

# Improving Vision-language Models with Perception-centric Process Reward Models

## Supplementary Material

### 887 7. Prompts and Verification Demonstrations

888 In this section, we present the comprehensive prompts uti-  
889 lized by our Perception-centric PRM, **Perceval**, for the  
890 error-finding schema described in Section 3.1 (refer to  
891 Boxes 7.4 and 7.5). Notably, this configuration is identi-  
892 cal to the prompt provided to strong models (e.g., Gemini-  
893 2.5-Pro) for the automated annotation and verification pro-  
894 cess, as detailed in Section 3.1. Given an image, a query,  
895 and a corresponding response, Perceval generates a detailed  
896 verification rationale. We provide representative examples  
897 of these outputs below: Box 7.1 illustrates a verification  
898 instance for a **correct response**, while Box 7.6 demon-  
899 strates the detection of **errors**. Furthermore, as outlined in  
900 Section 4.1, Perceval is integrated into the Reinforcement  
901 Learning (RL) training loop to optimize policy models. The  
902 specific prompt employed for the policy model during this  
903 phase is provided in Box 7.2

#### Box 7.1: Sample Verification Output: Correct Response

```
<think>
- Claims identified in response: The main claim is that the color of
the object is black. The supporting thought process also mentions
a black computer monitor.
- Checking each claim against visuals: The question asks for the
color of the object. The most prominent object in the image is
the computer monitor on the desk. Visually inspecting this object
confirms it is black. While there is another laptop on the right,
the response correctly focuses on the main subject of the image.
- Evidence found: The central computer monitor is clearly black.
The response accurately identifies this color. The thought process
correctly identifies the object in question.
- Issues identified: No issues were found. The response correctly
answers the question based on the visual evidence.
</think>
<answer>
The response is correct
</answer>
```

#### Box 7.2: Prompt for Policy Model

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put within <answer> </answer> tags.

#### Box 7.3: Sample Verification Output: Response with Errors

```
<think>
- Claims identified in response:
1. The man is sitting on a bicycle.
2. The setting is an outdoor, relaxed environment, possibly near
a beach or coastal area.
3. There is a promotional surfboard behind the man.
4. The man is riding the bicycle.
- Checking each claim against visuals:
1. ***"The man is sitting on a bicycle."***: This is correct. The
image clearly shows a man seated on a red bicycle.
2. ***"The setting is an outdoor, relaxed environment, possibly
near a beach or coastal area."***: This is a reasonable interpreta-
tion. The ground is unpaved dirt, there are trees, and the lighting
suggests a sunny day, consistent with an outdoor, rural or coastal
location.
3. ***"There is a promotional surfboard behind the man."***: This
is incorrect. A close inspection of the image reveals that the ob-
ject behind the man is a large, black, circular object with a tex-
tured, woven pattern. It strongly resembles a tire, likely a bicycle
tire, rather than a surfboard. The man is holding a small, light-
colored object, which appears to be a phone or a small camera,
not a surfboard.
4. ***"The man is riding the bicycle."***: This is incorrect. The
man is stationary; his feet are on the pedals, and he is holding the
handlebars. His posture and the lack of motion blur confirm he is
not riding the bike.
- Evidence found:
- The man is sitting on a bicycle.
- The object behind him is a large, black, textured wheel/tire, not
a surfboard.
- The man's feet are on the pedals, indicating he is stationary.
- Issues identified:
- The response incorrectly identifies the object behind the man as
a surfboard.
- The response incorrectly states that the man is riding the bicycle
when he is clearly stationary.
</think>
<answer>
[
"The presence of a promotional surfboard behind him suggests a
possible link to surfing or beach activities.",
"The man is riding a bicycle."
]
</answer>
```

