Visual Correspondence



Q: Which point on the second image corresponds to the point in the first image?

(A) Point A (B) Point B (C) Point C (D) Point D

Llama-3.2-11B-Vision-Instruct: Point D.

Llama-3.2-11B-Vision-Instruct (Ours trained) + VTS-V

 t_1 : I will use depth estimation to understand the spatial relationships and

identify the correct point

a₁: depth action {"image index":0} o_1 : image-1

Verifier : Continue.

 t_2 : I will now check the depth map for the second image to find the corresponding.

a2: depth_action {"image_index":1}

 o_2 : image-2



differ due to the change in camera perspective, they represent the same feature point in 3D space.

Result: B.

Counting



Q: How many skiers are wearing red? Select from the following choices. (A) 2 (B) 3 (C) 0 (D) 1



Qwen2.5-VL-7B-Instruct :



Owen2.5-VL-7B-Instruct (Ours trained) + VTS-V

t₁: To determine how many skiers are wearing red, I will first locate the skier in the image and then check if they are wearing red.

a₁: grounding_action{"image_index":0,"text":["skier"]} o_1 : image-1

😂 Verifier : Continue.

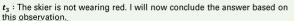
t2: I will zoom into the skier's bounding box to check if they are wearing red.

a2: zoomin_action{"image_index":0, "bounding_box:{"x_ min":372,"y min":228,"x max":418,"y max":284}}



Verifier: OK.

Result: C.



Fine-grained Recognition



Q: What's the color of traffic light?



Qwen2-VL-7B-Instruct :



Qwen2-VL-7B-Instruct (Ours trained) + VTS-V

t1: To determine the color of the traffic light. I need to identify and analyze the traffic light in the image. a1: grounding action{"image index":0,"text":["traffic light"]

 o_1 : image-1



 t_2 : I need to zoom into the identified bounding box to determine the color of the traffic light.

a₂: zoomin action{"image index":0."bounding box":{"x m in":871,"v min":397,"x max":880,"v max":407}} o_2 : image-2



Result: Green, I

t₃: The zoomed-in image clearly shows the color of the traffic light.

Forensic Detection

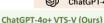


Q: You are given the two images. Please examine the details and tell which one of them is most likely to be a real photograph (A) the first image (B) the second image



 t_1 : To determine which image is most likely a real photograph, I will

ChatGPT-4o: A X



use the image to texts similarity action to compare each image with the description 'real photograph'.

a₁: image to texts similarity action{"image index":0,"texts":["real photograph"]}

o₁: Similarity scores: [0.1792]

Verifier : Continue.

 t_2 : I will now check the second image for comparison. a₂: image to texts similarity action{"image index":1,"texts":["real

photograph"]} o₂: Similarity scores: [0.1836]

Verifier : OK.

t₃: The second image has the highest similarity score of 0.1836, indicating it is most likely to be a real photograph.

Result: B.

