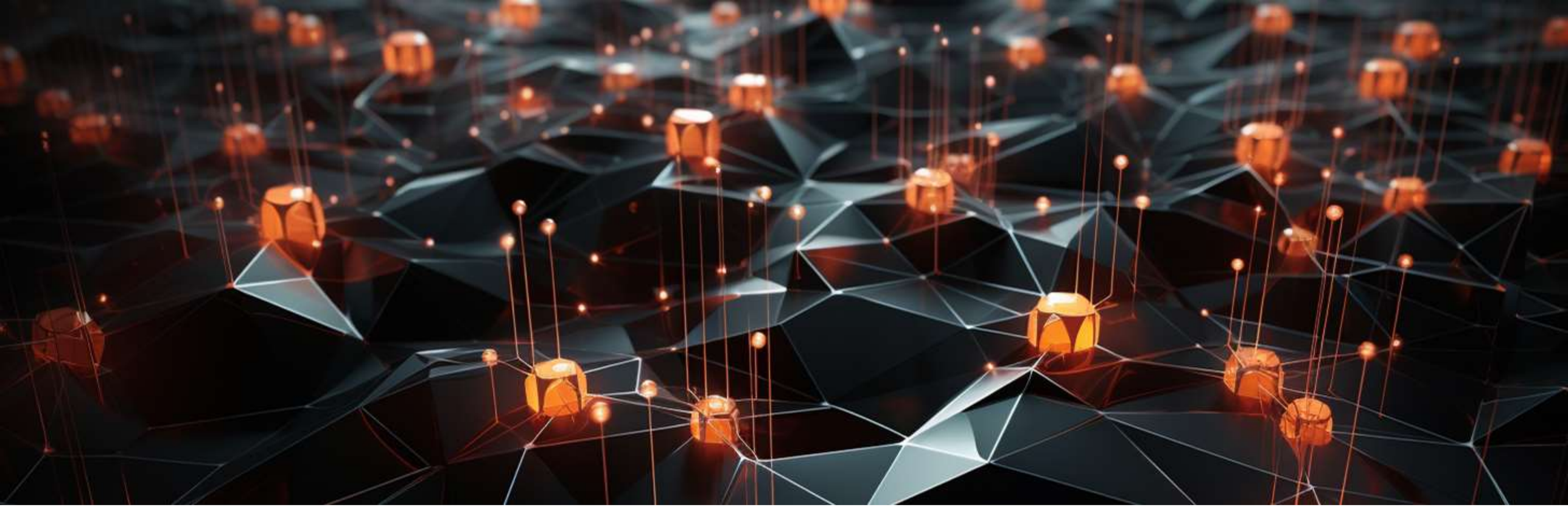


QWEN 2.5VL

UNIVERSAL RECOGNITION

What are these attractions? Please give their names in Chinese and English.

