Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMs

CVPR ‘24

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, Tsung-Yu Lin
Stony Brook University, Meta AI Research

Presented by: ***Kanchana Ranasinghe*** **- 04.02.2024**

Teaser

- Visual LLM
    - Process image + text
    - Generate text

Teaser

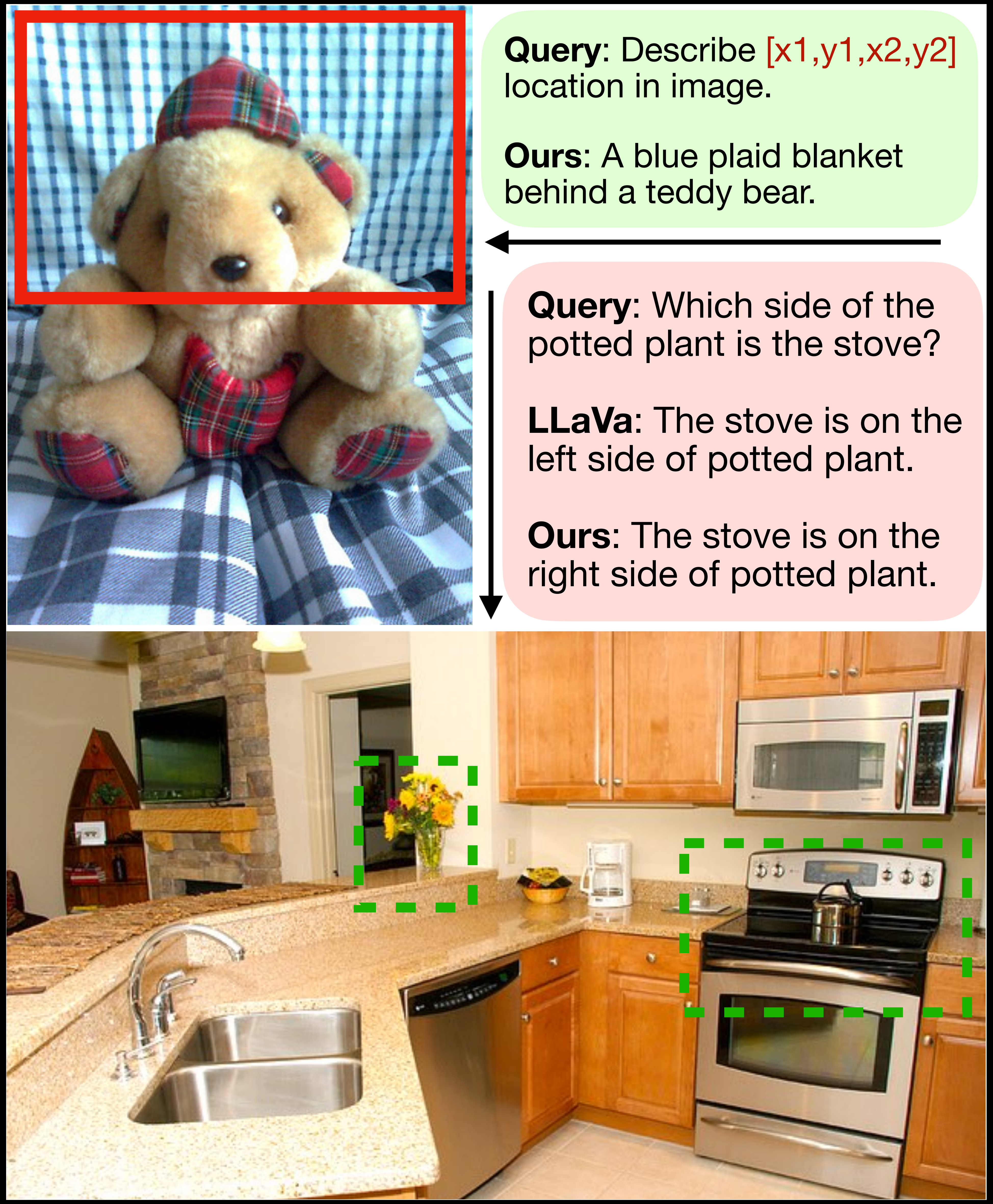
- Visual LLM
    - Process image + text
    - Generate text

- Location specific QnA
    - Process coordinates as text
    - Generate coordinates as text
    - Improves spatial awareness of QA

Teaser

- Visual LLM
  - Process image + text
  - Generate text

- Location specific QnA
  - Process coordinates as text
  - Generate coordinates as text
  - Improves spatial awareness of QA



Query: Describe [x1,y1,x2,y2] location in image.

Ours: A blue plaid blanket behind a teddy bear.

Query: Which side of the potted plant is the stove?

LLaVa: The stove is on the left side of potted plant.

Ours: The stove is on the right side of potted plant.

Agenda

1. Background
2. Motivation
3. Methodology
    3.1. Coordinate representation
    3.2. Instruction Fine-Tuning
    3.3. Pseudo-Data
4. Findings
    4.1. Improved VQA
    4.2. Novel Skills
5. Discussion

1. Background

Visual Question Answering (VQA) with LLaVA architecture

- BLIP-2

Li, Junnan et al. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.” International Conference on Machine Learning (2023).
Liu, Haotian et al. “Visual Instruction Tuning.” NeurIPS 2023

1. Background

Visual Question Answering (VQA) with LLaVA architecture

- BLIP-2

- LLaVA
    - LLM + Visual Encoder
    - Pre-train Adapter MLP
    - Instruction Fine-Tune LLM

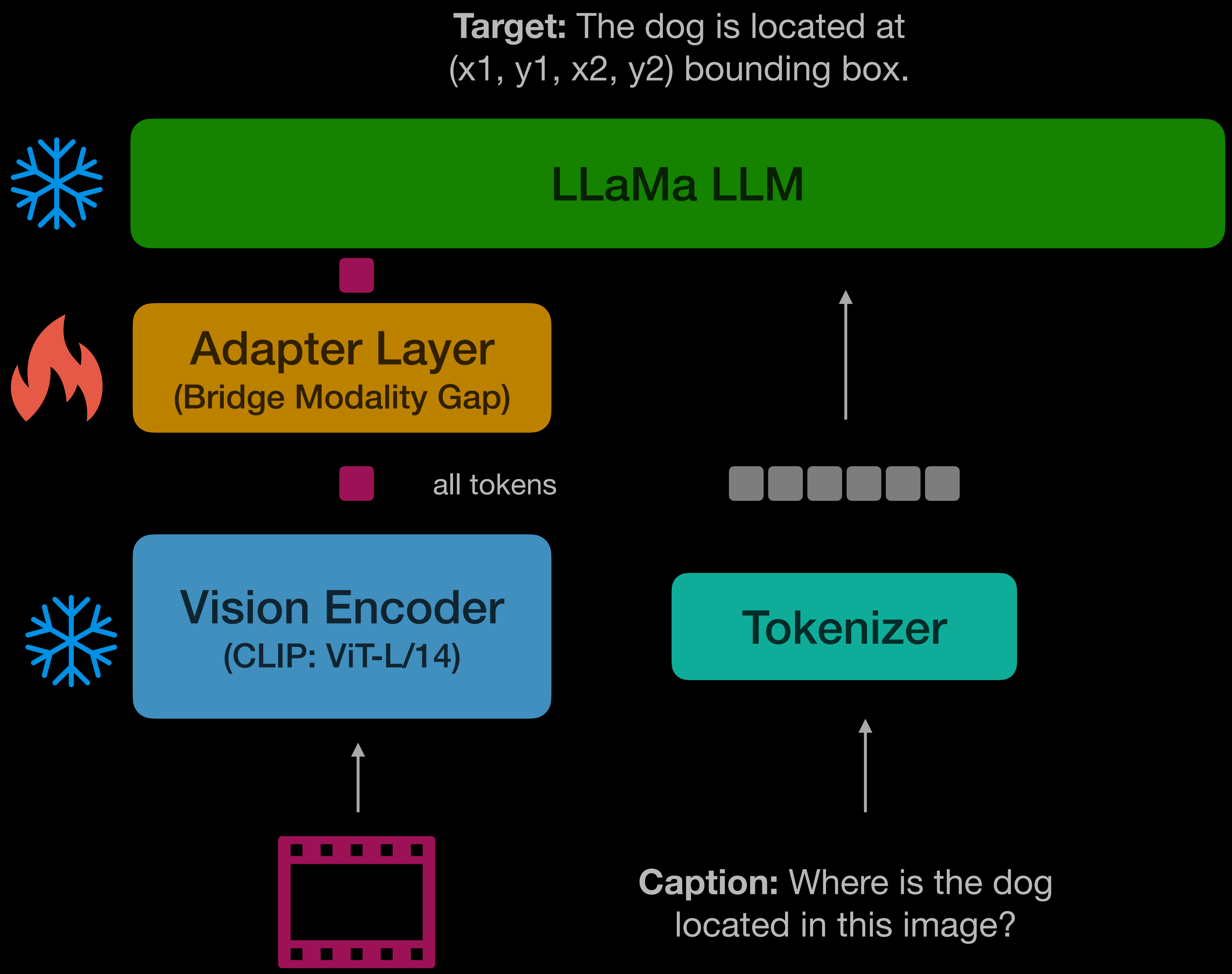
Li, Junnan et al. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.” International Conference on Machine Learning (2023).
Liu, Haotian et al. “Visual Instruction Tuning.” NeurIPS 2023

