

## I. MULTIMODAL REFERENCE VISUAL GROUNDING APPENDIX

### A. More Training Details

The results of varying numbers of training epochs for the weight adapter of NIDS-Net are presented in Table I. We select 640 epochs as it provides stable and robust performance. AP, AP50, and AP75 are metrics used to evaluate object detection models, where AP is the average precision, AP50 is the precision at an IoU threshold of 0.5, and AP75 is the precision at an IoU threshold of 0.75. The object detection results of not using the weight adapter in NIDS-Net are presented in Table II.

TABLE I  
DETECTION RESULTS WITH DIFFERENT NUMBERS OF TRAINING EPOCHS.

Number of Training Epoch	80	160	320	640
AP	59.9	58.6	56.3	<b>60.6</b>
AP50	64.8	63.4	60.9	<b>65.5</b>
AP75	64.6	63.1	60.5	<b>65.2</b>

TABLE II  
DETECTION RESULTS WITH AND WITHOUT THE WEIGHT ADAPTER.

Visual Encoder	Using Weight Adapter	AP	AP50	AP75
DINOv2 [1]	×	58.0	62.7	62.5
	✓	60.6	65.5	65.2
Perception Encoder [2]	×	66.0	71.7	71.1
	✓	<b>73.3</b>	<b>79.7</b>	<b>79.1</b>

### B. Visual Grounding Prompts

We present the following visual grounding prompts for LVLMs.

#### Visual Grounding Prompts

**Image:**  $\text{\$}\{image\}\text{\$}$

**Prompt:** Provide the bounding box coordinates of the region this sentence describes:  $\langle ref \rangle \text{\$}\{language\ instruction\} \text{\$} \langle /ref \rangle$

### C. Object Description Generation

Given one representative image of an object, we show the prompt to generate the object information via a Large Vision-Language Model (LVLM). GPT-4o and GPT-4o-mini employ the same prompts.

#### GPT-4o-mini Prompts

##### System Prompt:

You are an expert at structured data extraction. You will be given a picture. Please extract information and convert it into the given structure.

**Image:**



**Prompt:**

You are given an image of an item on a flat surface (on a table, ground, etc.). Please first carefully read and understand the image in detail. If there are multiple items, only carefully look through one of them. Then, describe the item in detail by following the steps and format below.

1. Shape: Please describe the shape or type of the item, such as a bottle, bag, round item, square item, etc.
2. Colors: Please describe all the colors on or in the item, such as label colors, text colors, cover colors, etc. The item may be covered by multiple colors. Please describe all of them one by one. For example, bottle: transparent, liquid in the bottle: black, the main color of the bag: green, the text on the item: black, etc.
3. Texts: Please extract all texts on the item with the position and color of the text. For example, "ingredients: on the surface, black". If there is no recognized text, please only output "None".
4. Function: Please describe the usage of the item in the given picture.
5. Summary of the item: Please summarize the above descriptions in sentences one-by-one.

**GPT-4o-mini Output**

```
{
  "shape": "bottle",
  "colors": [
    {
      "description": "the main color of the liquid",
      "color": "brown"
    },
    {
      "description": "the color of the bottle",
      "color": "transparent"
    },
    {
      "description": "the color of the label",
      "color": "white"
    }
  ],
}
```

```

    {
      "description": "the color of the text on the
        label",
      "color": "red"
    },
    {
      "description": "the color of the cap",
      "color": "gray"
    },
    {
      "description": "the color of the text below the
        bottle",
      "color": "black"
    }
  ],
  "texts": [
    {
      "text": "Coke",
      "position": "on the label, red"
    },
    {
      "text": "Diet",
      "position": "on the label, red"
    },
    {
      "text": "20oz",
      "position": "below the bottle, black"
    },
    {
      "text": "DIET",
      "position": "below the bottle, black"
    }
  ],
  "name": "002_coca-cola_soda_diet_pop_bottle",
  "function": "This item is a beverage container holding
    Diet Coke, a low-calorie soft drink.",
  "summary": "The item is a bottle in the shape of a
    typical soft drink container. It features a brown liquid
    inside, with a transparent bottle and a white label.
    The label has red text that reads 'Coke' and 'Diet'.
    Additionally, there is black text below the bottle
    indicating '20oz' and 'DIET'. The cap of the bottle is
    gray. This bottle is used to store and serve Diet Coke."
}

```

#### GPT-4o Prompts

##### System Prompt:

You are an expert at structured data extraction. You will be given a picture. Please extract information and convert it into the given structure.

##### Image:



**Prompt:**

You are given an image of an item on a flat surface (on a table, ground, etc.). Please first carefully read and understand the image in detail. If there are multiple items, only carefully look through one of them. Then, describe the item in detail by following the steps and format below.

1. Shape: Please describe the shape or type of the item, such as a bottle, bag, round item, square item, etc.
2. Colors: Please describe all the colors on or in the item, such as label colors, text colors, cover colors, etc. The item may be covered by multiple colors. Please describe all of them one by one. For example, bottle: transparent, liquid in the bottle: black, the main color of the bag: green, the text on the item: black, etc.
3. Texts: Please extract all texts on the item with the position and color of the text. For example, "ingredients: on the surface, black". If there is no recognized text, please only output "None".
4. Function: Please describe the usage of the item in the given picture.
5. Summary of the item: Please summarize the above descriptions in sentences one-by-one.

