

# See Further, Think Deeper: Advancing VLM’s Reasoning Ability with Low-level Visual Cues and Reflection

## Supplementary Material

In this supplementary material, we provide further details on the dataset construction process and experimental results. Section A introduces the dataset construction pipeline, including prompt design and data template formulation used during data generation. Section B presents qualitative visualizations of model outputs, highlighting representative cases that further demonstrate the strengths and limitations of ForeSight. Section C provides detailed metrics on the model’s performance in general VQA tasks.

### A. Data Construction Pipeline

**Visual Tools Annotation.** We propose a novel framework for Visual Cues and Visual Feedback chain (V<sup>2</sup>-CoT) data construction, designed to yield a high-quality, structured reasoning chain. The complex visual reasoning process is decoupled into discrete stages. Subsequently, manually configured rules are employed to integrate the data outputs from each stage, thereby assembling a comprehensive reasoning chain. This mechanism ultimately achieves a synergistic or closed-loop functionality encompassing “thinking with visual cues” and “reflection with visual feedback.” Our approach utilizes a specific, structured prompt that guides the model toward tool-based reasoning, shown in *Prompt 1*. This guidance enables the model to appropriately invoke designated visual tools, interpret the generated visual feedback, and seamlessly connect these intermediate steps to derive the final answer. This mechanism fundamentally realizes the principle of “thinking with visual cues”. The Canny tool, Zoom-In tool, and Color tool are referred to as <CANNY>, <ROI>, and <COLOR>, respectively. All tools are parameter-free, except for <ROI>. For tool-free data instances, *Prompt 2* instructs the model to output image-based deep reasoning exclusively to derive a definitive result. This collaborative process, which integrates automated annotation with rule validation, leads to a substantial improvement in both the quality and the completeness of the generated visual reasoning data.

**Visual Reflection Annotation.** The generation of “Reflection with Visual Feedback” chains is achieved through a dedicated, automated pipeline. This process is mainly accomplished by *Prompt 3*, which strategically guides the VLM to output a high-quality, efficient reflection chain. The principle for the reflection stage is strictly conditional: if the predicted answer coincides with the ground-truth answer, the reflection process must produce a correct verification result based on the provided mask image. Conversely, in the event of a mismatch, the reflection mechanism is mandated

to diagnose the error and propose a necessary correction. To ensure that the VLMs fully understand the reflection logic, our dataset specifically includes both the reasoning chains for reflecting correctly and the reasoning chains for reflecting incorrectly/proposing corrections.

**Resulting Data Format.** Following the two rounds of annotation (visual reasoning and visual reflection) the collected data is subjected to rigorous rule matching and cleaning. The final structured data annotations are then generated, as visually demonstrated in Fig. 1. The resulting dataset encapsulates a complete closed-loop reasoning. This process is initialized by a thought sequence that alternates between tool calls and the analysis of visual tool feedback, ultimately producing a draft answer. This draft answer subsequently serves as the basis for creating a mask on the original image, which provides the necessary visual feedback. Finally, the model leverages this generated visual feedback for in-depth reflection and validation, ultimately converging upon the final answer.

```
Response Template
<think>Reasoning step: analyze the image and
decide tool usage.</think>
<tool_call><tool_name>CANNY</tool_name>
<params>{}</params></tool_call>
<tool_response>Return the processed image
from the applied tool.</tool_response>
<think>Refine reasoning with updated visual
information.</think>
Iteratively call tools...
<draft_answer>Output initial bounding box
and feature prediction.</draft_answer>
<verify>Check correctness and consistency
using a mask.</verify>
<final_answer>Provide the verified final
bounding box and features.</final_answer>
```

Figure 1. An example of data annotation.

### B. Qualitative Analysis

**Cases of Visual Tools.** This section provides qualitative examples to demonstrate the effectiveness of our proposed method. Fig. 2, Fig. 3, Fig. 4, and Fig. 5 illustrate the outcomes of various tool combinations applied to natural images, while Fig. 6, Fig. 7, and Fig. 8 showcase these effects

## Prompt 1: Think With Tool Annotation Prompt

### Character description

You are an image analysis expert. You have the ability to analyze images, locate objects, and identify visual character.

### Terminology Notes:

- Target Character is a visual characteristic of an object in an image that is different from other objects.
- Target region is the coordinates of an object in an image that has different target characteristics from other objects.