1. Background

Visual Question Answering (VQA) with LLaVA architecture

- BLIP-2
- LLaVA
  - LLM + Visual Encoder
  - **Pre-train Adapter MLP**
  - Instruction Fine-Tune LLM

595K image-text pairs


Target: The dog is located at (x1, y1, x2, y2) bounding box.
LLaMa LLM
Adapter Layer
(Bridge Modality Gap)
all tokens
Vision Encoder
(CLIP: ViT-L/14)
Tokenizer
Caption: Where is the dog located in this image?

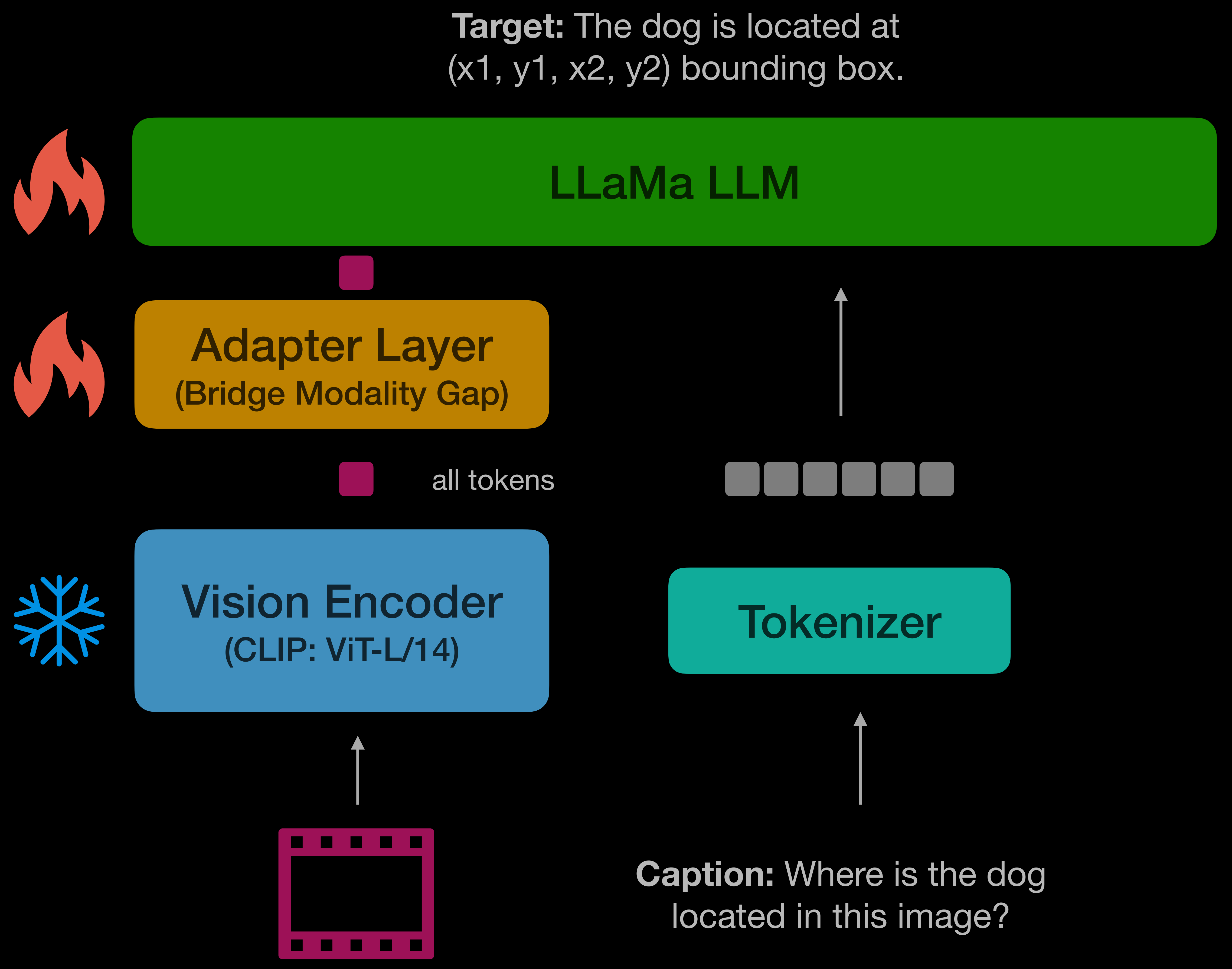
Li, Junnan et al. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.” International Conference on Machine Learning (2023).
Liu, Haotian et al. “Visual Instruction Tuning.” NeurIPS 2023

1. Background

Visual Question Answering (VQA) with LLaVA architecture

- BLIP-2
- LLaVA
  - LLM + Visual Encoder
  - Pre-train Adapter MLP
  - **Instruction Fine-Tune LLM**