**GPT-4o Output**

```
{
  "shape": "Container",
  "colors": [
    {
      "description": "Lid color",
      "color": "Yellow"
    },
    {
      "description": "Main body color",
      "color": "Yellow"
    },
    {
      "description": "Text color",
      "color": "White, Blue, Brown"
    }
  ],
}
```

```

    {
      "description": "Illustration color",
      "color": "Brown, Pink, White"
    }
  ],
  "texts": [
    {
      "text": "CHOCOLATE",
      "position": "Top left",
      "color": "White"
    },
    {
      "text": "Nesquik",
      "position": "Left side",
      "color": "Blue"
    },
    {
      "text": "made with REAL COCOA",
      "position": "Right side",
      "color": "White"
    },
    {
      "text": "no ARTIFICIALS",
      "position": "Right side",
      "color": "White"
    },
    {
      "text": "50 CALORIES",
      "position": "Bottom left",
      "color": "Blue"
    },
    {
      "text": "NET WT 20.1 OZ (1.25 LB) 570 g",
      "position": "Bottom",
      "color": "Blue"
    }
  ],
  "function": "The item is a container of chocolate-flavored powder for making chocolate milk or similar beverages.",
  "summary": "The item is a container with a yellow lid and body. It features white, blue, and brown text. Illustrations are in brown, pink, and white. Text includes 'CHOCOLATE', 'Nesquik', 'made with REAL COCOA', and more. It is used for making chocolate-flavored drinks.",
  "filename": "060_nesquik_chocolate_powder"
}

```

#### D. LLM Matching

For the two matching strategies, we present the following prompts to match the object descriptions of predicted objects with the referring expressions. GPT-4o and GPT-4o-mini are using the same prompts.

##### 1) Joint Matching:

**System Prompt:**

You are an expert in information matching. Your task is to match items from a given list of descriptions to corresponding inquiries based on relevance. Each inquiry only matches one item description and appears once in the final output.

Each item description includes positional information, where the first value represents the x-axis (horizontal position) and the second value represents the y-axis (vertical position). A higher x-axis value indicates the item is positioned further to the right. A higher y-axis value indicates the item is positioned lower.

Once you determine the matches, convert them into the specified output format.

**Prompt:**

Items' Description:

Item ID: 5:

- Description: `{"shape": "bottle", "colors": [{"description": "main color of the liquid", "color": "dark brown"}, {"description": "color of the bottle label", "color": "red"}, {"description": "color of the cap", "color": "burgundy"}, {"description": "color of the text on the label", "color": "white"}, {"description": "color of the establishment year", "color": "white"}], "texts": [{"text": "Dr Pepper", "position": "on the label, white"}, {"text": "Est. 1885", "position": "on the label, white"}, {"text": "23", "position": "on the label, white"}, {"text": "20 OZ", "position": "on the label, white"}], "name": "005_dr_pepper_soda_pop_bottle", "function": "This item is a beverage, specifically a carbonated soft drink, intended for consumption.", "summary": "The item is a bottle with a dark brown liquid inside, indicating a soda. The bottle label is red with white text. The cap is burgundy, and the item is identified as Dr Pepper, established in 1885. It contains 20 ounces of the beverage."}`
- Position: (438, 346)

Item ID: 6:

- Description: `{"shape": "bottle", "colors": [{"description": "main color of the bottle", "color": "orange"}, {"description": "cap color", "color": "blue"}, {"description": "label color", "color": "blue"}, {"description": "text color on the label", "color": "white"}, {"description": "text color for the size", "color": "black"}], "texts": [{"text": "FANTA", "position": "on the label, white"}, {"text": "orange", "position": "on the label, white"}, {"text": "20oz", "position": "below the bottle, black"}], "name": "006_fanta_orange_fruit_soda_pop_bottle", "function": "A carbonated soft drink flavored with orange.", "summary": "The item is a bottle in the shape of a standard soda bottle. It is predominantly orange with a blue cap and a blue label. The label features white text that reads 'FANTA' and 'orange'. Below the bottle, there is a black text indicating '20oz'. This item is a carbonated soft drink flavored with orange."}`
- Position: (327, 193)

Item ID: 7:

- Description: `{ "shape": "Bottle", "colors": [{ "description": "Main color of the bottle", "color": "blue" }, { "description": "Text color on the label", "color": "white" }, { "description": "Label color", "color": "black" } ], "texts": [{ "text": "POWER ADE", "position": "on the label, white" }, { "text": "MOUNTAIN BERRY BLAST", "position": "on the label, white" }, { "text": "50% MORE ELECTROLYTES", "position": "on the label, white" }, { "text": "20 oz", "position": "on the surface, black" } ], "name": "007_powerade_mountain_berry_blast", "function": "The item is a sports drink designed to provide hydration and electrolytes, commonly used during physical activity or exercise.", "summary": "The item is shaped like a bottle. The main color of the bottle is blue, with white text on the label and a black background. The texts include 'POWER ADE', 'MOUNTAIN BERRY BLAST', and '50% MORE ELECTROLYTES' in white, and '20 oz' in black on the surface. This is a Powerade Mountain Berry Blast drink, which is used for hydration and replenishment of electrolytes." }`
- Position: `(650, 316)`

Inquiries:

Inquiry ID: 1, Inquiry Content: the orange bottle.