### Task Description

- Using the provided auxiliary information, which includes the tools, target region, and target character, analyze and explain the reasons why the <Target Character> in the image's <Target Region> differs from the surrounding region. Explain why each tool from the auxiliary information was chosen and describe the results obtained.
- Use only the tools listed in the auxiliary information.
- Focus exclusively on the target region and feature.
- Provide a logical, coherent analysis following professional standards.

### Analysis Steps:

1. Initial Observation: Describe the general content of the image.
2. Tool Selection Reasoning: Explain why the first tool in the list was selected based on the initial observation.
3. Tool Application Results: Describe the results from using the first tool.
4. Subsequent Analysis: If multiple tools are listed, repeat steps 2 and 3 for each tool.
5. Conclusion: Based on the above findings, the target region and Target Character can be derived.

### Output Requirements:

1. Language: Use English.
2. Relevance: Ignore irrelevant information and avoid unnecessary details.
3. Format: Use natural and fluent language without timestamps or extra instructions.
4. Output Structure: Clearly separate each tool's description and analysis sequentially.
5. Output Format: Use natural, fluent language. Provide a complete explanation without timestamps or additional instructions. Use '<>' to mark tools; choose from: <CANNY>, <ROI>, <COLOR>. The sentence structure should be: "Based on the available information, we analyzed that we need to use <Tool 1>, and then use <Tool 2> for further analysis. <Attribution analysis> shows that the chosen tools are sufficient to draw a conclusion."
6. Do not include bounding boxes <Target Region> or Target Characters <Target Character> during the analysis; these are only shown in the final result.
7. Descriptions of different tools should be separated and presented sequentially.
8. The tags <think> and </think> cannot contain symbols such as <CANNY>, <ROI>, or <COLOR>.
9. Please output the result in the following format:  
<think><Description, need to use tool xx></think><tool xx><think><tool xx result description (whether other tools are used)></think>(if other tools are used: <tool \*\*><think><tool \*\* result description, final result></think>)

### Example:

- Example 1

## Prompt 2: Think Without Tool Annotation Prompt

### Character description

You are an image analysis expert. You have the ability to analyze images, locate objects, and identify visual character.

### Task Description

Using the provided image and auxiliary information (<Target Region>, <Target Character>), analyze and explain the reasons why the <Target Character> in the image’s <Target Region> differs from the surrounding region.

Note: Assume that the auxiliary information (<Target Region>, <Target Character>) is correct. Do not output information about other regions or features. Your output should simulate the process of inferring the auxiliary information based solely on the given image, without prior knowledge of the <Target Region> or <Target Character>. Ensure your explanation is logically sound, well-organized, and adheres to professional analysis standards.

### Analysis Steps:

Briefly describe the image content and explain the reasoning behind identifying the <Target Region> and <Target Character>, ultimately leading to the analysis of these features.

### Output Requirements:

1. Language: Use English.
2. Filter irrelevant information: Ignore any irrelevant special characters and redundant information.
3. Coherence: Provide only explanations relevant to target identification; avoid unnecessary descriptions.
4. Output Format: Use natural and fluent language. Provide a complete explanation; no timestamps or other instructions are needed.
5. Do not include the target region bounding box <Target Region> or target feature <Target Character> in the analysis process.
6. Sentence structure: <description>, conducting <causal analysis>, we can identify the result. Please output the result in the following format: <think><Description, Analysis, Conclusion></think>

### Example:

- Example 1

Table 1. More details on VQA. FP is Fine-grained Perception.

	Metric	Qwen2.5-VL 72B*	Qwen2.5-VL 7B*	ForeSight 7B	$\Delta$
MME	ALL	2433.7	2288.2	2326	+37.8
	Perception	1702.9	1675.3	1704	+28.7
	Reasoning	730.7	612.9	622	+9.1
MMBench	ALL	89.18	82.47	81.50	-0.97
	Attribute Reasoning	87.94	78.89	78.68	-0.21
	Coarse Perception	89.86	88.18	86.47	-1.71
	FP-Cross Instances	90.91	74.83	75.50	+0.67
	FP-Single Instances	92.83	87.37	88.39	+1.02
	Logical Reasoning	80.51	72.88	71.97	-0.91
	Relation Reasoning	86.96	80.87	81.58	+0.71

on synthetic images. Specifically, when the target region is perceptually ambiguous or difficult to discern, the Zoom-

in tool is employed (Fig. 3). To confirm subtle variations in the shape of the target object, the Canny tool provides crucial corroborating evidence (Fig. 2, Fig. 6). Furthermore, as illustrated in Fig. 7, challenging color differentiation in the original image is mitigated by color conversion, which renders objects of disparate colors clearly discernible, thus facilitating the model in reaching the correct conclusion. Finally, when a single visual tool proves insufficient to confirm a distinction, a strategic combination of multiple tools significantly aids the model’s judgment by aggregating complementary visual feedback (Fig. 4, Fig. 5, Fig. 8).