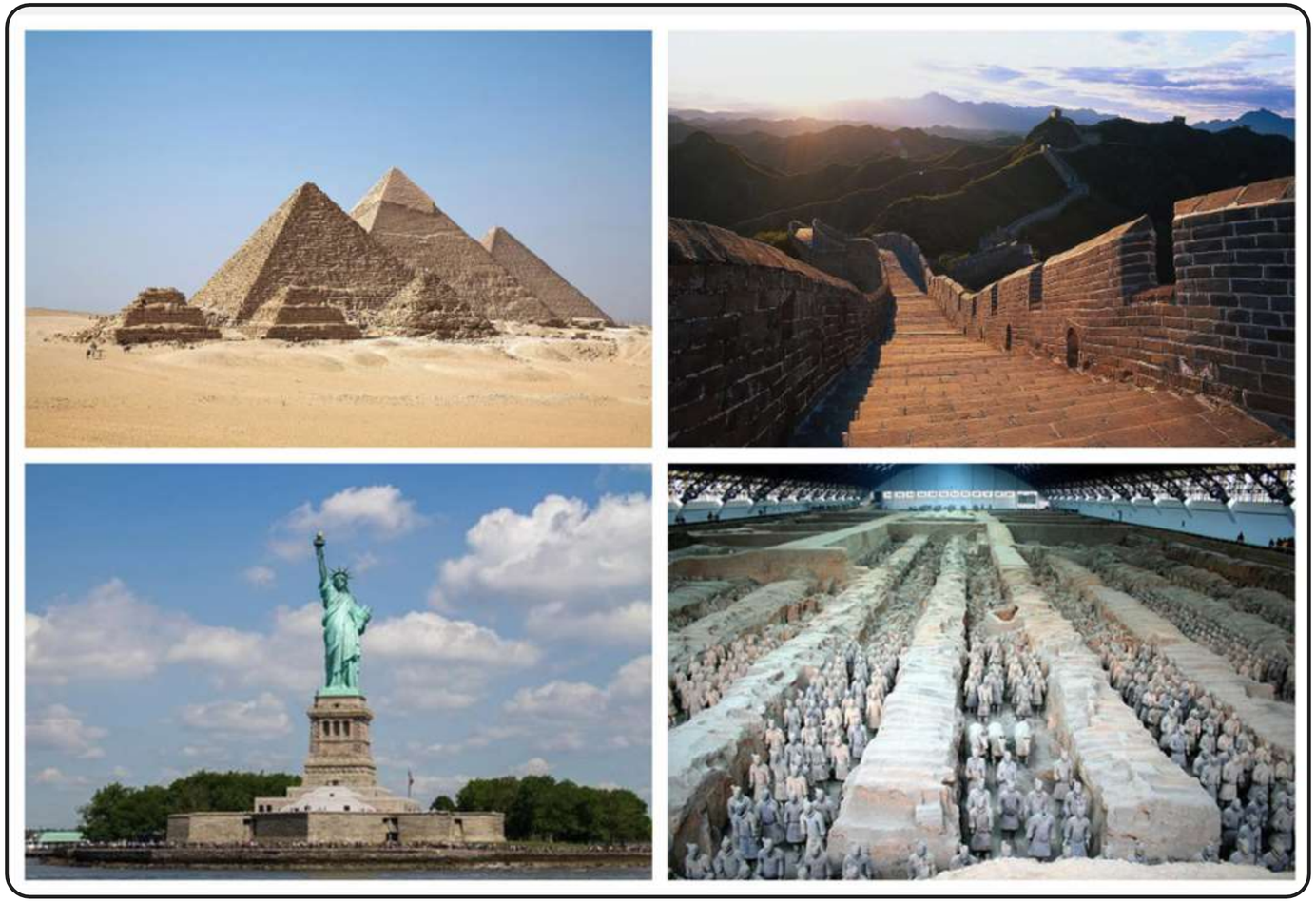
Qwen2.5-VL
These attractions are:

Top left: **The Great Pyramids of Giza**
(埃及的吉萨大金字塔)

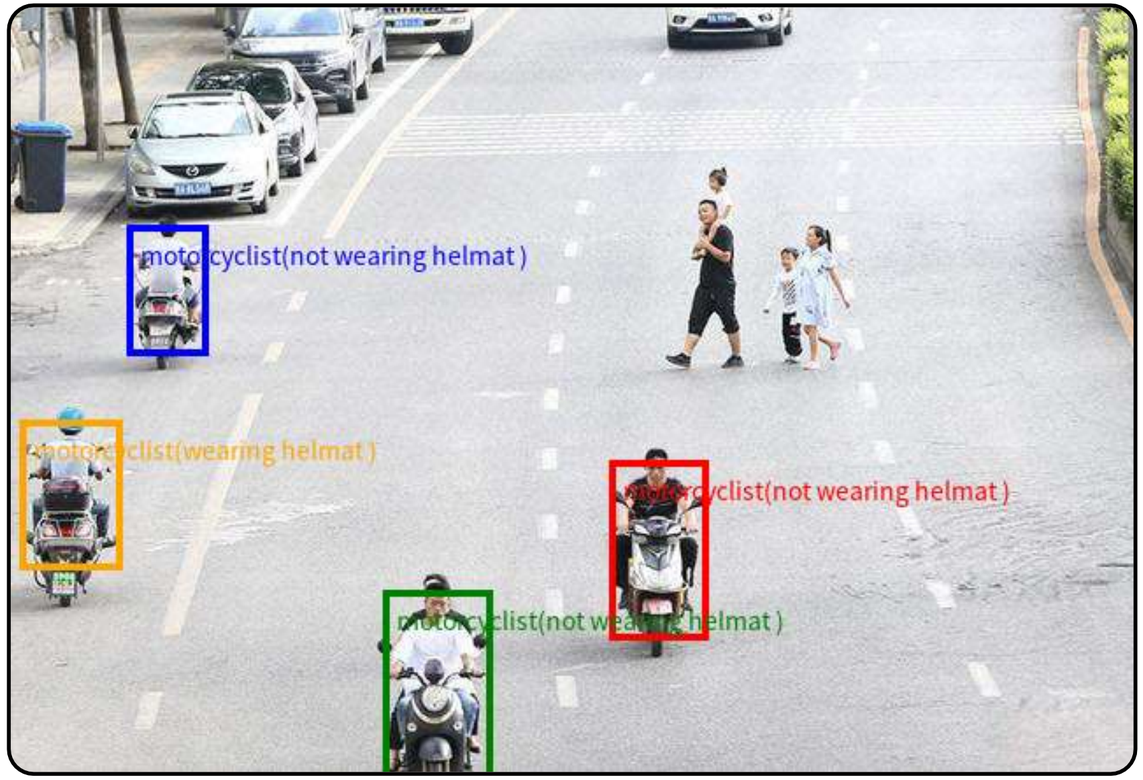
Top right: **The Great Wall of China**
(中国的长城)

