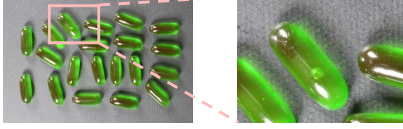


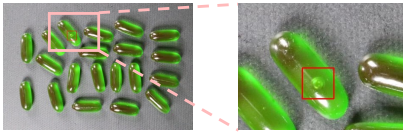
A1. Dataset Establishment

A1.1. How to highlight the anomaly?

Prompt Example: Bubble Inside the Capsule



User	Do you observe any anomalies in the image?
GPT-4o	The image shows green gel capsules that appear uniform in size, shape, and color. Upon inspection, no obvious anomalies such as broken capsules, irregular shapes, or unexpected colors are visible.
Let's draw a bounding box around the anomaly!	



User	Do you observe any anomalies in the image?
GPT-4o	Yes, there appears to be an anomaly in the image. One of the capsules, marked with a red box, has a noticeable defect. It seems to have a bubble or an air pocket inside, which is not observed in the other capsules. This could indicate a manufacturing defect or inconsistency in the capsule filling process.
If there is no bounding box, let's give a short hint to GPT-4o!	
User	Do you observe any anomalies in the image? (hint: a bubble in the capsule)
GPT-4o	The anomaly in the image is the presence of a visible air bubble inside one of the green gel capsules. While most capsules appear uniformly filled with liquid, the capsule located near the top center of the image has a small, clear air pocket within it. This could be a result of incomplete filling during the manufacturing process or an issue with the capsule sealing.

Table 7. Comparison of the GPT-4o [72] outputs with and without visual and textual hints for the anomaly.

As shown in Table 7, recent advanced MLLMs like GPT-4o fail to detect the anomalies in the image, so building the instruction tuning dataset using previous methods [8] is impractical. However, we observe that when the GPT-4o is provided some "hints", it presents impressive performance on anomaly reasoning or description. For example, a red bounding box drawn around the anomalous area enables GPT-4o to detect the tiny bubble inside the small capsule. This observation indicates that **the anomaly information is already contained in the visual tokens, and the failure of existing MLLMs is because the language model cannot effectively pick out the related tokens**, which is the major inspiration of our token-picking mechanism.

Most of the existing AD datasets, such as MVTec AD [2], contain anomaly masks for anomaly localization. Therefore,

we leverage these masks to generate the bounding boxes on the images. Specifically, the masks for an anomalous image are dilated and merged (if two masks are too close) before calculating the coordinates of the bounding boxes. Similarly, the image with bounding boxes drawn on it will serve as the visual prompt for GPT-4o. We also tried many other ways to utilize the anomaly masks, such as highlighting the mask area with different colors, consecutively providing the image and mask, and converting the normalized coordinates of the bounding box into a text prompt. None of them can as effectively guide the GPT-4o in finding anomalous features as drawing bounding boxes on the image.

A1.2. WebAD – The largest AD dataset

Existing industrial or medical anomaly detection datasets, such as MVTec AD [2] and BMAD [1], only contain a limited number of classes (< 20) and several different anomaly types for each class (most of the anomaly types are similar) due to the collection of these kinds of anomaly images involves extensive human involvements. This limitation hinders the ZSAD model from learning a generic description of anomaly and normal patterns. Also, the MLLMs cannot obtain enough knowledge of visual anomaly descriptions for unseen anomaly types. Therefore, more diverse data is required for a robust ZSAD & reasoning model. Many recent dataset works collect and annotate online images to enrich existing datasets and demonstrate their effectiveness in the training of current data-hungry deep learning models.

To collect the online images that can be utilized for anomaly detection, we design an automatic data collection pipeline by combining GPT-4o [72] and Google Image Search [26]. As shown in Figure 8, we first employ GPT-4o to list 400 class names commonly seen in our daily life. Then, for each class, the GPT-4o is asked to generate 10 corresponding anomalous and normal phrases based on the class name. The abnormality or normality descriptions indicated by these phrases are specifically suitable for the class name. These phrases will serve as the search prompts to query the image links in Google Image Search. However, the downloaded images are very "dirty" and contain many noise samples and duplications. For example, the collected anomaly set contains lots of normal images, and vice versa. A data-cleaning step is applied after the image collection.