**Cases of Visual Reflection.** Fig. 9 and Fig. 10 demonstrate that ForeSight acquires the capability to execute visual reflection based on visual feedback during the infer-

### Prompt 3: Visual Reflection Annotation Prompt

#### Role Description:

You are an image analysis expert capable of recognizing and describing image content based on visual character.

#### Task Description:

-Use the available information (predictions) and auxiliary information (ground truth) to analyze and interpret whether the predicted answers in the available information are correct. By judging the masked part of the expected answer in the image, determine if the prediction is accurate, referencing the correct answer in the auxiliary information.

-If the predicted answer matches the ground truth, it is correct. In this case, there should be no object of inconsistent visual character in the unmasked parts. The masked portion should represent a complete, single instance, and this instance should differ from the unmasked instances in the auxiliary information's features.

-If the predicted answer is different from the ground truth, the prediction result is wrong. In this case, there is a whole or part of the object with inconsistent visual character in the unmasked part.

-Note: Auxiliary information is the ground truth. Do not output the auxiliary information. When outputting, you should pretend you only used the available information and did not know the auxiliary information in advance. Ensure the explanation is logical and clear, and that the final output meets professional analysis requirements.

#### Analysis Steps:

1. First, analyze whether there are targets with inconsistent visual character in the unmasked part of the image in the available information. visual character mainly include Orientation, Color, Focus, Shape, Size, Location, and Pattern.

2. Secondly, analyze whether the masked part of the image in the available information can completely cover the instance.

3. If both of the above points are true, then the prediction is correct, and the output is that the prediction is correct. Otherwise, the prediction is incorrect, and the output should suggest a corrected answer.

#### Output Requirements:

1. Language: Use English.

2. Filter irrelevant information: Ignore any irrelevant special characters and redundant information.

3. Logical coherence: Provide only explanations relevant to object recognition; avoid unnecessary descriptions.

4. Output format: Use natural and fluent language. Provide a complete explanation; no timestamps or other instructions are needed. Do not output terms like "ground truth" or "correct answer" during the analysis.

5. Sentence structure: <Description of visual character consistency>, <Description of the mask image>, <Analysis and conclusion>.

6. Please output the result in the following format: <verify><Description, Analysis, Conclusion></verify>

#### Example:


- Example 1
- Example 2

ence phase, allowing it to successfully correct the incorrect draft answer and ultimately arrive at the correct conclusion. In particular, Fig. 7 illustrates that the model retains the ability to identify errors even when the bounding box provided does not fully encapsulate the object's boundaries,

thereby enabling the correction of the initial draft answer.

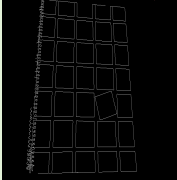
**Failed Cases and Limitations.** Fig. 5 presents an example of a tool combination where the Canny tool appears to be redundant, which consequently contributed to an erroneous size assessment in the generated answer. Fig. 4 shows

a shape misjudgment, and crucially, the reflection module failed to detect the incomplete object encapsulation by the bounding box. These limitations are attributed partly to the relatively small size and weak foundational capabilities of the Qwen2.5-VL-7B model employed in this study. Future work will focus on adopting models with superior basic visual capabilities and expanding the coverage to encompass a broader spectrum of vision tasks.




Carefully examine the input image: `<image>` and identify an object that is significantly different from the others.