100K image-conversation pairs

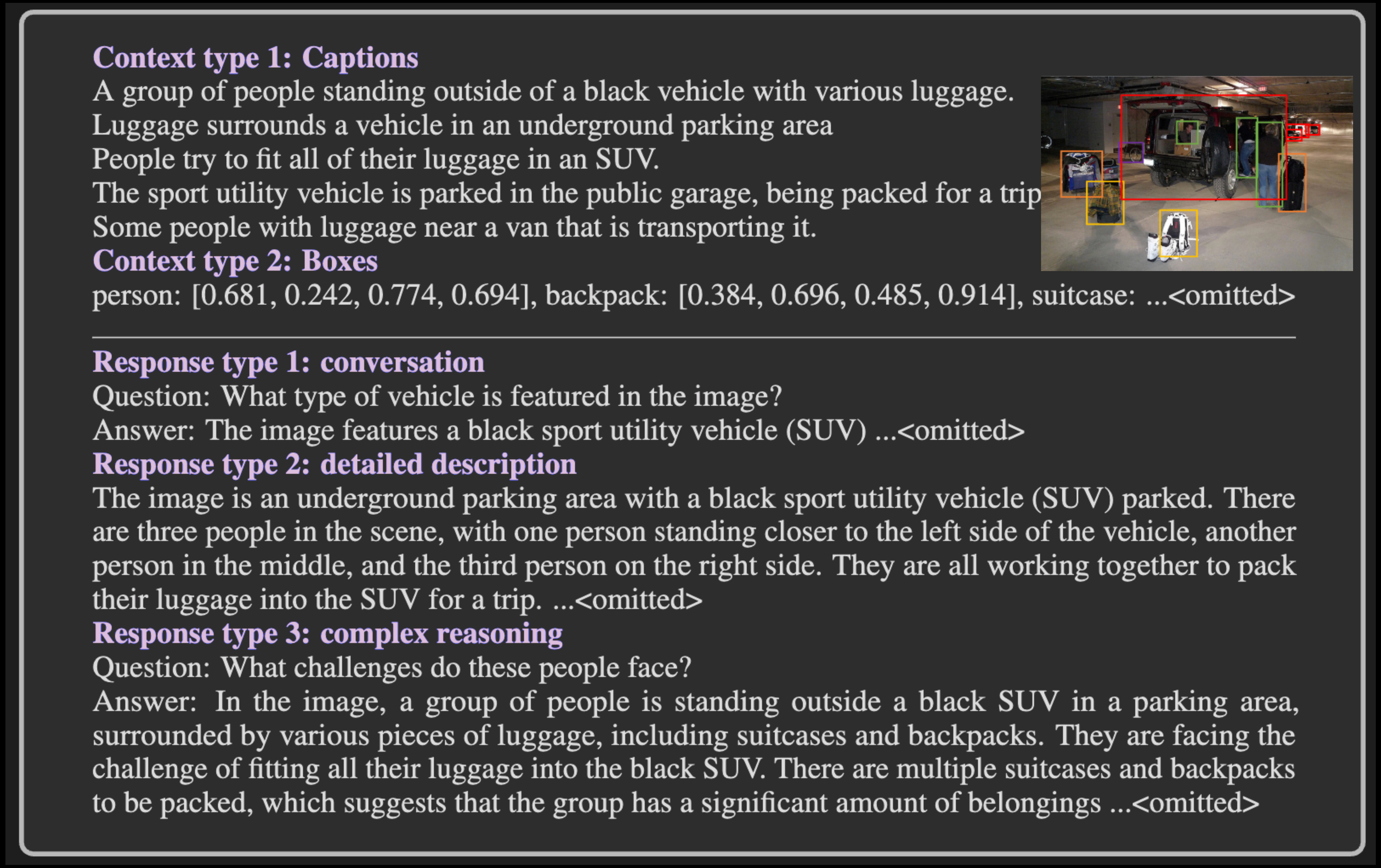


Li, Junnan et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." International Conference on Machine Learning (2023).
Liu, Haotian et al. "Visual Instruction Tuning." NeurIPS 2023

1. Background

Visual Question Answering (VQA) with LLaVA architecture

- BLIP-2
- LLaVA
    - LLM + Visual Encoder
    - Pre-train Adapter MLP
    - **Instruction Fine-Tune LLM**



Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

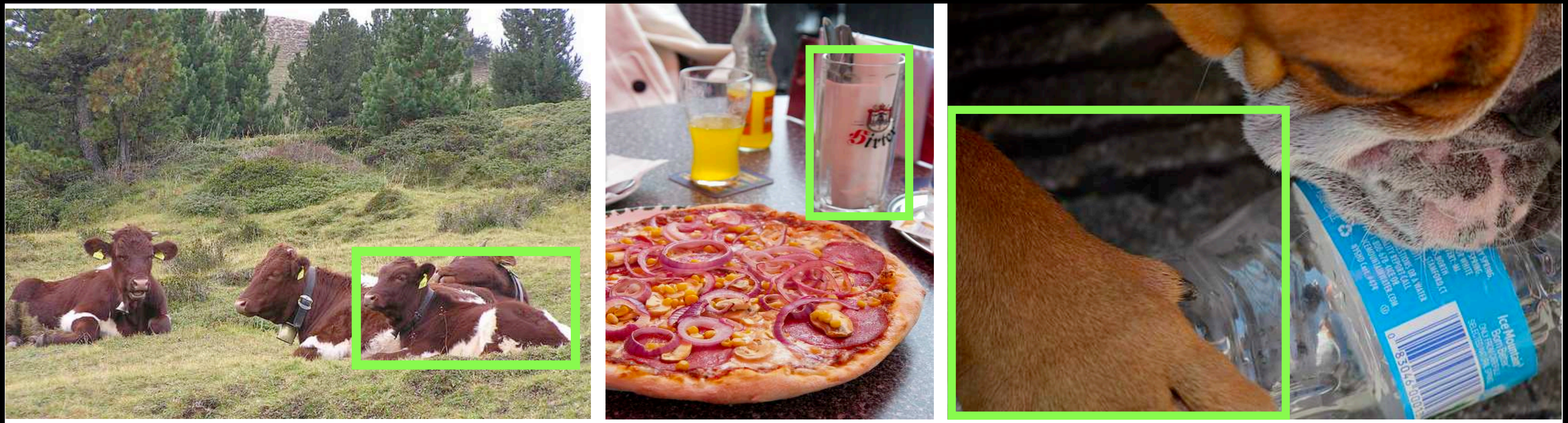
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Li, Junnan et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." International Conference on Machine Learning (2023).
Liu, Haotian et al. "Visual Instruction Tuning." NeurIPS 2023

2. Motivation

Revisit Proposed Task: Spatial Coordinates as Text for QnA

Prompt: Describe the region described by (x1,y1, x2, y2) bounding box.

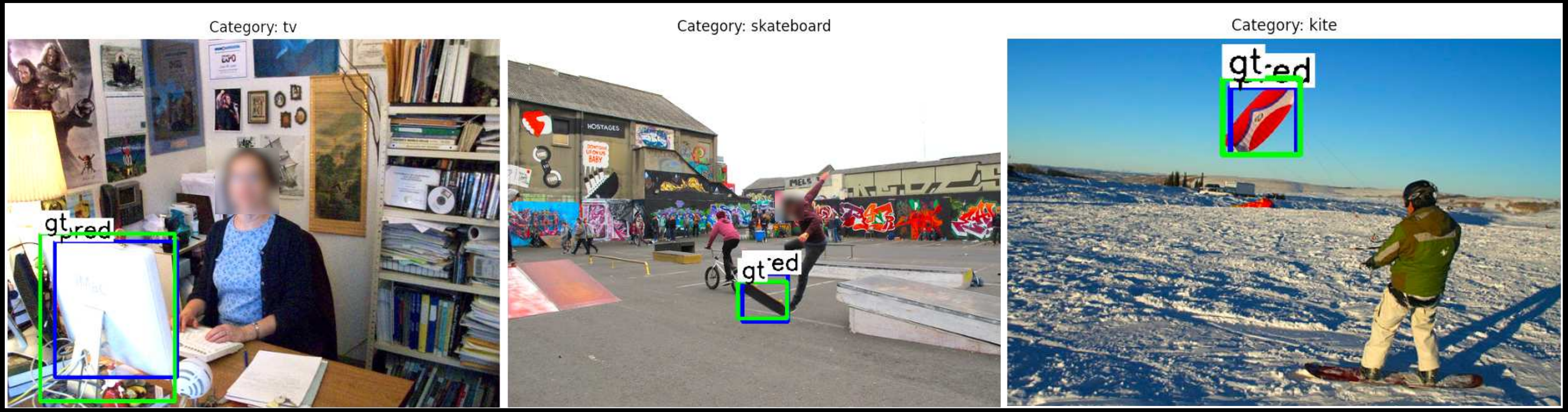


There is a cow that is lying down on a grassy hillside, surrounded by other cows and trees.

There is a cup that is a tall glass, placed on a table next to a pizza.