Since the duplications mainly occur within a specific class, we extract the CLIP [73] features for all the images in the class and compare the cosine similarity of these features. If the similarity value is larger than 0.99, then one of the images will be removed. To deal with the problematic grouping of anomaly and normal images, we combine the image and its corresponding search prompt and give them to GPT-4o for normal and anomaly classification. In the system prompt, we explicitly tell the GPT-4o that the search prompt is just a hint and not always correct and ask GPT-4o

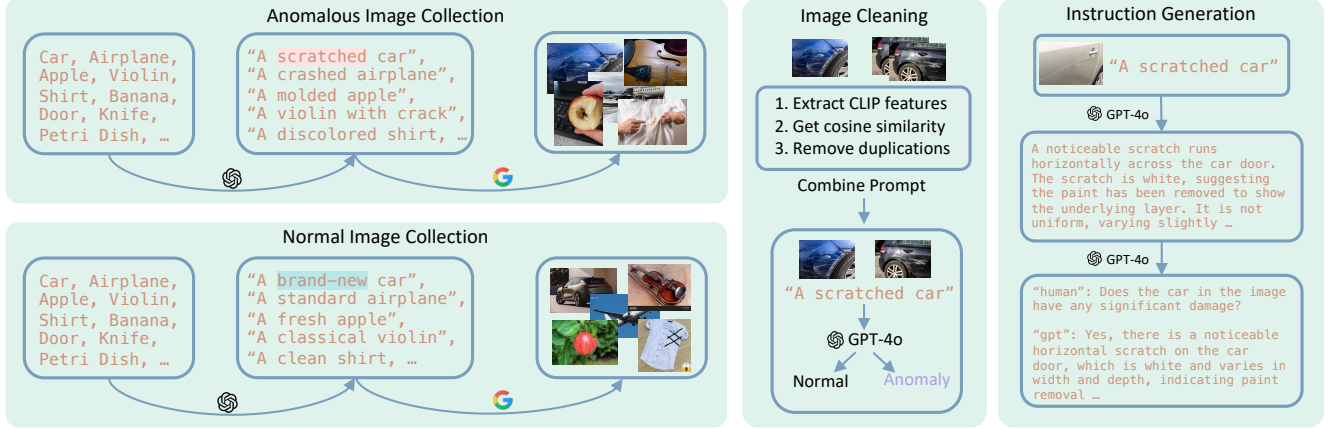


Figure 8. Automatic data collection pipeline for WebAD. The entire pipeline is fully automatic at an affordable cost (API usage). Other advanced open-sourced MLLMs can be applied to replace GPT-4o for further reduction of cost.

to determine the normality and abnormality by itself. This step will remove the images with incorrect labels and the artificial images, such as cartoons or art. Some samples in the collected WebAD dataset are shown in Figure 9. In total, WebAD contains around 72k images from 380 classes and more than 5 anomaly types for each class.

A1.3. Instruction Data Generation

For existing datasets, we manually combine the anomaly type and the class name to create the short anomaly prompt (hint). Then, the image with or without the bounding boxes and the corresponding short prompt are utilized to prompt GPT-4o for the generation of detailed descriptions of the image and the anomalies. These descriptions contain all the information required for instruction-following data. The in-context learning strategy is implemented to generate the multi-round conversation data (see Figure 10). Questions designed to elicit a one-word answer are utilized to balance the distribution of the normal and anomaly samples.

A2. Training Details

In the professional training stage, we leverage AdamW [65] to be the optimizer and CosineAnnealingWarmRestarts [64] as the learning rate scheduler. The initial learning rate is set to be $1e-4$, and the restart iteration is half of the single epoch. The anomaly expert is trained on 8 H100 GPUs for 2 epochs (2 hours), and the total batch size is 128. In the instruction tuning stage, we follow the default training setting of *LLaVA-OneVision* [44] (reduce the batch size to 128), and the total training time for 0.5B and 7B models are 7 hours and 50 hours on 8 H100, respectively. When sampling the instruction data from the original recipe of *LLaVA-OneVision*, we put more emphasis on low-level image understanding and 3D multi-view Q&A, considering that anomaly detection originates from the low-level feature differences and the

3D anomaly detection requires multi-image understanding. Besides, for more knowledge in the medical domain, the model is also fed with the data from *LLaVA-Med* [45].

A3. Experimental Results

A3.1. Anomaly Detection

Similar to previous ZSAD works, the detailed image-level AUROC results for the anomaly expert of Anomaly-OV on VisA [115] and MVTec AD [2] are provided in Table 8.