Inquiry ID: 2, Inquiry Content: the middle one.

Inquiry ID: 3, Inquiry Content: bottle with a black cap. You are given a few inquiries.

Please find matched item for each inquiry and list all answers in the given format.

GPT-4o / GPT-4o-mini Output

```
{
  "matches": [
    {
      "inquiry_id": 1,
      "item_id": 6
    },
    {
      "inquiry_id": 2,
      "item_id": 5
    },
    {
      "inquiry_id": 3,
      "item_id": 7
    }
  ]
}
```

## 2) Independent Matching:

GPT-4o / GPT-4o-mini

### System Prompt:

You are an expert in information matching. Your task is to match items from a given list of descriptions to the given inquiry based on relevance. Each inquiry only matches one item

description and appears once in the final output.

Each item description includes positional information, where the first value represents the x-axis (horizontal position) and the second value represents the y-axis (vertical position). A higher x-axis value indicates the item is positioned further to the right. A higher y-axis value indicates the item is positioned lower.

Once you determine the matches, convert them into the specified output format.

**Prompt:**

Items' Description:

Item ID: 5:

- Description: `{"shape": "bottle", "colors": [{"description": "main color of the liquid", "color": "dark brown"}, {"description": "color of the bottle label", "color": "red"}, {"description": "color of the cap", "color": "burgundy"}, {"description": "color of the text on the label", "color": "white"}, {"description": "color of the establishment year", "color": "white"}], "texts": [{"text": "Dr Pepper", "position": "on the label, white"}, {"text": "Est. 1885", "position": "on the label, white"}, {"text": "23", "position": "on the label, white"}, {"text": "20 OZ", "position": "on the label, white"}], "name": "005_dr_pepper_soda_pop_bottle", "function": "This item is a beverage, specifically a carbonated soft drink, intended for consumption.", "summary": "The item is a bottle with a dark brown liquid inside, indicating a soda. The bottle label is red with white text. The cap is burgundy, and the item is identified as Dr Pepper, established in 1885. It contains 20 ounces of the beverage."}`
- Position: (438, 346)

Item ID: 6:

- Description: `{"shape": "bottle", "colors": [{"description": "main color of the bottle", "color": "orange"}, {"description": "cap color", "color": "blue"}, {"description": "label color", "color": "blue"}, {"description": "text color on the label", "color": "white"}, {"description": "text color for the size", "color": "black"}], "texts": [{"text": "FANTA", "position": "on the label, white"}, {"text": "orange", "position": "on the label, white"}, {"text": "20oz", "position": "below the bottle, black"}], "name": "006_fanta_orange_fruit_soda_pop_bottle", "function": "A carbonated soft drink flavored with orange.", "summary": "The item is a bottle in the shape of a standard soda bottle. It is predominantly orange with a blue cap and a blue label. The label features white text that reads 'FANTA' and 'orange'. Below the bottle, there is a black text indicating '20oz'. This item is a carbonated soft drink flavored with orange."}`
- Position: (327, 193)

Item ID: 7:

- Description: `{"shape": "Bottle", "colors": [{"description": "Main color of the bottle", "color": "blue"}, {"description": "Text color on the label", "color": "white"}, {"description": "Label color", "color": "black"}], "`



```
texts": [{"text": "POWER ADE", "position": "on the label, white"}, {"text": "MOUNTAIN BERRY BLAST", "position": "on the label, white"}, {"text": "50% MORE ELECTROLYTES", "position": "on the label, white"}, {"text": "20 oz", "position": "on the surface, black"}], "name": "007_powerade_mountain_berry_blast", "function": "The item is a sports drink designed to provide hydration and electrolytes, commonly used during physical activity or exercise.", "summary": "The item is shaped like a bottle. The main color of the bottle is blue, with white text on the label and a black background. The texts include 'POWER ADE', 'MOUNTAIN BERRY BLAST', and '50% MORE ELECTROLYTES' in white, and '20 oz' in black on the surface. This is a Powerade Mountain Berry Blast drink, which is used for hydration and replenishment of electrolytes."}
- Position: (650, 316)
```

Inquiry:

the orange bottle.

You are given an inquiry. Please find the best matched item and output the answer in the given format.

GPT-4o / GPT-4o-mini Output

```
{"item_id": 7}
```

## REFERENCES

- [1] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv:2304.07193*, 2023.
- [2] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed *et al.*, “Perception encoder: The best visual embeddings are not at the output of the network,” *arXiv preprint arXiv:2504.13181*, 2025.