Bottom left: **The Statue of Liberty**
(美国的自由女神像)

Bottom right: **The Terracotta Army**
(中国的秦始皇兵马俑)



OBJECT DETECTION



Detect all motorcyclists in the image and return their locations in the form of coordinates.

The format of output should be like {“bbox_2d”: [x1, y1, x2, y2], “label”: “motorcyclist”, “sub_label”: “wearing helmet” # or “not wearing helmet”}.

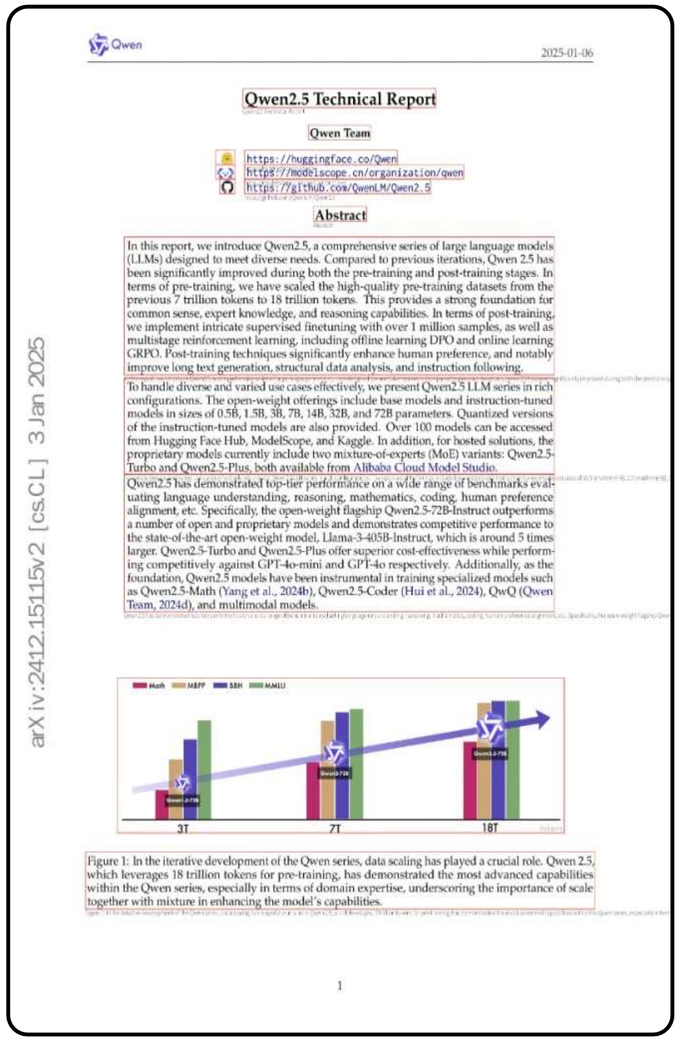

```
[
    {"bbox_2d": [341. 258. 397. 360], "label": "motorcyclist", "sub_label": "not wearing helmat "},
    {"bbox_2d": [212. 332. 274. 448], "label": "motorcyclist", "sub_label": "not wearing helmat "},
    {"bbox_2d": [66. 124. 112. 198]. "label": "motorcyclist", "sub_label": "not wearing helmat "},
    {"bbox_2d": [5. 235. 63. 320], "label": "motorcyclist", "sub_label": "wearing helmat "}
]
```