**Box 7.4: Error-Finding Prompt for Perceval (Part 1 of 2)**

```

""""# Image/Video Response Verification Task
## Your Role
You are a rigorous fact-checker who verifies text responses
against visual content.
## Task Overview
You will receive:
- One or more images/videos
- Questions about the visual content
- Text responses to those questions
Your job is to verify if the text responses accurately describe
what's shown in the images/videos.
## Verification Process
### Step 1: Analyze the Text Response
- Read through the text response carefully
- Identify all specific claims made in the response
- Break down claims into verifiable elements (objects, colors,
numbers, actions, locations, etc.)
- Note what the response says should be visible in the im-
ages/videos
### Step 2: Verify Claims Against Visual Content
- For each claim identified in Step 1, check the images/videos
- Look for evidence that supports or contradicts each claim
- Pay attention to:
- Factual inaccuracies (wrong colors, numbers, objects, etc.)
- Claims about things not visible in the images/videos
- Misinterpretations of what's shown
- Missing critical details that should have been mentioned
### Step 3: Document Your Process Record your verification
steps in '<think>' tags, including:
- What claims you identified in the text response
- How you checked each claim against the images/videos
- What you observed that supports or contradicts each claim
- Your reasoning for any inconsistencies found
[The prompt continues in the next box]

```

**Box 7.5: Error-Finding Prompt for Perceval (Part 2 of 2)**

```

[Continued from previous box]
## Output Format
**Required Structure:**
""
<think>
[Your detailed verification process]
- Claims identified in response: [list key claims to verify]
- Checking each claim against visuals: [go through each claim
systematically]
- [If using zoom tool: mention which areas you examined more
closely]
- Evidence found: [what supports or contradicts each claim]
- Issues identified: [list any problems found]
</think>
<answer>
[Python list of exact problematic sentences/phrases from the orig-
inal response, OR "The response is correct"]
</answer>
""
**For inconsistencies found:**
""python
[
"exact sentence or phrase from response that is incorrect",
"another exact sentence or phrase that has problems",
# etc. - use the original wording, do not paraphrase
]
""
**If no issues found:**
"" The response is correct ""
## Input Format
- **Question and Visual content:** question
- **Response to verify:** response
## Important Guidelines
- Be thorough but focus on significant inconsistencies
- Don't flag minor stylistic differences or reasonable interpreta-
tions
- Only report clear factual errors or missing critical information
- Use the zoom tool when you need to examine specific regions
more closely
- In the final answer, include the exact original sentences/phrases
that are problematic - do not paraphrase or rewrite them
—
**Now please provide your verification following this for-
mat **""

```

**Box 7.6: Sample Verification Output: Response with Errors**

```
<think>
- Claims identified in response:
1. The man is sitting on a bicycle.
2. The setting is an outdoor, relaxed environment, possibly near
a beach or coastal area.
3. There is a promotional surfboard behind the man.
4. The man is riding the bicycle.
- Checking each claim against visuals:
1. ***"The man is sitting on a bicycle."***: This is correct. The
image clearly shows a man seated on a red bicycle.
2. ***"The setting is an outdoor, relaxed environment, possibly
near a beach or coastal area."***: This is a reasonable interpreta-
tion. The ground is unpaved dirt, there are trees, and the lighting
suggests a sunny day, consistent with an outdoor, rural or coastal
location.
3. ***"There is a promotional surfboard behind the man."***: This
is incorrect. A close inspection of the image reveals that the ob-
ject behind the man is a large, black, circular object with a tex-
tured, woven pattern. It strongly resembles a tire, likely a bicycle
tire, rather than a surfboard. The man is holding a small, light-
colored object, which appears to be a phone or a small camera,
not a surfboard.
4. ***"The man is riding the bicycle."***: This is incorrect. The
man is stationary; his feet are on the pedals, and he is holding the
handlebars. His posture and the lack of motion blur confirm he is
not riding the bike.
- Evidence found:
- The man is sitting on a bicycle.
- The object behind him is a large, black, textured wheel/tire, not
a surfboard.
- The man's feet are on the pedals, indicating he is stationary.
- Issues identified:
- The response incorrectly identifies the object behind the man as
a surfboard.
- The response incorrectly states that the man is riding the bicycle
when he is clearly stationary.
</think>
<answer>
[
"The presence of a promotional surfboard behind him suggests a
possible link to surfing or beach activities.",
"The man is riding a bicycle."
]
</answer>
```