LoRA | 94.44%(+61.11%) | 0.0764(+0.0667) | -0.1915(+0.2382) | 0.0404(+0.1486) |
| | Full | 97.22%(+63.89%) | 0.2090(+0.1993) | -0.1016(+0.3281) | 0.1576(+0.2658) |
| samples: 36 | w.o. CoT | 100.00%(+66.67%) | 0.0952(+0.0855) | -0.0959(+0.3338) | 0.0597(+0.1679) |
| | w.o. coordinate | 83.33%(+50.00%) | - | - | - |
| | Pre-trained | 64.50% | 0.0000 | 0.0000 | 0.0000 |
| VisA | LoRA | 89.50%(+25.00%) | 0.0110(+0.0110) | -0.0353(-0.0353) | -0.0230(-0.0230) |
| | Full | 94.33%(+29.83%) | 0.0216(+0.0216) | -0.0175(-0.0175) | -0.0033(-0.0033) |
| samples: 476 | w.o. CoT | 77.31%(+12.81%) | 0.0220(+0.0220) | 0.0086(+0.0086) | 0.0113(+0.0113) |
| | w.o. coordinate | 86.76%(+22.26%) | - | - | - |
| | Pre-trained | 59.79% | 0.0228 | -0.6799 | -0.3205 |
| Self | LoRA | 97.45%(+37.66%) | 0.2786(+0.2558) | 0.0658(+0.7457) | 0.2168(+0.5373) |
| | Full | 98.87%(+39.08%) | 0.2959(+0.2731) | 0.0642(+0.7441) | 0.2359(+0.5564) |
| samples: 1768 | w.o. CoT | 85.47%(+25.68%) | 0.3554(+0.3326) | 0.1683(+0.8482) | 0.3040(+0.6245) |
| | w.o. coordinate | 96.21%(+36.42%) | - | - | - |

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction of our paper precisely mirror its contributions and scope. Specifically, we have made three key contributions, and experiments we conducted on the dataset were elaborated in detail.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a "Limitation" section in our paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our dataset has proved its validity and generalization through relevant experiments.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our relevant datasets, experimental codes and checkpoint have all been open-sourced, and others can reproduce them easily.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our paper provide open access to the data and code, with sufficient documents and scripts to faithfully reproduce the main experimental results. The related link is `https://github.com/Zhaolutuan/AnomalyCoT`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The script in the code link we provided presents the relevant Settings in detail.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the limitations of the parameter quantity and computing resources of the large model, it is very difficult for us to conduct multiple experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The introduction of relevant computing resources is in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in all aspects of our paper strictly complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the introduction section of our paper, we pointed out the positive impact of our research on IAD and MLLM.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., Pre-Trained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The dataset and checkpoint we released are related to the IAD domain and do not involve such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In our research, relevant detailed information and documentation are provided along with the new assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in our paper include datasets, codes and related checkpoints, all of which have detailed documentation and scripts for use.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our paper elaborates in detail on the participation of the large model in the dataset, including the processing of the data and related benchmark experiments.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.