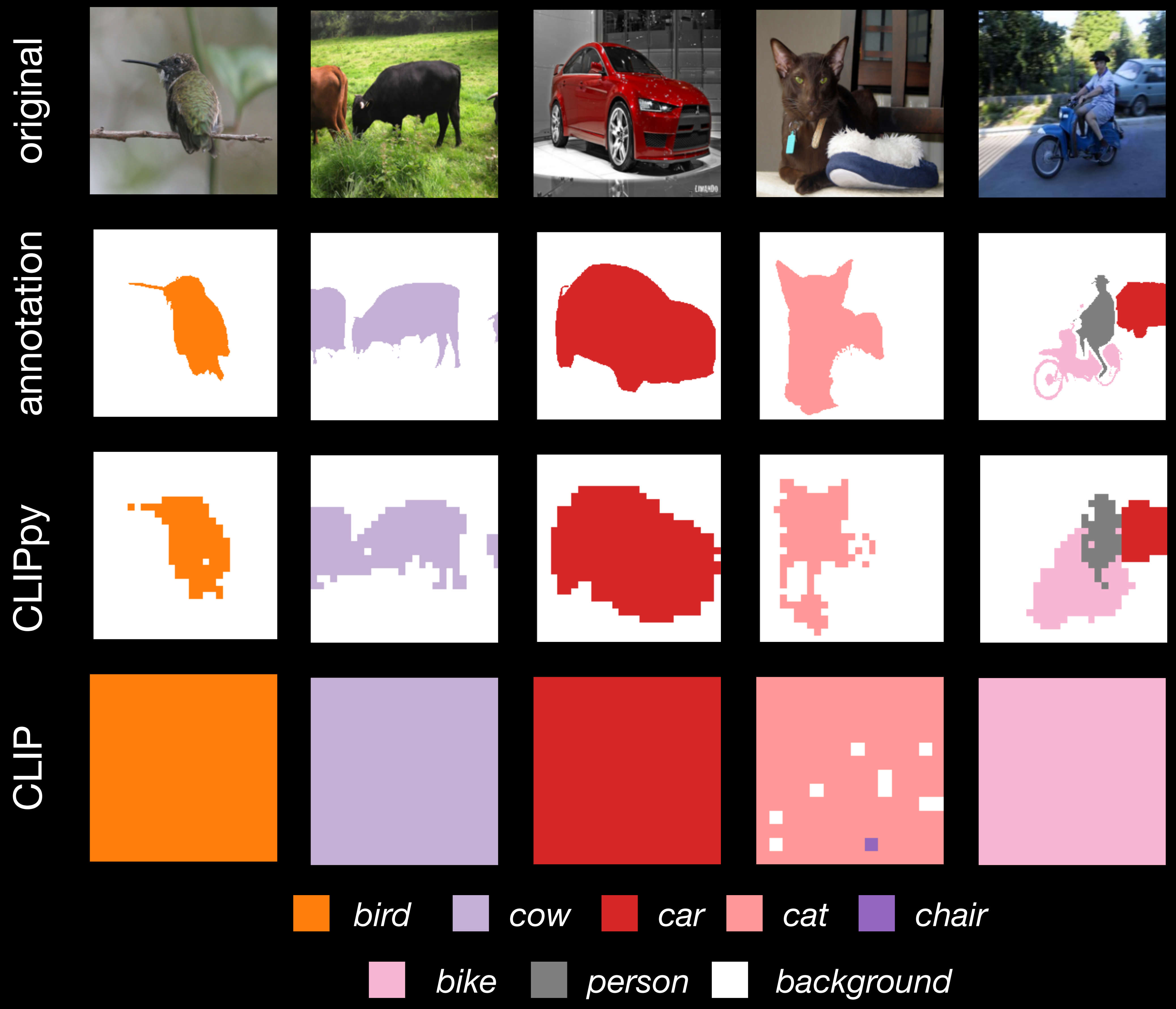
There is a dog that is a brown and white dog, and it is standing next to a bottle of water, possibly drinking from it.

Prompt: Where is the {category} object located in the image?



2. Motivation

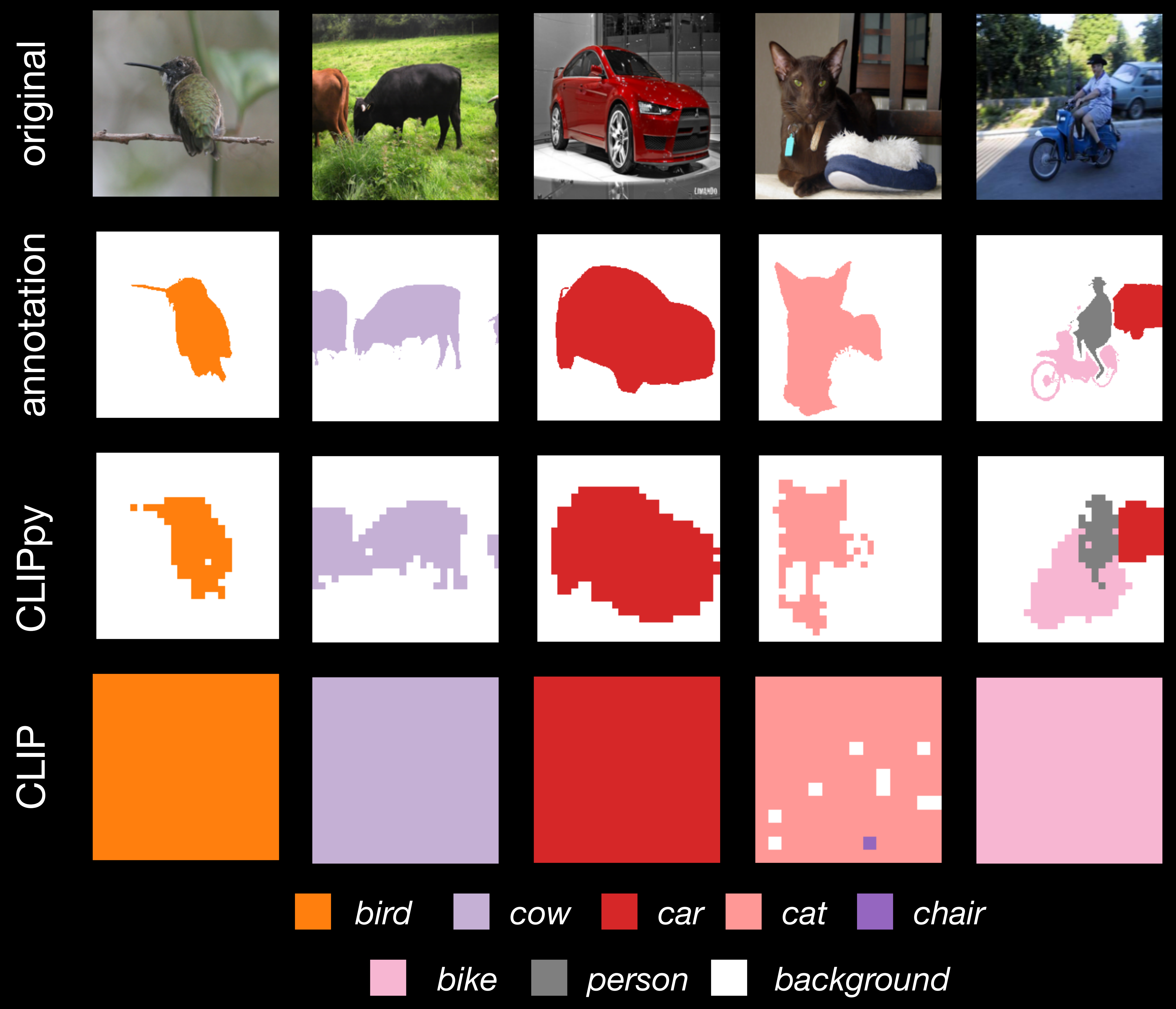
Prior Work: CLIPpy$_{\text{(ICCV '23)}}$ localization improves robustness

