Report 3

Project 3: Object Detection and Human–Object Interaction Analysis

TJ DiMeola
Course: Computer Vision CSCI 581
Instructor: Dr. Hawk Wang
Date: November 16, 2025

Project Objective

The goal of this project is to deepen understanding of object detection pipelines and extend that knowledge to higher-level visual reasoning.

Part 1: Lightweight Object Detection

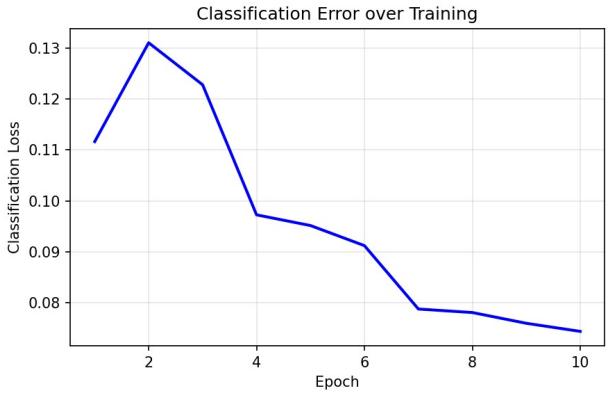
Implement and train a lightweight detection model. Here Faster R-CNN with MobileNet is applied to two banana detection datasets:

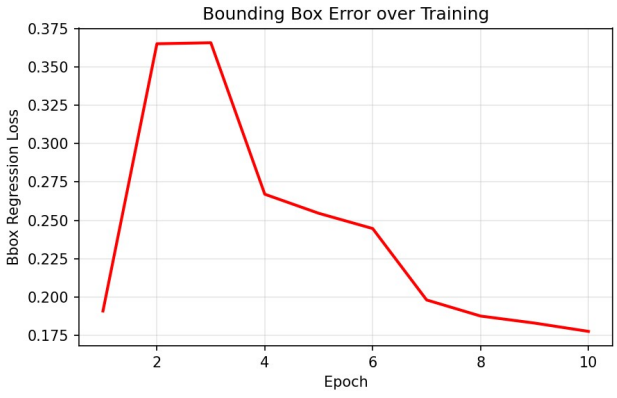
- banana-detection (d2l)
- banana_classification (Kaggle)

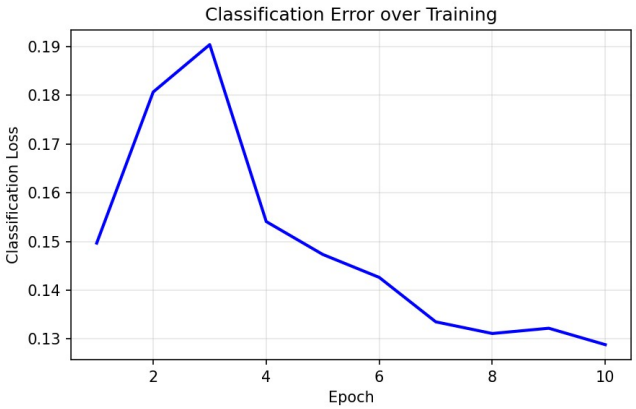
> Since the Kaggle banana dataset (https://www.kaggle.com/datasets/atrithakar/banana-classification) had no labels (bounding box data in csv form), I used qwen to generate bounding boxes for the data and then calculated the bbox values to match the csv format of the d2l dataset.
>
> This VLM-as-annotator approach generated 921 accurate labels allowing me to train on a much larger mixed-dataset leading to greatly improved in-the-wild image detection (from 1/7 to 6/7 success rate). This effectively demonstrates that VLMs can serve as effective zero-shot data labelers for training traditional detectors."