OBJECT DETECTION

Locate the person who act bravely, report the bbox coordinates in JSON format.

```
[
        {"bbox_2d": [74, 58, 526, 619], "label": "person who act bravely"}
]
```

DOCUMENT PARSING



Qwen 2025-01-06

Qwen2.5 Technical Report

Qwen Team

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen2.5

Abstract

In this report, we introduce Qwen2.5, a comprehensive series of large language models (LLMs) designed to meet diverse needs. Compared to previous iterations, Qwen 2.5 has been significantly improved during both the pre-training and post-training stages. In terms of pre-training, we have scaled the high-quality pre-training datasets from the previous 7 trillion tokens to 18 trillion tokens. This provides a strong foundation for common sense, expert knowledge, and reasoning capabilities. In terms of post-training, we implement intricate supervised finetuning with over 1 million samples, as well as multistage reinforcement learning, including offline learning DPO and online learning GRPO. Post-training techniques significantly enhance human preference, and notably improve long text generation, structural data analysis, and instruction following.

To handle diverse and varied use cases effectively, we present Qwen2.5 LLM series in rich configurations. The open-weight offerings include base models and instruction-tuned models in sizes of 0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B parameters. Quantized versions of the instruction-tuned models are also provided. Over 100 models can be accessed from Hugging Face Hub, ModelScope, and Kaggle. In addition, for hosted solutions, the proprietary models currently include two mixture-of-experts (MoE) variants: Qwen2.5-Turbo and Qwen2.5-Plus, both available from Alibaba Cloud Model Studio.

Qwen2.5 has demonstrated top-tier performance on a wide range of benchmarks evaluating language understanding, reasoning, mathematics, coding, human preference alignment, etc. Specifically, the open-weight flagship Qwen2.5-72B-Instruct outperforms a number of open and proprietary models and demonstrates competitive performance to the state-of-the-art open-weight model, Llama-3-405B-Instruct, which is around 5 times larger. Qwen2.5-Turbo and Qwen2.5-Plus offer superior cost-effectiveness while performing competitively against GPT-4o-mini and GPT-4o respectively. Additionally, as the foundation, Qwen2.5 models have been instrumental in training specialized models such as Qwen2.5-Math (Yang et al., 2024b), Qwen2.5-Coder (Hui et al., 2024), QwQ (Qwen Team, 2024d), and multimodal models.

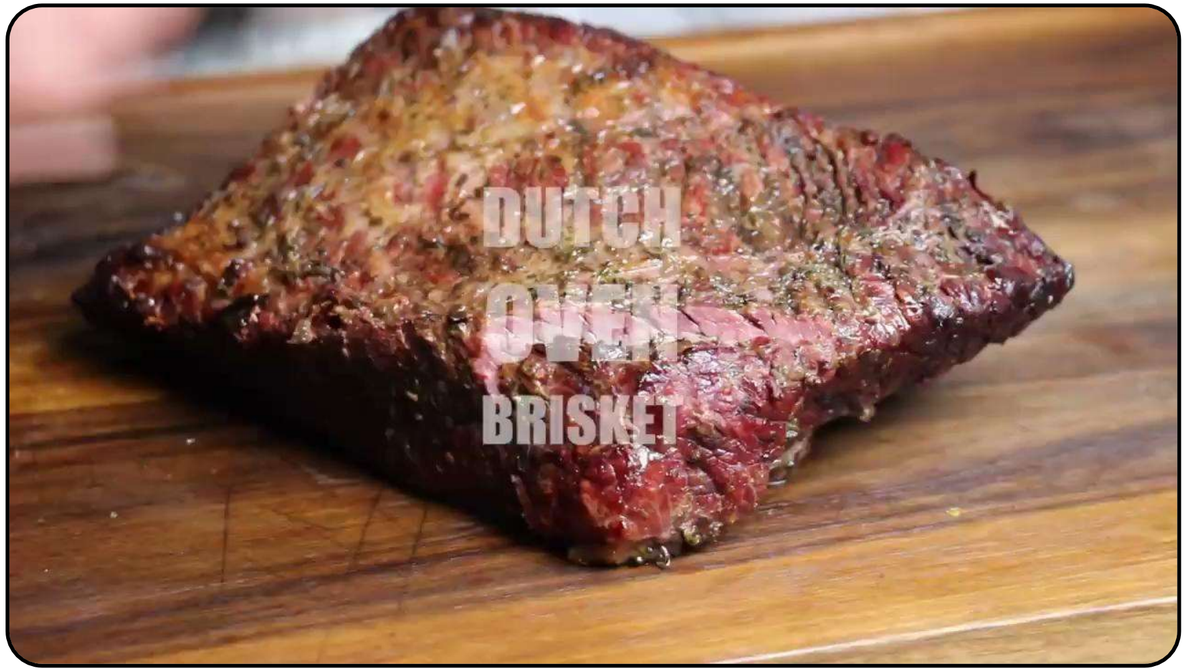
Figure 1: In the iterative development of the Qwen series, data scaling has played a crucial role. Qwen 2.5, which leverages 18 trillion tokens for pre-training, has demonstrated the most advanced capabilities within the Qwen series, especially in terms of domain expertise, underscoring the importance of scale together with mixture in enhancing the model's capabilities.