`<think>`Initial observation reveals a green spiral bag placed on a white textured surface. The target region appears to focus on a section of the bag where the grid pattern is slightly tilted. To better understand this feature, we need to use the edge detection tool.`</think>`  
`<tool_call>` {"name": "CANNY", "arguments": {}} `</tool_call>`



`<think>`The application of the edge detection tool highlights the outlines of the squares and their edges, showing clear distinctions between them. However, the orientation of the target region's squares appears irregular compared to the rest, indicating a unique pattern. Based on the available information, we analyzed that the edge detection tool is sufficient to draw a conclusion.`</think>`  
`<draft_answer>`([549, 382, 684, 507], 'orientation')`</draft_answer>`



`<verify>`Correct. The verification tool indicates that the unmasked parts of the image maintain consistent visual features such as orientation, color, and shape across all instances. The masked area completely covers the intended instance without leaving any part of it visible. Therefore, the prediction is correct, and the final answer can be given directly.`</verify>`  
`<final_answer>`([549, 382, 684, 507], 'orientation')`</final_answer>` ✓

Final answer visualization  
 GT is blue Prediction is red


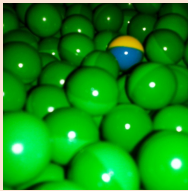


Figure 2. An example of using the Canny tool in natural image.

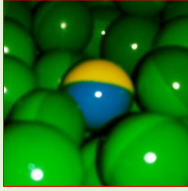
### C. Details of VQA

Tab. 1 provides more detailed VQA results. MME assesses perception (max 2000) and cognition (max 800), while

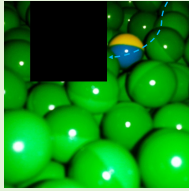


Carefully examine the input image: `<image>` and identify an object that is significantly different from the others.

`<think>`Initial observation reveals an image filled with green candies, with one distinctively different candy in the center. The target region appears slightly unclear, necessitating a closer look. To address this, we need to use the zoom in tool to focus on the target region for better analysis.`</think>`  
`<tool_call>` {"name": "ROI", "arguments": {"bbox\_2d": [539, 167, 735, 353]}} `</tool_call>`



`<think>`After applying the zoom in tool, the central area becomes clearer, revealing a candy with a unique yellow and blue pattern among the surrounding green candies. This distinct coloration makes it stand out significantly. The chosen tool is sufficient to draw a conclusion.`</think>`  
`<draft_answer>`([137, 0, 539, 421], 'color,pattern')`</draft_answer>`



`<verify>`Incorrect. The verification tool indicates that the unmasked part of the image shows a candy with a distinct yellow and blue pattern, which is visually inconsistent with the surrounding green candies. The predicted coordinates do not align with the unique object's location. Additionally, the masked area does not completely cover the intended instance as per the visual features described. Therefore, the prediction is incorrect and requires adjustment to accurately reflect the target object.`</verify>`  
`<final_answer>`([539, 167, 735, 353], 'color,pattern')`</final_answer>` ✓ ✓

Final answer visualization  
 GT is blue Prediction is red

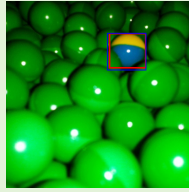
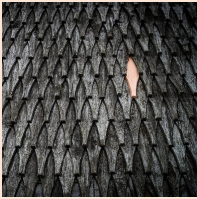
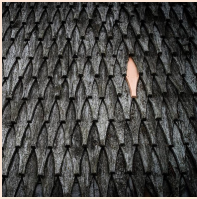




Figure 3. An example of using the Zoom-In tool in natural image.

MMBench evaluates 6 dimensions via iteratively questioning. ForeSight improves MME Reasoning by 9.1 and MMBench Relation Reasoning by 0.71, indicating that its benefits extend beyond basic perception grounding to visual reasoning. The gains mainly come from enhanced shape and positional accuracy, whereas the relatively modest overall



Carefully examine the input image:  and identify an object that is significantly different from the others.

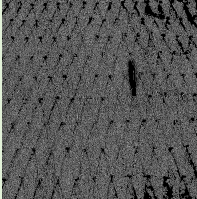


Carefully examine the input image:  and identify an object that is significantly different from the others.

**Figure 4: Using Canny and ROI tools in natural image.**

*<think>*Initial observation reveals a tree with a unique pattern of leaves arranged along its edges. A distinct section of the tree appears to have a lighter-colored, more elongated shape compared to the surrounding darker, smaller leaves. To better understand this pattern, we need to use the edge detection tool to highlight the contours and shapes within the target region.*</think>*

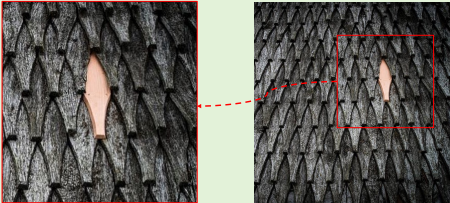
*<tool\_call>* {"name": "CANNY", "arguments": {}} *</tool\_call>*



*<think>*The application of the edge detection tool has successfully outlined the edges of the tree's leaves, providing a clear distinction between the target region and its surroundings. The target region exhibits a distinct rectangular shape with smooth sides, differing significantly in size and contour compared to the surrounding elements.