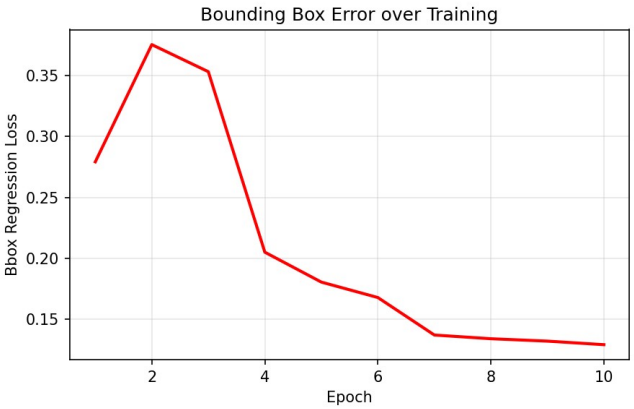
Tasks and Deliverables:

- Plot training loss curves (class error and bounding box error). Below there are two TL curves. The first based on banana-prediction (d2l) alone, the second based on both the d2l and the Kaggle datasets.



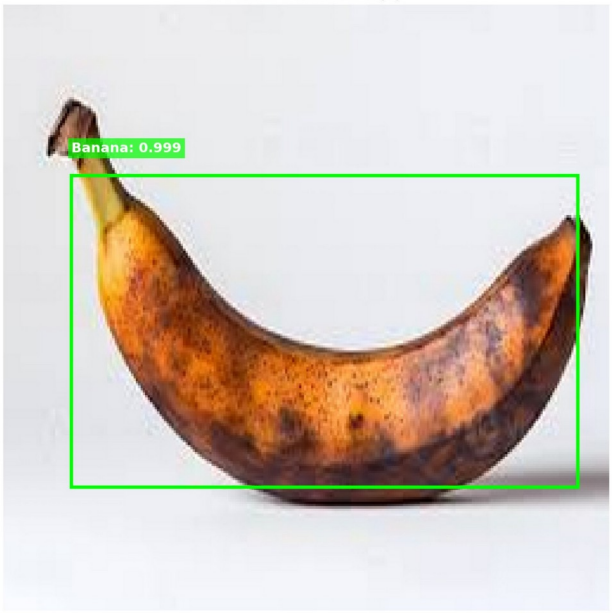






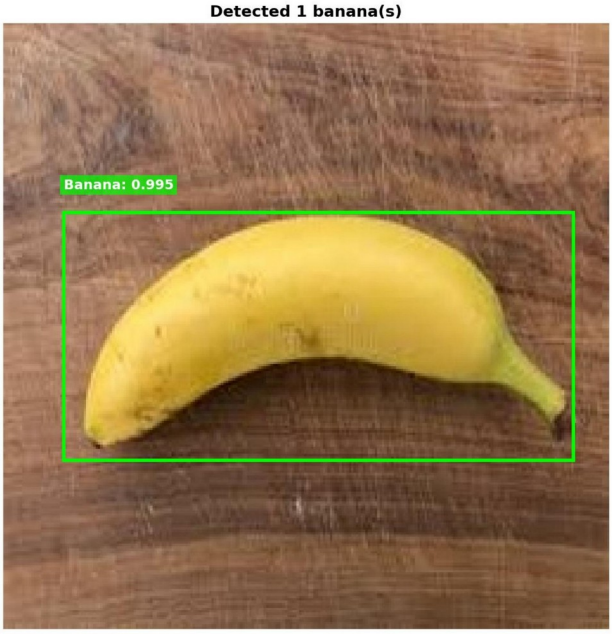
- Show 5 sample detections: See next page.
- Test the the detection model with random (in-the-wild) banana images
  - **Were there any failure cases?** The d2l val set had no failure cases, on the Kaggle set, failed in all cases. The model generated by mixed training (d2l + Kaggle) failed on only the last image (the banana on the grass and gravel).
  - **Discuss the possible reasons that caused the detection to fail:** The d2l case failed on all “real” bananas because it was really just a single banana. The transforms were poor (they could at least elongated it, reducing the arc in several instances). The mixed case only failed on the image #7 because the background was simply to complex for Faster R-CNN to really parse.