1

```
<html><body><h1 data-bbox="879 283 1605 348">Qwen2.5 Technical Report</h1><p data-bbox="1124 406 1360 456"></p>

<address class="author"><p>Qwen Team</p></address>

<div class="image" data-bbox="778 491 850 541"><img data-bbox="778 491 850 541"/></div>

<p data-bbox="885 491 1460 541">https://huggingface.co/Qwen</p>

<div class="image" data-bbox="778 541 850 589"><img data-bbox="778 541 850 589"/></div>

<p data-bbox="885 541 1708 589">https://modelscope.cn/organization/qwen</p>

<div class="image" data-bbox="792 589 850 639"><img data-bbox="792 589 850 639"/></div>

<p data-bbox="885 589 1584 639">https://github.com/QwenLM/Qwen2.5</p>

<h2 data-bbox="1143 681 1344 733">Abstract</h2>
```

.

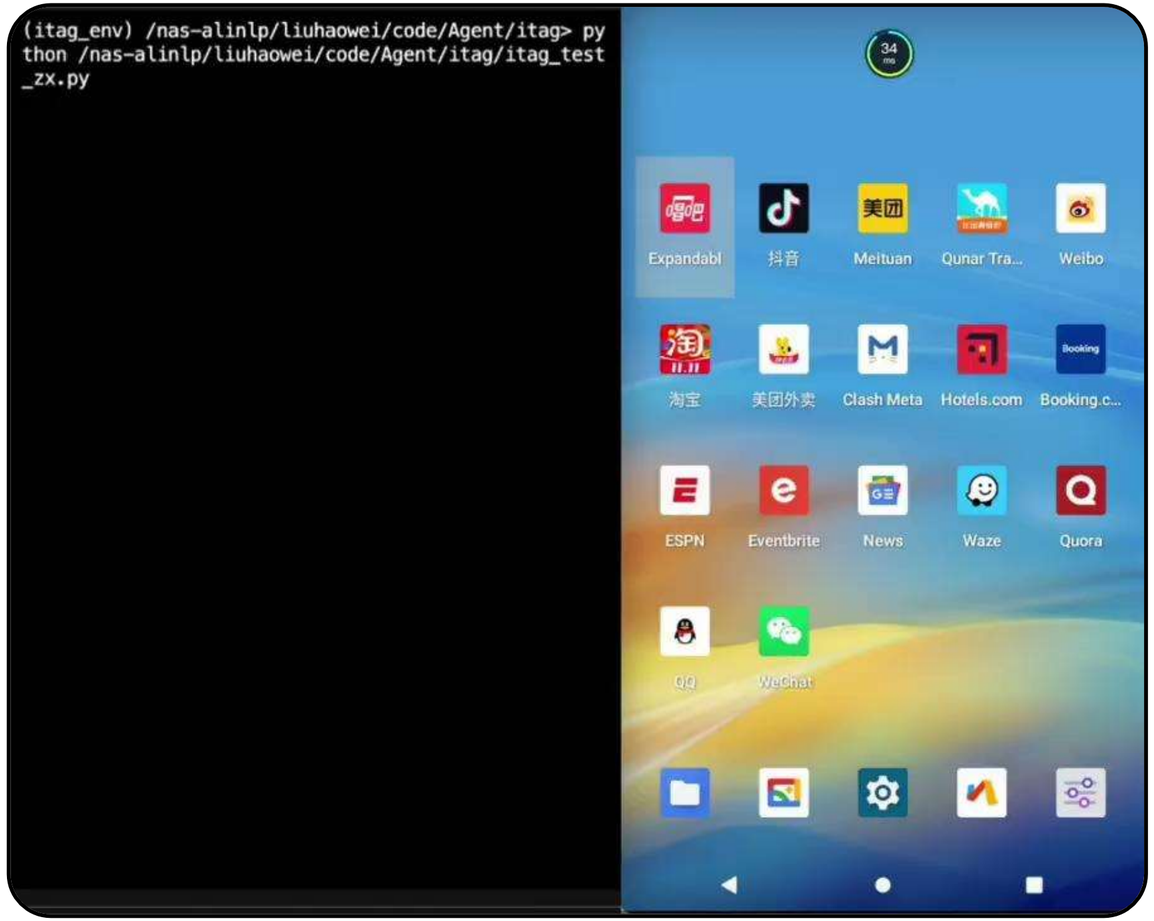
LONG VIDEO ANALYSIS



Localize a series of activity events in the video, output the start and end timestamp for each event, and describe each event with sentences. Provide the result in JSON.


```
[
 {
      "start_time": "00:21.00",
      "end_time": "00:49.00",
      "description": "A person removes a piece of meat from its packaging and cuts off the fat."

  },
  {
      "start_time": "00:50.00",
      "end_time": "01:27.00",
      "description": "The person seasons the meat with salt and pepper on both sides."
  },

  {
      "start_time": "01:28.00",
      "end_time": "01:36.00",
      "description": "The seasoned meat is placed on a grill."

]
```

AGENT ACTIONS



Please help me book a one-way ticket in the bookings app. The starting point is Chongqing Jiangbei Airport and the ending point is Beijing Capital Airport on January 28th

Usage

Qwen 2.5VL Chat Interface

```
chat_response = client.chat.completions.create(
    model="Qwen/Qwen2.5-VL-7B-Instruct",