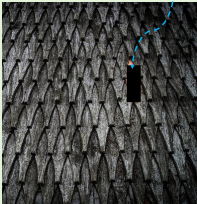
However, the central area remains somewhat ambiguous, necessitating a closer inspection.*<think>*

*<tool\_call>* {"name": "ROI", "arguments": {"bbox\_2d": [627, 304, 696, 480]}} *</tool\_call>*



*<think>*Utilizing the zoom in tool allows us to focus on the target region, revealing finer details. The rectangular shape within the tree is more prominent, and its lighter color contrasts sharply with the darker wood around it. The analysis confirms that the chosen tool is sufficient to draw a conclusion about the target feature.*</think>*

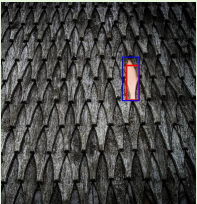
*<draft\_answer>* ([627, 304, 696, 480], 'color,shape') *</draft\_answer>*



*<verify>*Correct. The verification tool indicates that all objects in the unmasked part of the image maintain consistent visual features such as color and shape. The masked area completely covers the intended instance without leaving any part of it visible. Therefore, the prediction is correct, and the final answer can be given directly.*</verify>*

*<final\_answer>* ([627, 304, 696, 480], 'color,shape') *</final\_answer>*


Final answer visualization  
GT is blue Prediction is red



**Figure 5: Using all tools in natural image.**


*<think>*Initial observation reveals a grid of square tiles arranged neatly, predominantly in a light pink hue. The central tile stands out due to its distinct color and size compared to the surrounding ones. To analyze the target region's features such as color, shape, and size, edge detection is necessary to delineate the shapes and contours accurately.*</think>*

*<tool\_call>* {"name": "CANNY", "arguments": {}} *</tool\_call>*



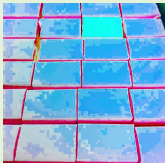
*<think>*The application of the Canny tool highlights the edges of the squares, providing clear outlines that help distinguish the target region's boundaries. However, the central area remains somewhat ambiguous due to the uniformity of the colors, necessitating further analysis with the ROI tool.*</think>*

*<tool\_call>* {"name": "ROI", "arguments": {"bbox\_2d": [441, 69, 715, 206]}} *</tool\_call>*




*<think>*Utilizing the ROI tool allows for a closer inspection of the target region, revealing a clearer view of the square's distinct features. The magnified view confirms the presence of a larger square that contrasts sharply with the surrounding ones. Its unique size and position make it a focal point within the grid. To further emphasize the target feature, a color analysis tool is required.*</think>*

*<tool\_call>* {"name": "COLOR", "arguments": {}} *</tool\_call>*



*<think>*Applying the COLOR tool enhances the contrast between the target square and its surroundings by converting the image into a blue hue, making the target square's edges more prominent. This adjustment facilitates a detailed examination of the target region's size, which is indeed larger than the surrounding squares. Based on the available information, we analyzed that we need to use the edge detection tool, and then use the zoom-in tool for further analysis. The target square in the target region has a distinct size compared to the surrounding squares, indicating its uniqueness within the grid.*</think>*

*<draft\_answer>* ([441, 69, 715, 206], 'color,size') *</draft\_answer>*



*<verify>*Correct. The verification tool indicates that all the unmasked parts of the image maintain consistent visual features such as orientation, shape, and pattern. The masked area completely covers the intended instance without leaving any part of it visible. Therefore, the prediction is correct, and the final answer can be given directly.*</verify>*

*<final\_answer>* ([441, 69, 715, 206], 'color,size') *</final\_answer>*

Final answer visualization  
GT is blue Prediction is red





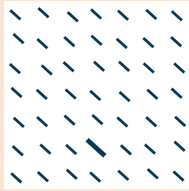
Figure 5. An example of using all tools in natural image.

Figure 4. An example of using a combination of Canny and ROI tools in natural image.

improvement in VQA is likely due to the stronger reliance of some tasks on common sense and the performance drop in coarse perception caused by tool-induced changes in image style and quality.