Detected 1 banana(s)

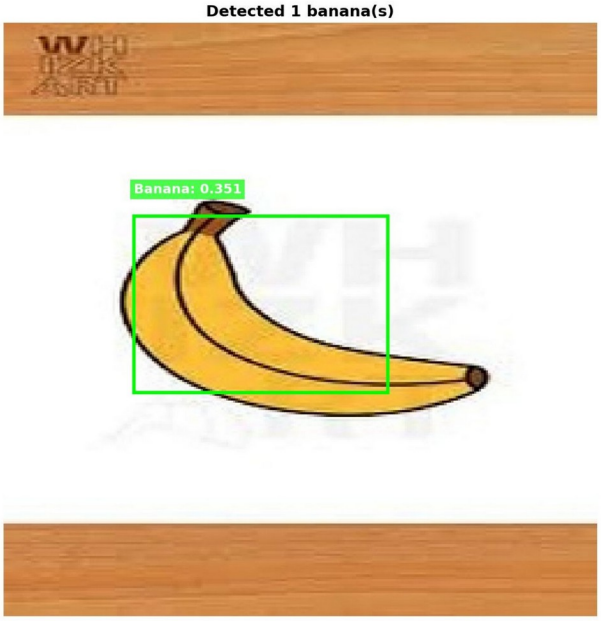


Detected 1 banana(s)

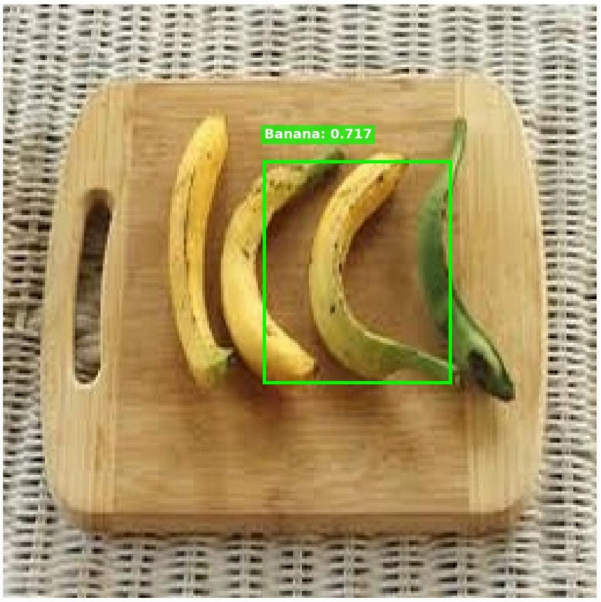
Detected 1 banana(s)



Detected 1 banana(s)



Detected 1 banana(s)