    messages=[
        {"role": "system",
         "content": "You are a helpful assistant."
        },
        { "role": "user",
          "content": [
              {
                 "type": "image_url",
                 "image_url": {"url": "https://example.com/image.png"
              },
              },
              {
                                    "type": "text",
                                    "text": "What is the text in the image?"
                              }
          ],
        },
    ],
)
```

QWEN 2.5VL: ROLE

SYSTEM: Sets global behaviour

```
{
  "role": "system",
  "content": "You are an assistant that answers briefly, returns currency values in USD, "
          + "and uses JSON when a user asks for structured data."
}
```

QWEN 2.5VL: ROLE

USER: Sends an image and text

```
{
  "role": "user",
  "content": [
    { "type": "image",
      "image": "https://example.com/receipt-coffee.jpg" },
    { "type": "text",
      "text": "How much did the latte cost? Please give the answer as JSON." }
  ]
}
```

MODEL CARD

Name	Qwen/Qwen2.5-VL-3B-Instruct
Summary	Multi-modal LLM capable of image, video and tool-augmented chat.
Parameters	3 billion
Release Date	April 2025
Developer	Alibaba Cloud (Qwen team)
License	Qwen license (research/non-commercial); 3 B & 72 B sizes are not Apache 2.0.

WILL IT FIT?

Model size	Precision	VRAM needed*	Consumer cards that clear it
Qwen 2.5-VL-3B	8-bit / BF16	6 – 8 GB	RTX 3060 12 GB, RTX 4060 8–16 GB, laptop 4070
Qwen 2.5-VL-7B	8-bit	12 – 14 GB	RTX 4070 12 GB (tight), RTX 4070 Ti 16 GB
	BF16	18 – 20 GB	RTX 3090 / 4090 24 GB
Qwen 2.5-VL-32B / 72B	Any	≥ 32 GB (32B) / ≥ 64 GB (72B) or multi-GPU	Prosumer/enterprise (RTX 6000 Ada 48 GB, dual 4090s, etc.)