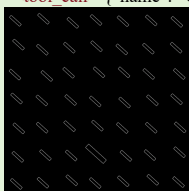


Carefully examine the input image: <image> and identify an object that is significantly different from the others.



<think>Initial observation reveals a patterned image with dark blue rectangular shapes scattered across a white background. The target region appears to focus on one of these rectangles, specifically examining its size relative to others. Edge detection is essential to delineate the boundaries of the rectangles accurately.</think>

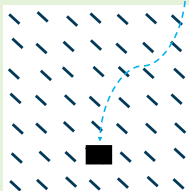
<tool\_call> {"name": "CANNY", "arguments": {}} </tool\_call>



<think>The application of the Canny edge detection tool has successfully outlined all the rectangular shapes within the image, including the target region. The edges are now clearly defined by white lines against a black background, allowing for precise measurement of the rectangle's dimensions.

The analysis confirms that the chosen tool is sufficient to draw a conclusion.</think>

<draft\_answer>[[431, 715, 568, 813], 'size']</draft\_answer>



<verify>Correct. The verification tool indicates that all unmasked objects maintain consistent visual features such as orientation, color, and size. The masked area in the predicted image completely covers the intended instance without leaving any part of it unmasked. Therefore, the prediction is correct, and the final answer can be given directly.</verify>

<final\_answer>[[431, 715, 568, 813], 'size']</final\_answer>

Final answer visualization

GT is blue Prediction is red



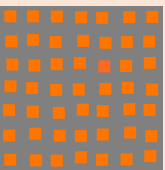


Figure 6. An example of using the Canny tool in synthetic image.

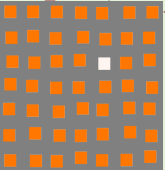


Carefully examine the input image: <image> and identify an object that is significantly different from the others.



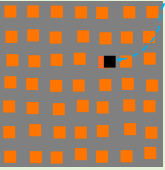
<think>First, a brief description of the image is provided. The image features a grid of orange squares on a gray background. The target region is a small section within this grid. To analyze the color feature of this region, we need to use the color conversion tool.</think>

<tool\_call> {"name": "COLOR", "arguments": {}} </tool\_call>



<think>After applying the color conversion tool, the target region's square appears distinctly lighter than the surrounding squares. This indicates a variation in color intensity or hue within the specified area. No further tools are needed as the color difference is clearly visible.

<draft\_answer>[[627, 323, 696, 392], 'color']</draft\_answer>



<verify>Incorrect. The verification tool indicates that the unmasked part contains a square with a different color compared to the rest, located at the top right section of the image. This suggests that the predicted answer is incorrect. Additionally, the masked part does not correspond to the intended target as per the visual features described. Therefore, the prediction is incorrect and requires adjustment to properly identify the target instance.</verify>

<final\_answer>[[598, 323, 666, 392], 'color']</final\_answer>

Final answer visualization

GT is blue Prediction is red

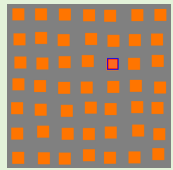



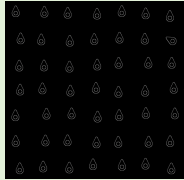
Figure 7. An example of using the Color tool in synthetic image.



Carefully examine the input image: <image> and identify an object that is significantly different from the others.

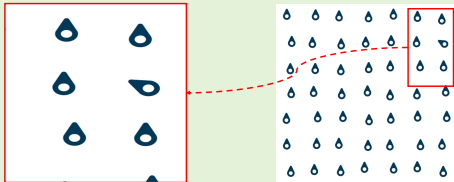
<think>Initial observation reveals a grid of water droplet icons arranged uniformly across a white background. The target region appears to focus on the droplets near the corner, specifically examining their orientation. Given the need to analyze shapes and contours, the first step involves using edge detection to delineate the boundaries of the droplets accurately.</think>

```
<tool_call>{"name": "CANNY", "arguments": {}} </tool_call>
```



<think>Applying the Canny edge detection tool highlights the contours of the droplets, making their shapes and orientations more apparent. The edges are now clearly defined, allowing us to distinguish individual droplets from their surroundings. However, the target region's droplet, located at the corner, is slightly tilted compared to others, which is evident by observing its position relative to the others. To better analyze this specific droplet's orientation, we need to use the ROI tool.

```
<tool_call>{"name": "ROI", "arguments": {"bbox_2d": [872, 196, 931, 245]}} </tool_call>
```