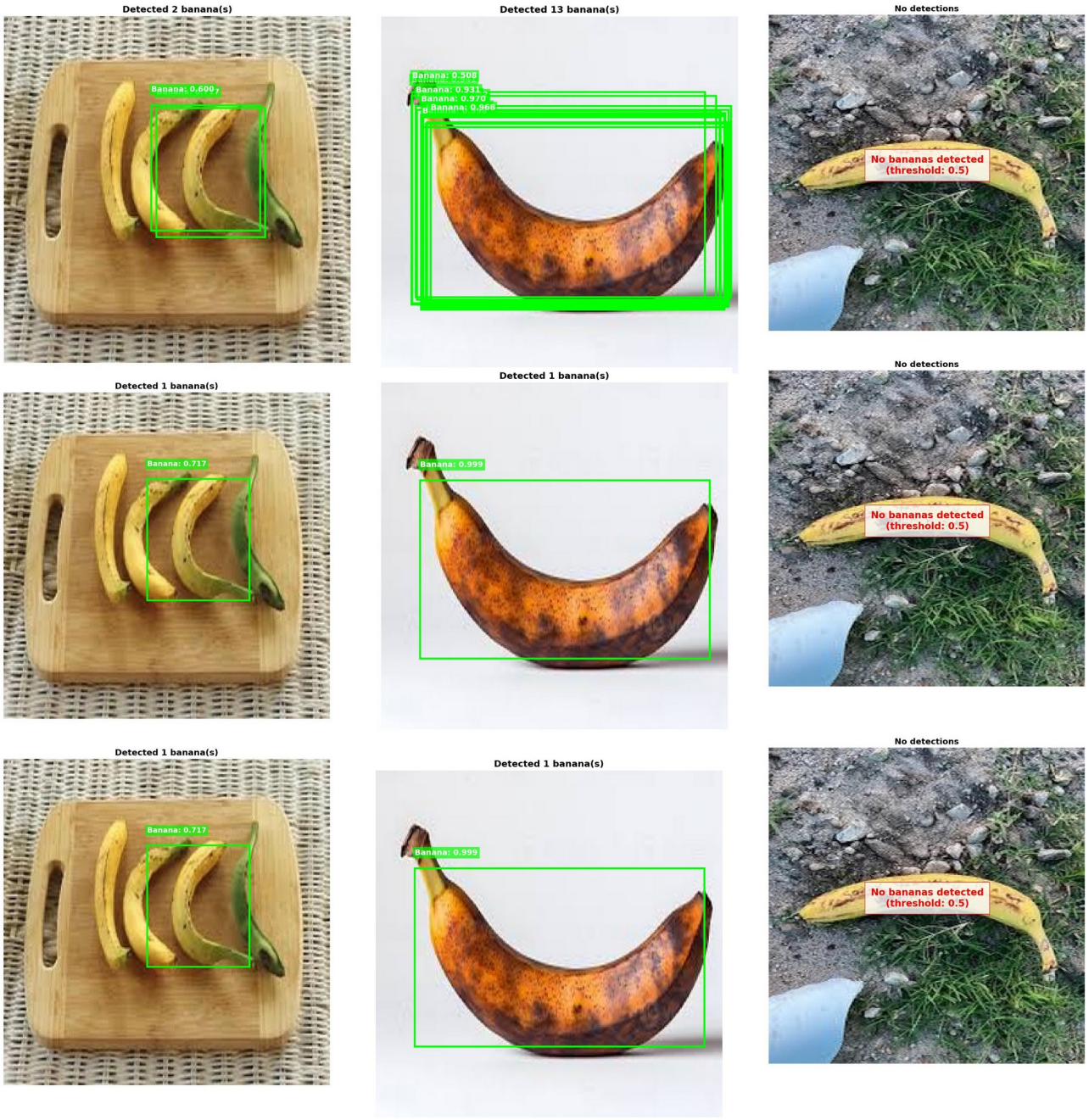


No detections

Part 2: Non-Maximum Suppression (NMS)

Tasks and Deliverables

- Visualize outputs before/after NMS



- **Compare with PyTorch's NMS implementation. Any difference?** I did not find any differences when these two implementations were run on the same test (zero-shot) images.
- **Discuss its purpose and limitations:** NMS eliminates redundant overlapping bounding boxes by keeping only the highest-confidence detection and suppressing nearby boxes above a specified IoU threshold. This prevent preventing multiple detections of the same object (see the first line of the image above, where no nms was provided). Such fixed IoU thresholds fail in crowded scenes because potentially valid detections can be suppressed when objects genuinely overlap. NMS is a greedy algorithm that can't adapt to varying

object densities and may discard correct boxes if a slightly higher-confidence but incorrect box overlaps them.

Part 3: Human–Object Interaction (HOI) Analysis using VLMs

Perform zero-shot HOI analysis using VLMs on a subset of HICO-DET dataset.

Tasks and Deliverables

- Use one (or more) open-source or closed-source VLMs to predict human–object interactions. Here I used Claude Sonnet 4.5 and Qwen VL Max.

```
Analyze this image and identify all human-object interactions.

For each interaction, provide the answer in this exact format:
<verb object>

Where:
- verb = the action being performed (e.g., riding, eating, holding,
sitting on)
- object = the object being interacted with (e.g., bicycle, pizza,
umbrella, chair)

Rules:
1. Only report interactions you can CLEARLY see in the image
2. Use simple, concrete verbs (not complex phrases)
3. One interaction per line
4. List only the interactions, nothing else

Example output format:
riding bicycle
holding umbrella
eating pizza

Now analyze this image:
```

- Come up with your prompt to guide the VLMs to predict <interaction object>. For example, <hold apple>, <ride bicycle>. Prompt used:
- Identify a few failure cases where VLMs fail to prediction the HOI classes for the given images? If so, discuss the possible reasons.