A3.2. Anomaly Reasoning

Table 9 to 13 presents more comparison results of GPT-4o [72], *LLaVA-OneVision* [44], and Anomaly-OV on AD & reasoning. Anomaly-OV shows better performance in the detection and description of the visual anomalies in the images. Table 14 demonstrates the low-level and complex reasoning capability of Anomaly-OV for an in-the-wild image, indicating a comprehensive understanding of the anomaly.

A4. Limitation and Future Work

Limitation. As shown in Table 15, sometimes, Anomaly-OV fails to provide an accurate classification of the target object, describes the anomaly by a general word (wax missing is described by "crack"), or presents wrong reasoning with hallucination. Also, there is still a large space for improvement in the detection performance of Anomaly-OV. Besides, the images contained in VisA-D&R are from the industrial domain, so more benchmarks in other domains, such as 3D and medical anomaly detection, are required to evaluate a unified AD & reasoning model.

Future Work. The detection performance of Anomaly-OV is highly determined by the anomaly expert (see Table 4), so a more advanced design of the expert model is recommended

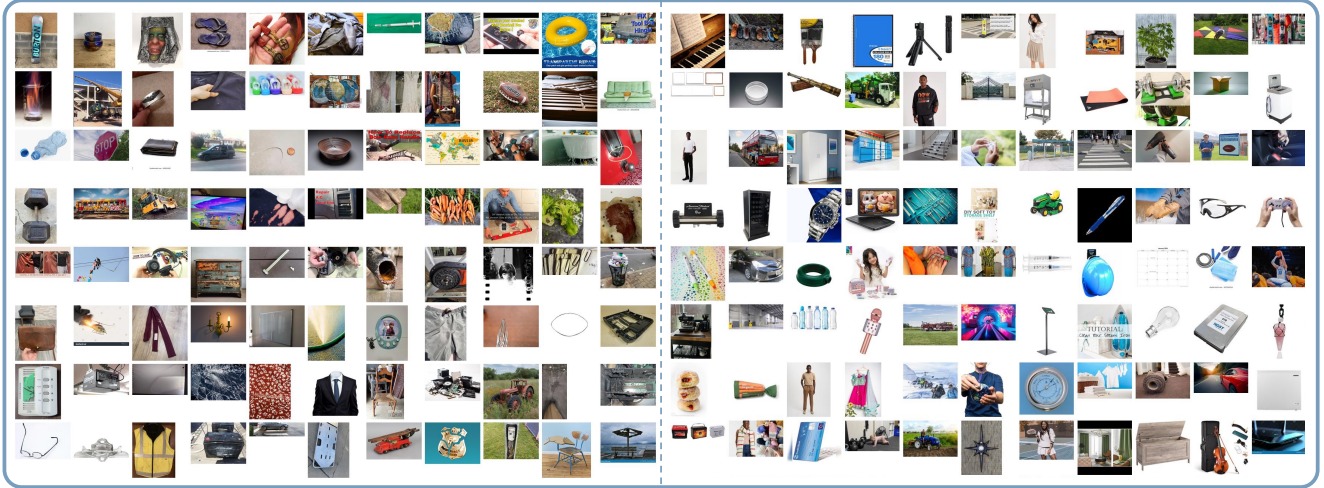


Figure 9. Overview of the gallery for in-the-wild image samples in WebAD. The images on the left side are anomalous, while the right side is for normal images. The links to download these images will be released to avoid copyright issues.

```

system_prompt = """
You are an intelligent visual anomaly assistant, and you are seeing a single image. What you see are provided with a comprehensive image description and a deep analysis of the anomalies contained, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a user asking about this photo. The answers should be in a tone that the assistant is seeing the image and answering the question.
Ask diverse questions and give corresponding answers.

The questions should be related to the anomalies in the image, including the presence of the anomalies, the reasoning, the nature, the location, the possible causes, the potential impacts, the solutions for current state, and the mitigation suggestions in the future. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image.
Do not ask any question that cannot be answered confidently.

The conversation should be coherent and informative, covering various aspects of the image and the anomalies present. The conversation should be concise, engaging and informative, providing a detailed understanding of the image and the anomalies detected.

The output should be strictly formatted into JSON format.
"""

----- the in-context prompt -----
"""
Example1 = ### Comprehensive Image Description:\n
The image shows two oranges placed on a wooden table. The oranges are partially wrapped in a clear plastic bag. The orange on the left has a clear and significant mold growth covering nearly half of its surface. The mold is white and green ...

Conversation1 =
{
  "conversation": [
    {
      "user": "Please describe the image you are seeing.",
      "assistant": "The image features two oranges on a wooden table, partially wrapped in a clear plastic bag. The left orange has a significant mold growth covering nearly half of its surface, while the right orange appears healthy with no visible signs of mold."
    },
    {
      "user": "Are there any anomalies for the oranges?",
      "assistant": "Yes, the anomaly in the image is the mold growth on the left orange, characterized by a fuzzy, white, and green layer that differs from the healthy appearance of the right orange."
    },
    {
      "user": "Why do you think the left orange is anomalous?",
      "assistant": "The left orange is anomalous due to the mold growth covering a substantial portion of its surface, indicating deterioration and unsuitability for consumption."
    }
  ]
}

Example2 = **Comprehensive Image Description**:\n
The image depicts a storefront window with shattered glass. It is part of a larger facade that includes additional glass panels on either side. The main window is severely damaged with multiple cracks spreading outward from a central impact point...

Conversation2 = """