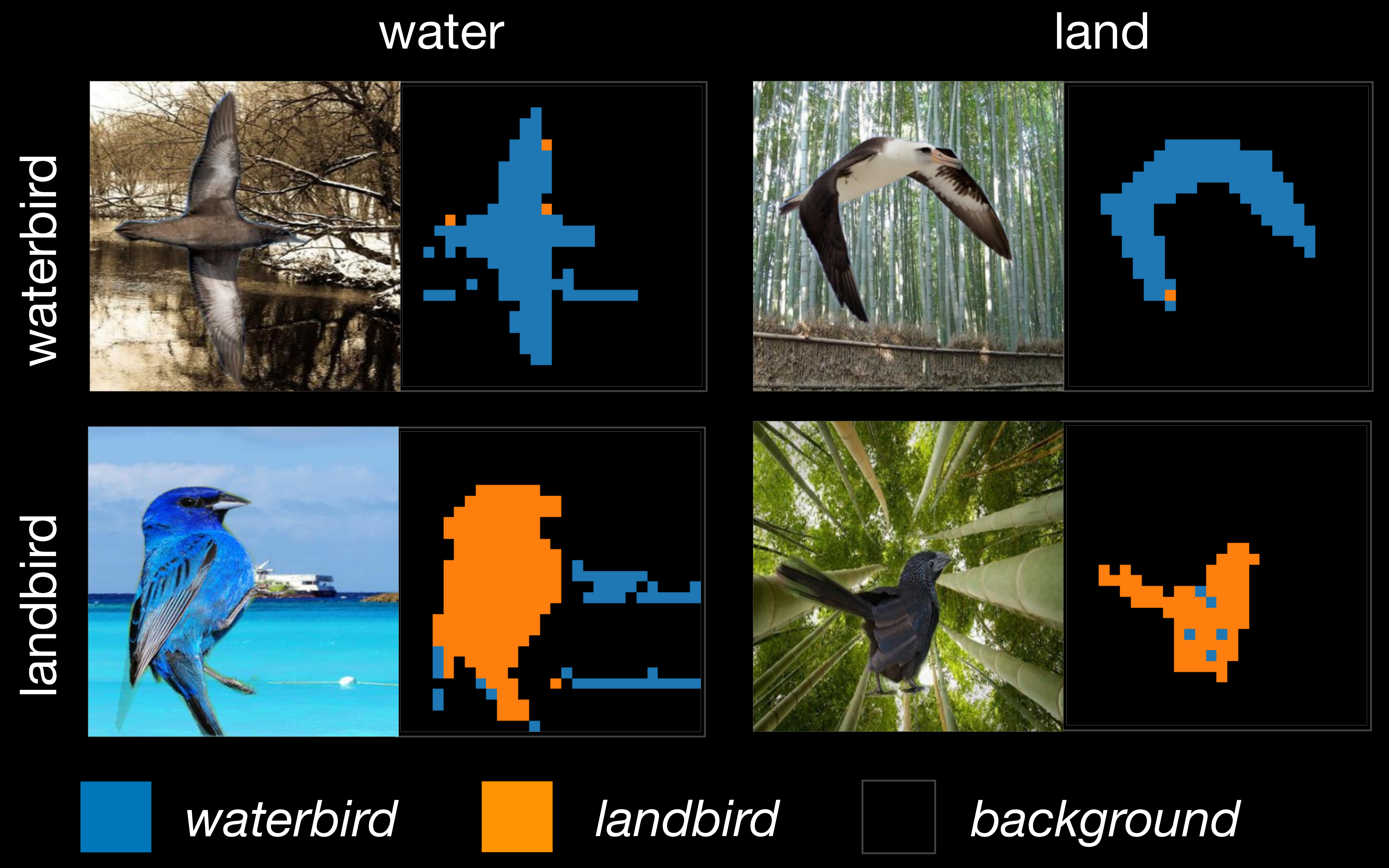


Ranasinghe, Kanchana et al. "Perceptual Grouping in Contrastive Vision-Language Models." 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2022): 5548-5561.

2. Motivation

Prior Work: CLIPpy$_{\text{(ICCV '23)}}$ localization improves robustness





CLIP	water	land	Δ
waterbird	80.2	48.1	**-32.1**
landbird	38.8	71.7	**-32.9**

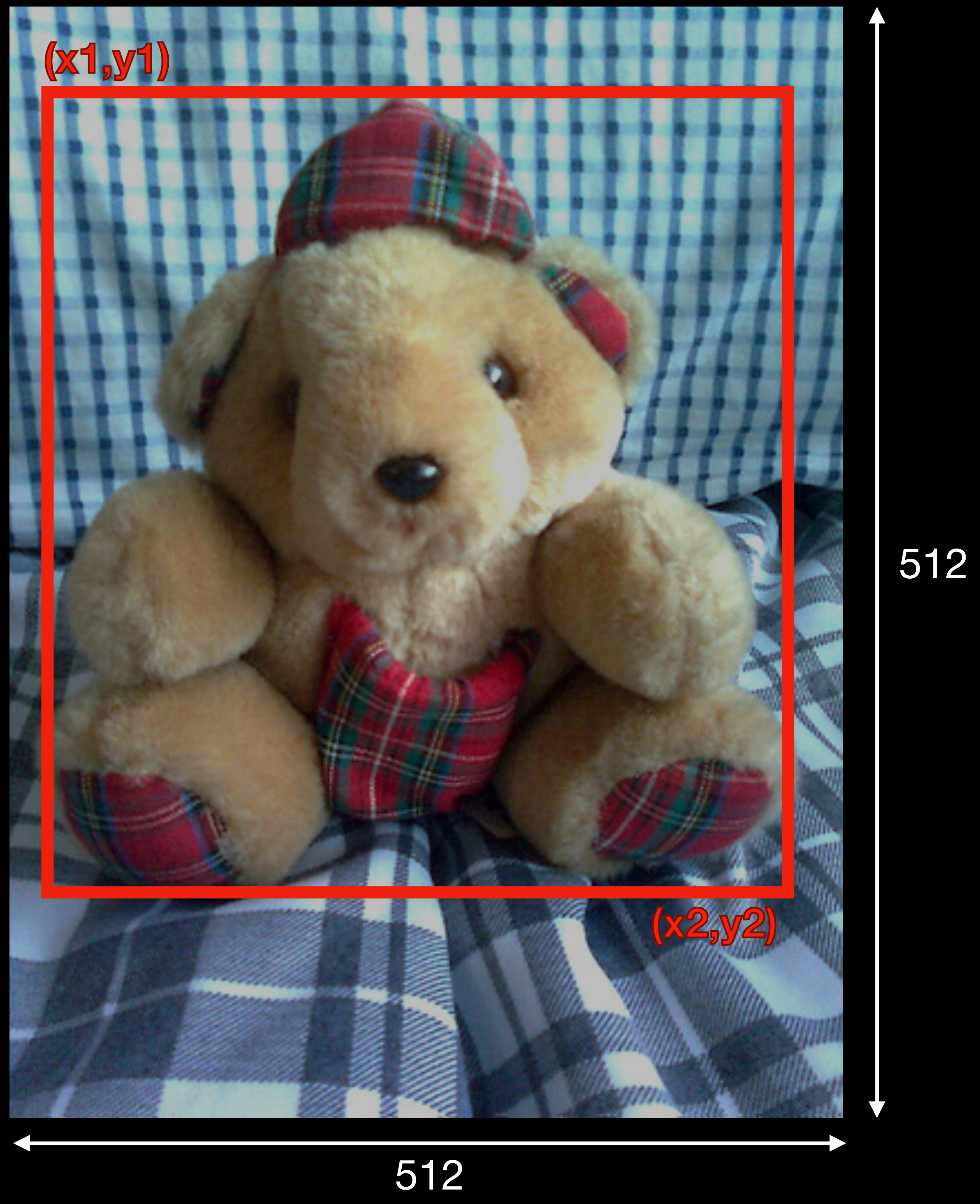
CLIPpy	water	land	Δ
waterbird	76.9	74.9	**-2.0**
landbird	80.0	84.1	**-4.1**

Ranasinghe, Kanchana et al. “Perceptual Grouping in Contrastive Vision-Language Models.” 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2022): 5548-5561.