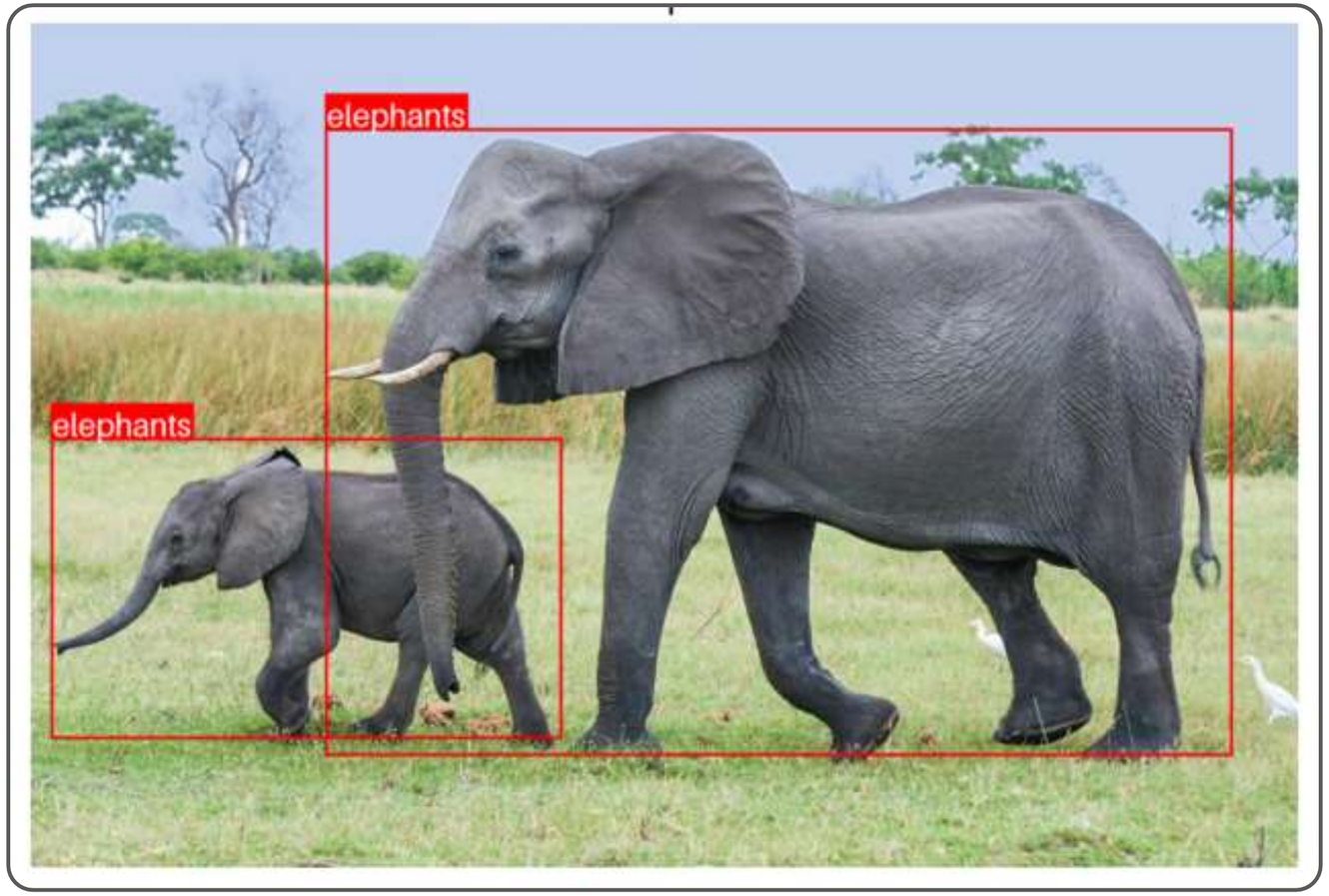
IMAGE CAPTIONING

INPUT

OUTPUT

The image depicts a serene and picturesque scene of a white dog sitting on a stone pathway near a stunning lake. The lake has crystal-clear turquoise water that reflects the surrounding landscape, including the mountains in the background. The mountains are covered with lush greenery, and the sky above is partly cloudy, adding to the tranquil atmosphere

OBJECT DETECTION & SPATIAL REASONING APP



Thank You