<think>Utilizing the ROI tool allows us to zoom into the target region, providing a clearer view of the droplet's orientation. It becomes evident that this droplet is indeed tilted to the right, differing from the other vertically aligned droplets. No further tools are necessary as the orientation feature is now evident.</think>

```
<draft_answer>([872, 196, 931, 245], 'orientation')</draft_answer>
```

<verify>Correct. The verification tool indicates that all unmasked water droplets maintain consistent visual features such as orientation, color, and size. The masked area in the predicted image completely covers the intended instance without affecting the uniformity of the surrounding objects. Therefore, the prediction is correct, and the final answer can be given directly.

</verify>

```
<final_answer>([872, 196, 931, 245], 'orientation')</final_answer>
```



Final answer visualization  
GT is blue Prediction is red

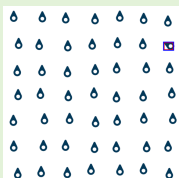
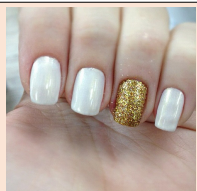


Figure 8. An example of using the Canny and ROI tool in synthetic image.



Carefully examine the input image: <image> and identify an object that is significantly different from the others.

<think>The image shows a close-up of a hand with nails painted, most of which are in a white color. One nail stands out due to its unique golden, glittery texture. This nail's distinct color and pattern make it visually separate from the surrounding white nails. The golden nail is the focal point due to its intricate glittery details, which contrast sharply with the plain white of the other nails. We can conclude that the difference is in the color and pattern, where the golden nail has a distinct golden color and glittery texture that sets it apart from the white nails.</think>

```
<draft_answer>([88, 255, 304, 500], 'color.pattern')</draft_answer>
```

<verify>Incorrect. The verification tool indicates that the unmasked part includes a nail with a white color and pattern, which is consistent with the intended target. However, the predicted coordinates do not align with the intended target as they mask a different area, specifically the nail with the white color and pattern instead of the targeted nail with the white color and pattern. Therefore, the prediction is incorrect and needs adjustment to match the intended target.</verify>

```
<final_answer>([529, 363, 755, 627], 'color.pattern')</final_answer>
```

Final answer visualization

GT is blue Prediction is red

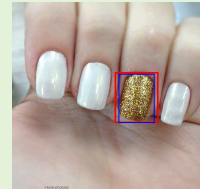
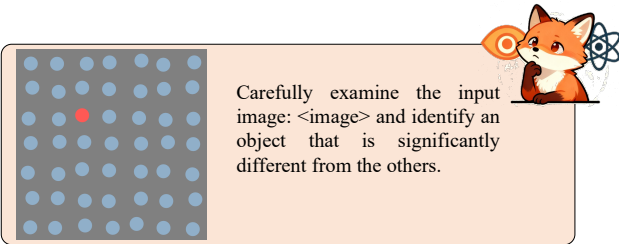


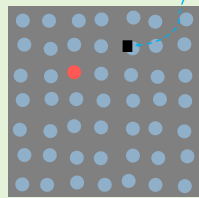
Figure 9. An example of visual reflection in natural image.



Carefully examine the input image: `<image>` and identify an object that is significantly different from the others.

`<think>`The image displays a grid of uniformly sized circles arranged in rows and columns against a neutral gray background. Most of the circles share a consistent light blue hue. However, one circle deviates from this uniformity by exhibiting a distinct red color. This particular circle is the focal point of interest, as it stands out due to its unique color feature. The primary distinguishing factor between this red circle and the surrounding blue ones is its color. Therefore, the conclusion is that the unique color of the target region sets it apart from the rest of the image.`</think>`

`<draft_answer>([[588, 176, 637, 235], 'color')</draft_answer>`



`<verify>`Incorrect. The verification tool confirms that the unmasked part contains a red circle which stands out due to its distinct color compared to the blue circles. However, the masked area does not correspond to the intended target as it is located at a different position. Therefore, the prediction is incorrect and needs to be corrected by adjusting the coordinates to match the location of the red circle.

`<final_answer>([[304, 304, 372, 372], 'color')</final_answer>` ✓

Final answer visualization

GT is blue Prediction is red

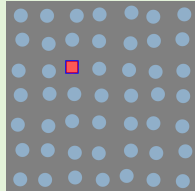


Figure 10. An example of visual reflection in synthetic image.