3. Methodology

3.1. Coordinate representation

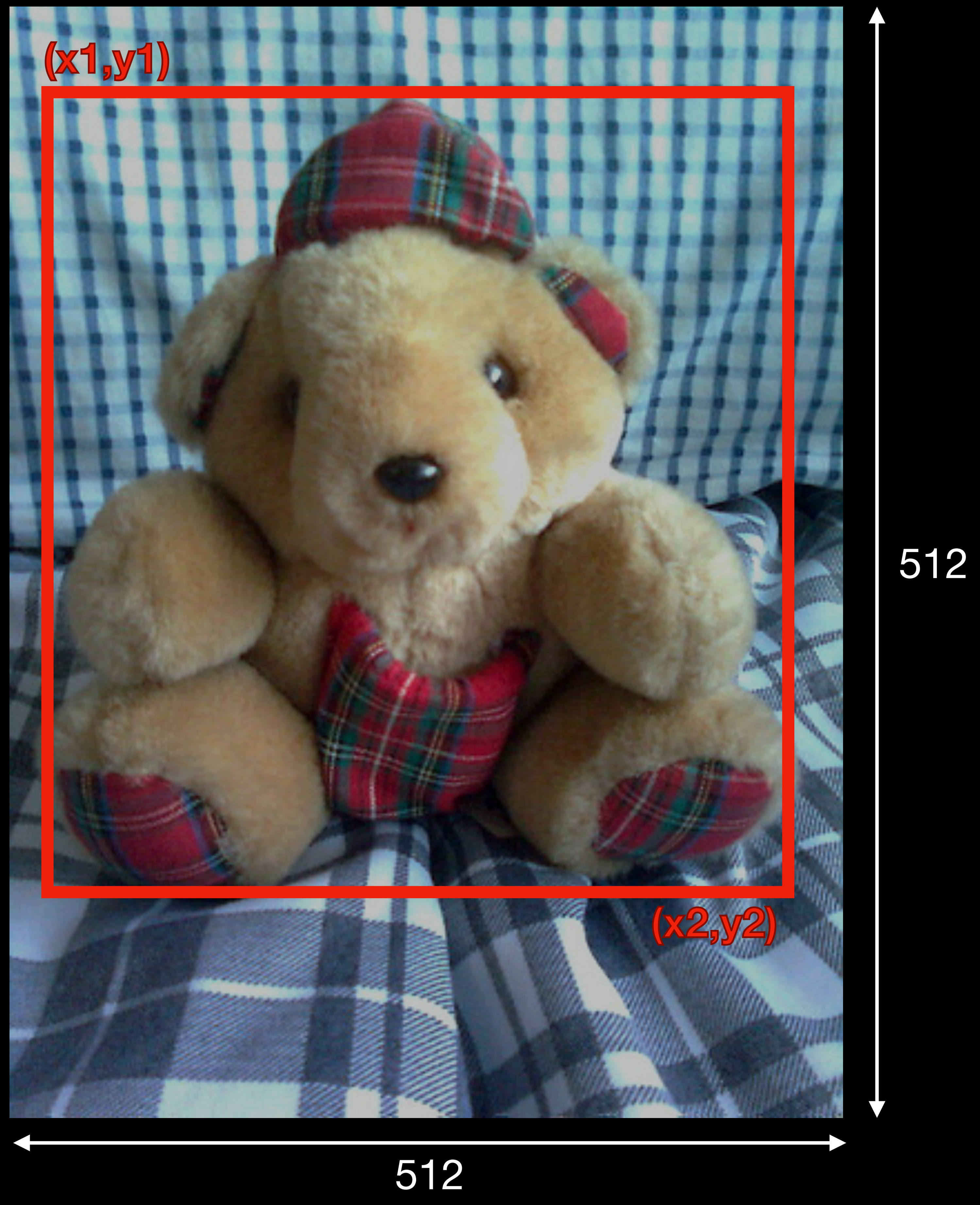
A. Normalized Floating Point Values
   e.g. [0.019, 0.114, 0.920, 0.786]



3.1. Coordinate representation

A. Normalized Floating Point Values
   e.g. [0.019, 0.114, 0.920, 0.786]

B. Integer Valued Binning (across image dimensions)
   e.g. ROUND $\left([\, 0.019, 0.114, 0.920, 0.786 \,] \times n_b\right)$
   = [4, 26, 206, 176] for $n_b = 224$

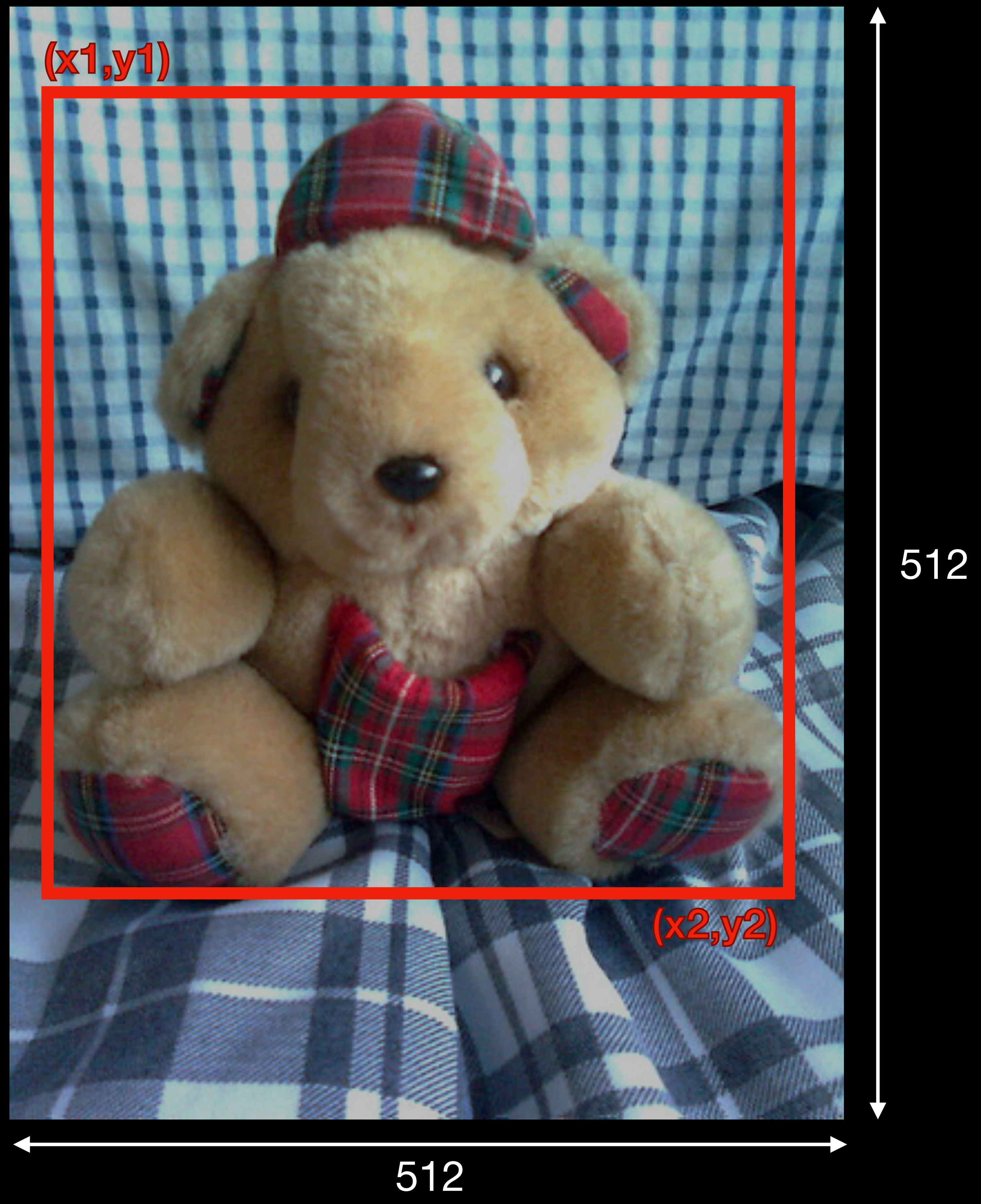


3.1. Coordinate representation

A. Normalized Floating Point Values
e.g. [0.019, 0.114, 0.920, 0.786]

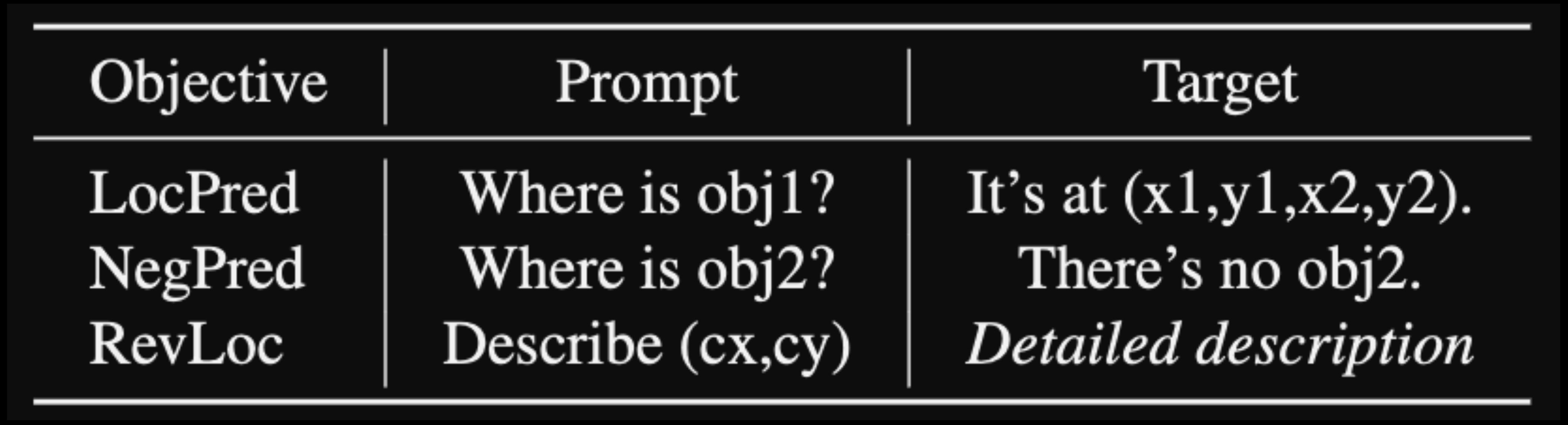
B. Integer Valued Binning (across image dimensions)
e.g. ROUND $\left([\ 0.019, 0.114, 0.920, 0.786\ ] \times n_b\right)$
= [4, 26, 206, 176] for $n_b = 224$