1. The HICO-DET ground truth is poorly done

- Doesn't capture all HOIs
- Sometimes just identifies and object rather than an interaction
- has odd underscores "_" between verb pairs that fool analytics into think VLM got the category wrong

1. Qwen does better on HOI metrics (where comparisons with gt can be trusted)

- +45% better precision (16.1% vs 11.1%)
- +31% better F1 (0.144 vs 0.110)
- 2 perfect scores vs Claude's 0

2. Claude seems to over-predict (where comparisons with gt can be trusted)

- Claude: 76 predictions (41% MORE than GT's 54)
- Qwen: 62 predictions (15% more than GT)
- 44% "Claude-only" predictions (46 interactions) vs Qwen's 27%
- Classic precision/recall tradeoff: Claude appears to be recall-focused (doesn’t miss anything!), Qwen is precision-focused (only say what its sure of)

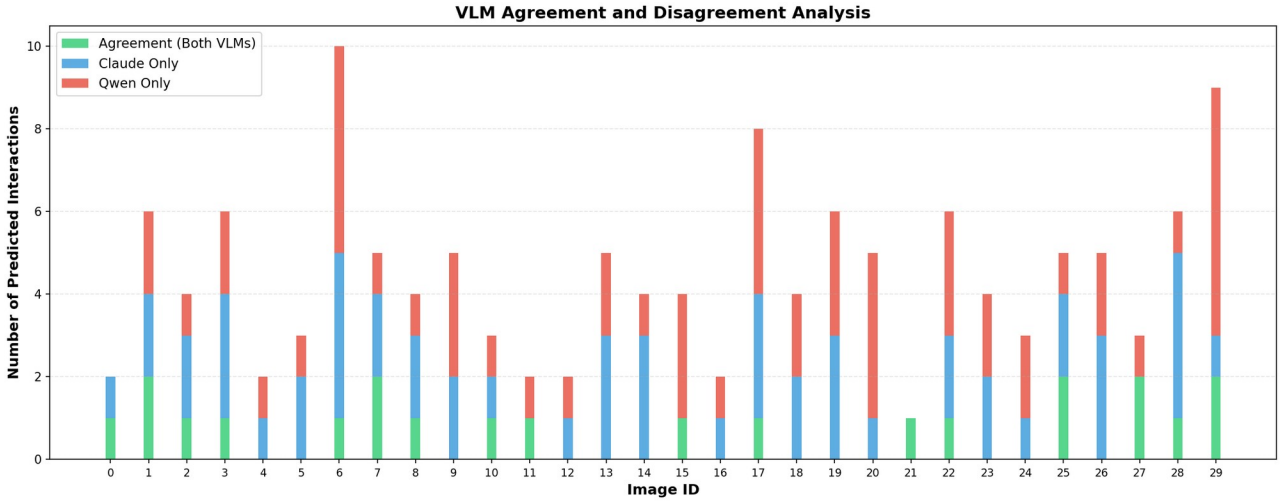
Failure Cases

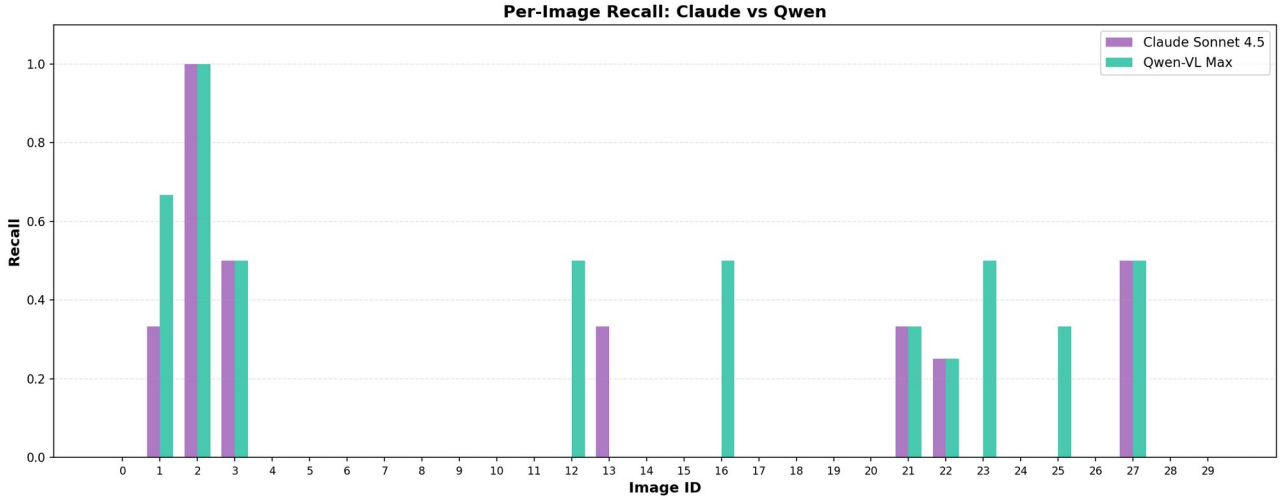