```

Figure 10. Prompt template for generating multi-round conversation in Anomaly-Instruct-125k (modified from the template of LLaVA [57]).

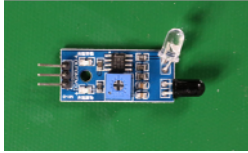
in future research. One can change the base model to other open-sourced MLLMs to resolve the wrong classification issue. Also, we found that the diversity of the anomaly type is very limited in existing industrial anomaly datasets (mainly 'crack' or 'broken'), causing the assistant to fail to provide fine-grained anomaly reasoning or description for

unseen anomaly features. Therefore, a more diverse industrial anomaly detection dataset is urgently required. Similar to other traditional MLLMs, Anomaly-OV only utilizes the output visual tokens from the last layer of the visual encoder as the input for LLM. However, anomaly detection is highly dependent on low-level visual clues. Hence, **forwarding**

VisA AUROC	capsules 98.6 pcb2 87.7	fryum 93.4 pcb3 86.6	cashew 93.9 pcb4 90.0	macaroni1 91.1	macaroni2 70.3	candle 96.5	pipe fryum 99.0	chewinggum 96.4	pcb1 90.3
MVTec AD AUROC	bottle 95.6 carpet 98.9	cable 88.8 metal nut 77.7	grid 99.2 wood 100.0	capsule 92.2 zipper 99.6	tile 95.5 leather 100.0	pill 92.3 toothbrush 94.2	screw 98.0	transistor 88.0	hazelnut 91.2

Table 8. Per-class image-level AUROC of the anomaly expert of Anomaly-OV on VisA and MVTec AD.

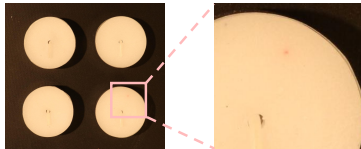
PCB Example: Bent Transparent LED



User	Is there an anomaly in the image?
GPT-4o	The image shows an infrared sensor module, which appears to be intact and properly assembled. There doesn't seem to be any visible anomaly or damage to the components.
LLaVA-OV	The image shows an electronic circuit board with various components, including a red LED that is not illuminated. There doesn't appear to be anything obviously anomalous in the image.
Ours	Yes, the anomaly in the image is the bent LED on the circuit board, which is not aligned properly with the board's surface.

Table 9. Additional results on VisA-D&R (PCB).

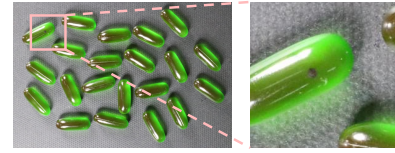
Candle Example: Pink Discoloration



User	Is there an anomaly in the image I should know about?
GPT-4o	Yes, there is an anomaly in the image. The candle in the top left corner has its wick facing the opposite direction compared to the other three candles.
LLaVA-OV	Based on the image, there is no apparent anomaly among the candles. They are all similar in size and shape, with a consistent hole in the center for the wick. The coloration appears uniform across the candles, and there are no visible signs of melting or irregularities that would suggest an anomaly.
Ours	Yes, the anomaly is the presence of a small red spot on the bottom right candle, which contrasts with the otherwise uniform white color.

Table 10. Additional results on VisA-D&R (Candle).

Capsule Example: Dark Spot



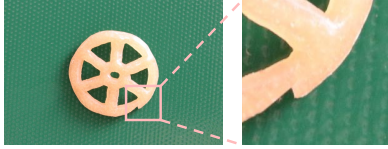
User	Is there an anomaly present in the image?
GPT-4o	Yes, there is an anomaly in the image. One of the capsules is a different color; it is brown instead of green.
LLaVA-OV	Based on the image alone, there is no clear indication of an anomaly. The pills appear to be uniformly shaped and colored, with no visible defects or irregularities. However, without additional context or a point of reference for size, it's difficult to determine if there are any anomalies that might not be immediately apparent from the image.
Ours	Yes, there is an anomaly present in the image. One of the capsules has a darker spot on its surface, which contrasts with the otherwise uniform green color.