C. Deviation from Image-Grid based Anchors
e.g. [0, 4, 3, 11, 6, 0]



3.2. Instruction Fine-Tuning Objectives

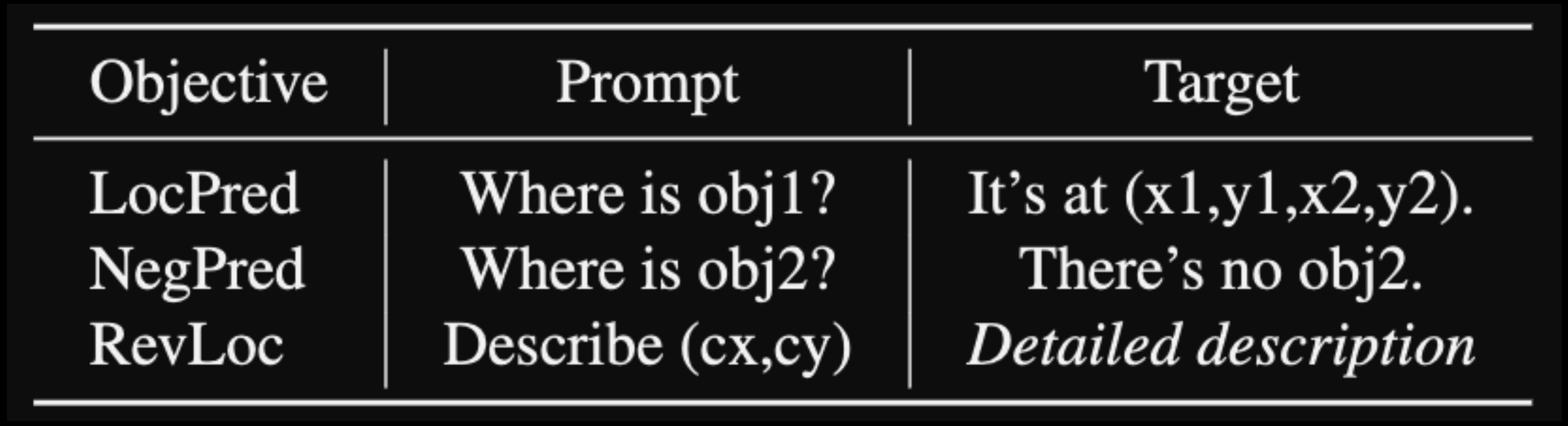
- Three distinct train objectives for instruction fine-tune stage



Objective	Prompt	Target
LocPred	Where is obj1?	It’s at (x1,y1,x2,y2).
NegPred	Where is obj2?	There’s no obj2.
RevLoc	Describe (cx,cy)	*Detailed description*

3.2. Instruction Fine-Tuning Objectives

- Three distinct train objectives for instruction fine-tune stage



Objective	Prompt	Target
LocPred	Where is obj1?	It’s at (x1,y1,x2,y2).
NegPred	Where is obj2?	There’s no obj2.
RevLoc	Describe (cx,cy)	*Detailed description*

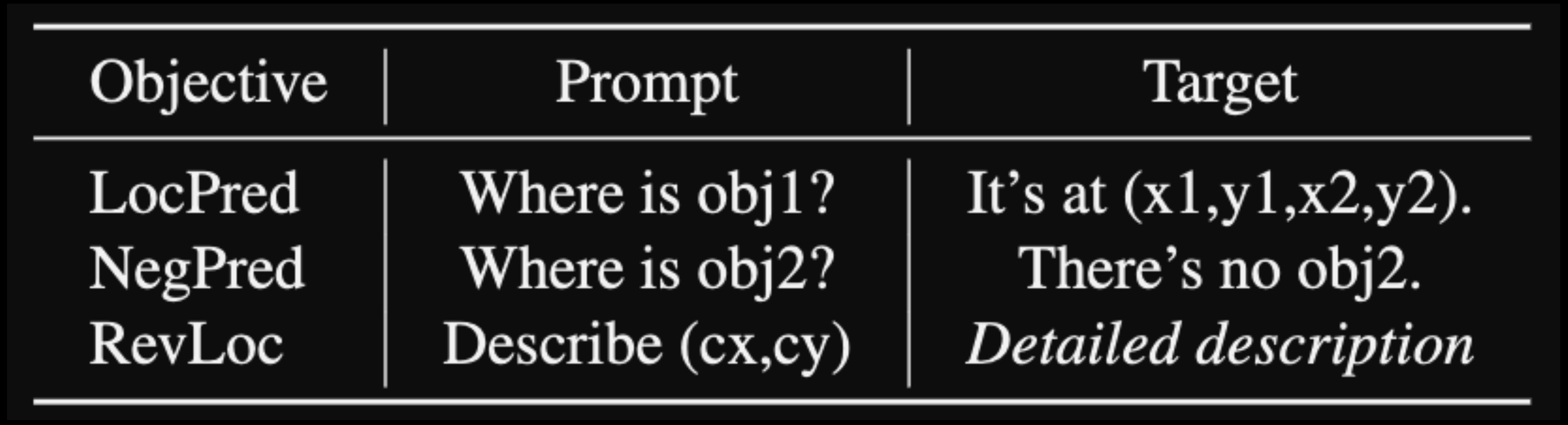
- LocPred
  generate coordinate outputs

- NegPred
  avoid hallucination

- RevLoc
  process coordinate inputs

3.2. Instruction Fine-Tuning Objectives

- Three distinct train objectives for instruction fine-tune stage



Objective	Prompt	Target
LocPred	Where is obj1?	It's at (x1,y1,x2,y2).
NegPred	Where is obj2?	There's no obj2.
RevLoc	Describe (cx,cy)	*Detailed description*

- LocPred
  generate coordinate outputs

- NegPred
  avoid hallucination

- RevLoc
  process coordinate inputs

PROMPT: "Where is the object described {category} located in image in terms of (x1,y1,x2,y2) bbox?"

LocPred TARGET: "It is located at {location} bbox."
NegPred TARGET: "There is no such object in the image."

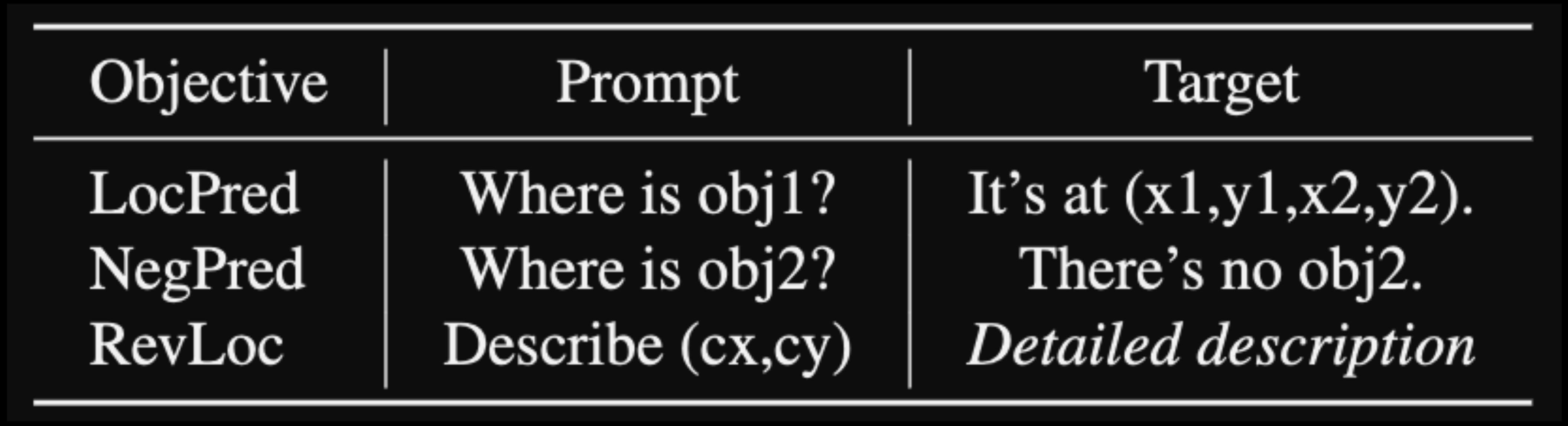
PROMPT: "Describe the object located at {loc} bbox?"
RevLoc TARGET: "There is a {category/description}."