- The three tables on the following page show the groundtruth vs the responses of the 2 VLMs. One of the most frustrating things here is that the ground truth of this dataset is not always right. E.g., image hoi_result_021.png, where one of the gt options is “umbrella.” That is simply an object, not an HOI. Additionally, the gt is formatted in such a way as relationships are often parsed: “sit_on” with an underscore.

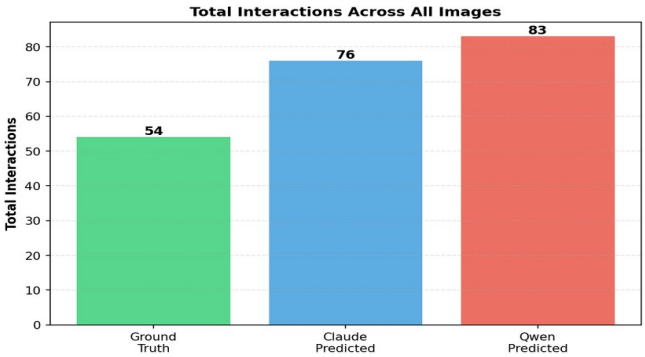
Failure Case Fixes

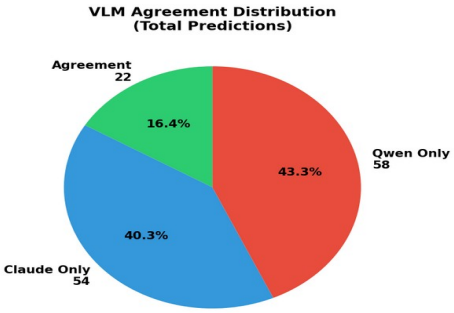
- The prompted was changed to better help the VLMs understand the *format* of the gt.
- But with the gt so bad, it made little sense to do much else. This assisted the metrics marginally.

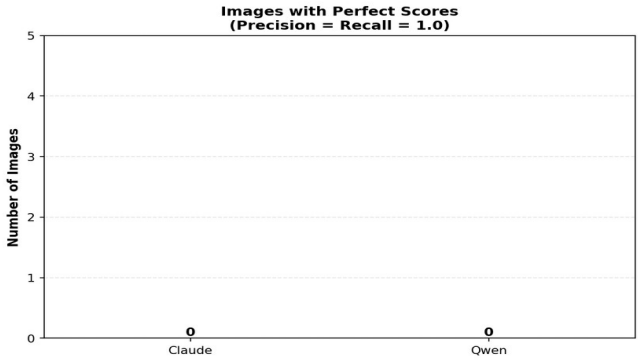
HOI Metrics and Evaluation Tables











Summary Statistics

Metric	Claude Sonnet 4.5	Qwen-VL Max
Avg Precision	0.100	0.192
Avg Recall	0.108	0.169
F1 Score	0.104	0.180
Perfect Scores	0	0
Total Predictions	76	83