Table 11. Additional results on VisA-D&R (Capsules).

recent paper: "Dense Connector for MLLMs" [96]) should be a possible solution for performance improvement.

multi-level features from different layers to the LLM (as

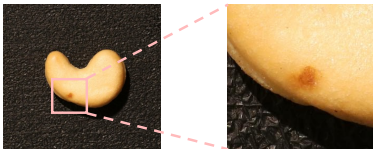
Fryum Example: Break Edge



User	Is there any anomaly in the image?
GPT-4o	The image shows a wheel-shaped snack on a green surface. There doesn't appear to be any obvious anomaly.
LLaVA-OV	The image shows a circular object with a pattern that resembles a wheel or a gear, placed on a green textured surface. There doesn't appear to be any obvious anomaly in the image; it seems to be a simple photograph of an object against a background. If there is something specific you are referring to as an anomaly, please provide more details or clarify what aspect you are questioning.
Ours	Yes, there is an anomaly in the image. The pasta wheel has a noticeable crack on its lower right side, disrupting its otherwise smooth and continuous surface.

Table 12. Additional results on VisA-D&R (Fryum).

Cashew Example: Brown Spot



User	Does the image contain any anomaly?
GPT-4o	The image shows a cashew nut on a textured black surface. There are no visible anomalies; the cashew appears normal.
LLaVA-OV	Based on the image provided, there are no visible anomalies. The object appears to be a single, intact cashew nut with a smooth surface and a natural brown coloration that is typical for cashews. There are no signs of damage, such as cracks or holes, and the background does not show any irregularities that would suggest an anomaly in the context of the image.
Ours	Yes, there is a noticeable anomaly on the cashew nut: a small brown spot located near the center of the nut.

Table 13. Additional results on VisA-D&R (Cashew).

In-the-Wild Example: Graffiti Road Sign



User	Describe the anomaly in the image.
Anomaly-OV	The anomaly in the image is the graffiti on the highway sign, which includes black spray paint marks and scribbles over the word 'Seattle' and other parts of the sign.
User	What is the reason for the anomaly?
Anomaly-OV	The graffiti is likely the result of vandalism, where individuals have used spray paint to deface public property for personal expression or as a form of protest.
User	How does the anomaly affect the functionality of the object?
Anomaly-OV	The graffiti affects the sign's functionality by potentially obstructing the visibility of important information, such as the word 'Seattle,' which could confuse drivers and lead to navigational errors.
User	How to avoid the anomaly in the future?
Anomaly-OV	To avoid such anomalies, regular maintenance and cleaning of signs should be implemented. Additionally, community awareness programs can help reduce vandalism by promoting respect for public property.

Table 14. In-the-wild results for an unseen object (Road Sign).

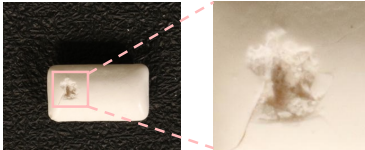
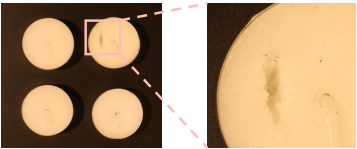
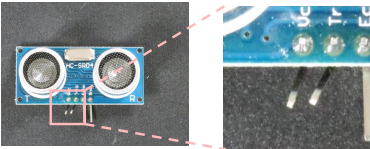
Failure Example 1: Wrong Classification	
	
User	Is there any anomaly visible in the image?
Anomaly-OV	Yes, there is a noticeable anomaly on the soap bar: a hole with jagged edges located towards the left side.
Failure Example 2: General Description	
	
User	Are there any anomalies present in the image with the set of candles?
Anomaly-OV	Yes, there is an anomaly present in the top right candle. It has a noticeable crack on its surface, which is not typical for a smooth, intact candle.
Failure Example 3: Wrong Reasoning	
	
User	Are there any anomalies in the image?
Anomaly-OV	Yes, the anomaly in the image is the absence of a crystal oscillator on the HC-SR04 ultrasonic sensor module.

Table 15. Failure results of Anomaly-OV on VisA-D&R.