Failure Analysis: Images 0-9

Image	Ground Truth	Claude Predicted	Qwen Predicted	Both Right ✓✓	Both Wrong ✗✗
0	bench sit_on	sit bench, lean railing	sit bench	∅	bench sit_on
1	horse hold, horse ride, horse walk	ride horse, lead horse, hold reins, stand grass	ride horse, carry clipboard, hold reins, walk horse	ride horse	horse walk, horse hold, horse ride
2	boat ride	driving boat, ride boat, operate boat	ride boat, sit boat	ride boat	boat ride
3	motorcycle ride, motorcycle sit_on	race motorcycle, ride motorcycle, operate sidecar, driving vehicle	sit sidecar, ride motorcycle, hold handlebar	ride motorcycle	motorcycle ride, motorcycle sit_on
4	backpack carry, backpack wear	hold jar	hold cup	∅	backpack wear, backpack carry
5	bench lie_on	touch desk, lie desk	sleep bench	∅	bench lie_on
	dining_table eat_at	drink glass, hold pen,	place napkin, hold cup,		dining_table eat_at

Failure Analysis: Images 10-19

Image	Ground Truth	Claude Predicted	Qwen Predicted	Both Right ✓✓	Both Wrong ✗✗
10	kite carry, kite hold	ride bicycle, hold decoration	ride bicycle, hold handlebar	∅	kite carry, kite hold
11	sports_ball carry, sports_ball hold, sports_ball throw	hold ball	throw ball, hold ball	∅	sports_ball throw, sports_ball carry, sports_ball hold
12	skateboard flip, skateboard jump	ride skateboard	jump skateboard	∅	skateboard flip, skateboard jump
13	elephant hold, elephant ride, elephant watch	stand water, ride elephant, watch elephants	sit elephant, hold stick	∅	elephant ride, elephant watch, elephant hold
14	potted_plant no_interaction	this image of what appears to be a flower market, i can identify the following human-	no human-object interactions are clearly visible in the image.	∅	potted_plant no_interaction
15	tie wear	hold bouquet	stand girl, hold bouquet, stand boy, sit ground	∅	tie wear
16	baseball_glove hold, baseball_glove wear	hold glove	hold baseball_glove	∅	baseball_glove wear, baseball_glove hold
17	train board, train ride	stand platform, boarding train, waiting train, photographing train	walk platform, wait train, hold bag, stand platform, hold coat	∅	train board, train ride
18	bus wash	hold pole, washing bus	hold mop, clean bus	∅	bus wash
19	bed lie_on	covered blanket, lie bed, resting head pillow	sleep bed, lie pillow, rest blanket	∅	bed lie_on

Failure Analysis: Images 20-29

Image	Ground Truth	Claude Predicted	Qwen Predicted	Both Right ✓✓	Both Wrong ✗✗
20	person teach	hold object	laugh woman, hold thread, talk woman, hold needle	∅	person teach
21	umbrella carry, umbrella hold, umbrella stand_under	hold umbrella	hold umbrella	hold umbrella	umbrella carry, umbrella hold, umbrella stand_under
22	kite fly, kite hold, kite launch, kite pull	watch kite, stand grass, hold kite	sit grass, walk grass, stand stroller, hold kite	hold kite	kite pull, kite launch, kite fly, kite hold
23	baseball_bat hold, baseball_bat wield	stand base, hold bat	hold baseball_bat, stand home_plate	∅	baseball_bat wield, baseball_bat hold
24	giraffe pet	touch giraffe	feed giraffe, hold leaf	∅	giraffe pet
25	boat ride, boat row, boat sit_on	hold sail, sit boat, rowing boat, hold oar	row boat, sit boat, hold oar	∅	boat row, boat sit_on, boat ride
26	motorcycle no_interaction	stand near motorcycle, view motorcycle, displaying motorcycle	stand ground, hold motorcycle	∅	motorcycle no_interaction
27	horse ride, horse straddle	ride horse, sit saddle	ride horse, hold reins, sit saddle	ride horse	horse ride, horse straddle
28	laptop hold, laptop read	working table, using computer, sit chair, using laptop, typing laptop	use laptop, sit chair	∅	laptop read, laptop hold
29	dining_table eat_at, dining_table sit_at	hold glass, sit table, sit chair	hold spoon, hold plate, eat food, hold knife, drink wine, hold fork, sit chair, hold glass	∅	dining_table eat_at, dining_table sit_at