3.2. Instruction Fine-Tuning Objectives

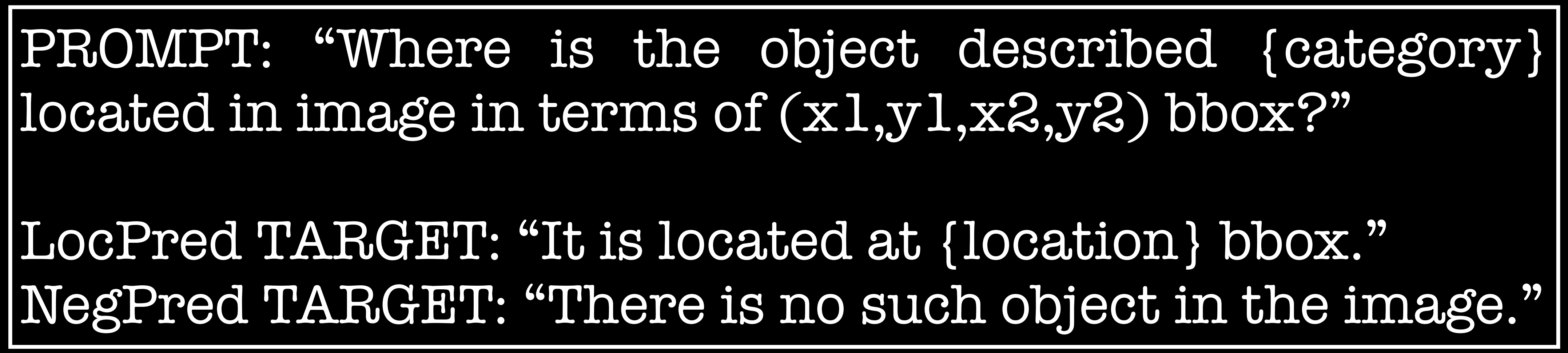
- Three distinct train objectives for instruction fine-tune stage

- LocPred
  generate coordinate outputs
- NegPred
  avoid hallucination
- RevLoc
  process coordinate inputs

Train with these objectives for preliminary model

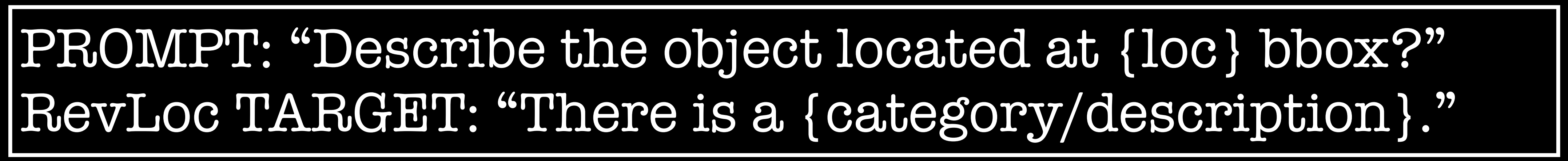


Objective	Prompt	Target
LocPred	Where is obj1?	It's at (x1,y1,x2,y2).
NegPred	Where is obj2?	There's no obj2.
RevLoc	Describe (cx,cy)	*Detailed description*



PROMPT: "Where is the object described {category} located in image in terms of (x1,y1,x2,y2) bbox?"

LocPred TARGET: "It is located at {location} bbox."
NegPred TARGET: "There is no such object in the image."



PROMPT: "Describe the object located at {loc} bbox?"
RevLoc TARGET: "There is a {category/description}."

3.3. Pseudo Data

Generate dataset: Object Location + Description Pairs

3.3. Pseudo Data

Generate dataset: Object Location + Description Pairs

- Inputs: Images with object bounding box annotations
    - GT or Object-Detector

3.3. Pseudo Data

Generate dataset: Object Location + Description Pairs

- Inputs: Images with object bounding box annotations
    - GT or Object-Detector

- V-LLM based object description
    - Contextual descriptions (describe object relative to surroundings)
    - Use images with single instance of object category (crop / filter)

3.3. Pseudo Data

Generate dataset: Object Location + Description Pairs

- Inputs: Images with object bounding box annotations
    - GT or Object-Detector

- V-LLM based object description
    - Contextual descriptions (describe object relative to surroundings)
    - Use images with single instance of object category (crop / filter)

PROMPT: “Describe the {category} in this image using one short sentence, referring to its visual features and spatial position relative to other objects in image.”

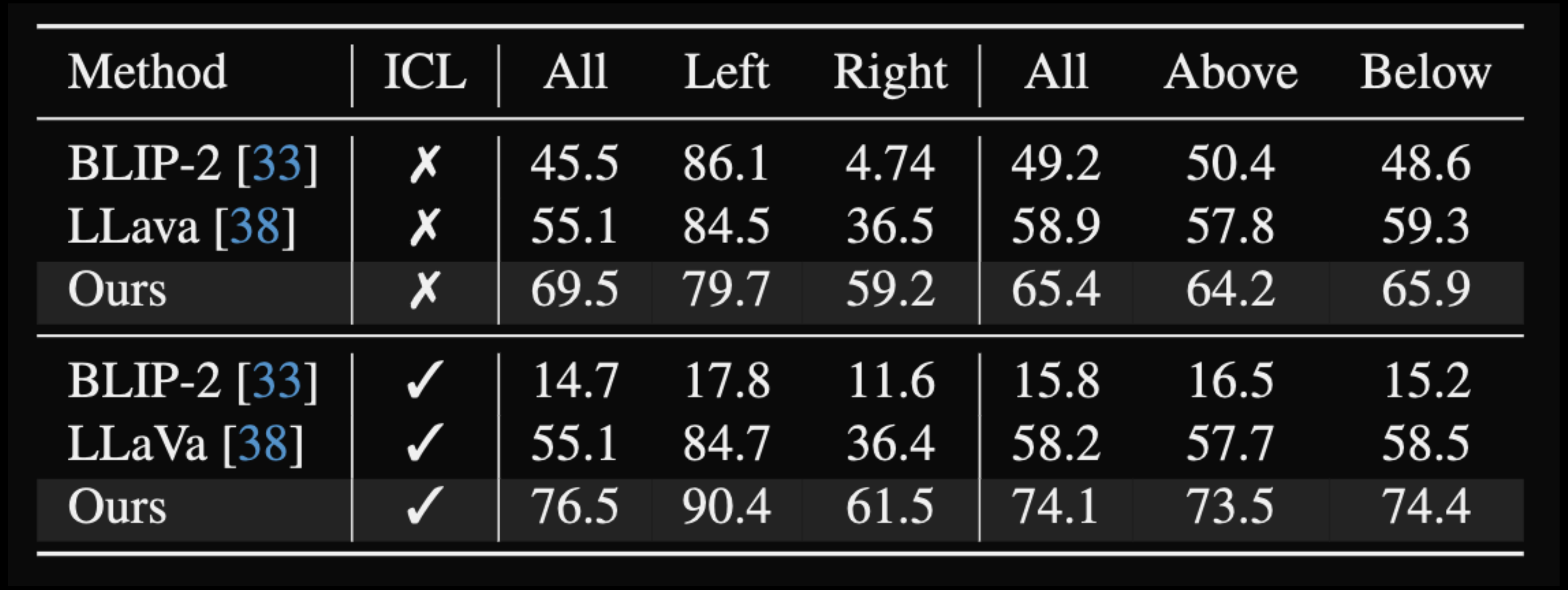
- Pre-training stage similar to LLaVA
- Fine-tuning using our proposed objectives and generated data

Resulting Model termed “**LocVLM**”

4. Findings

4.1. Improved VQA

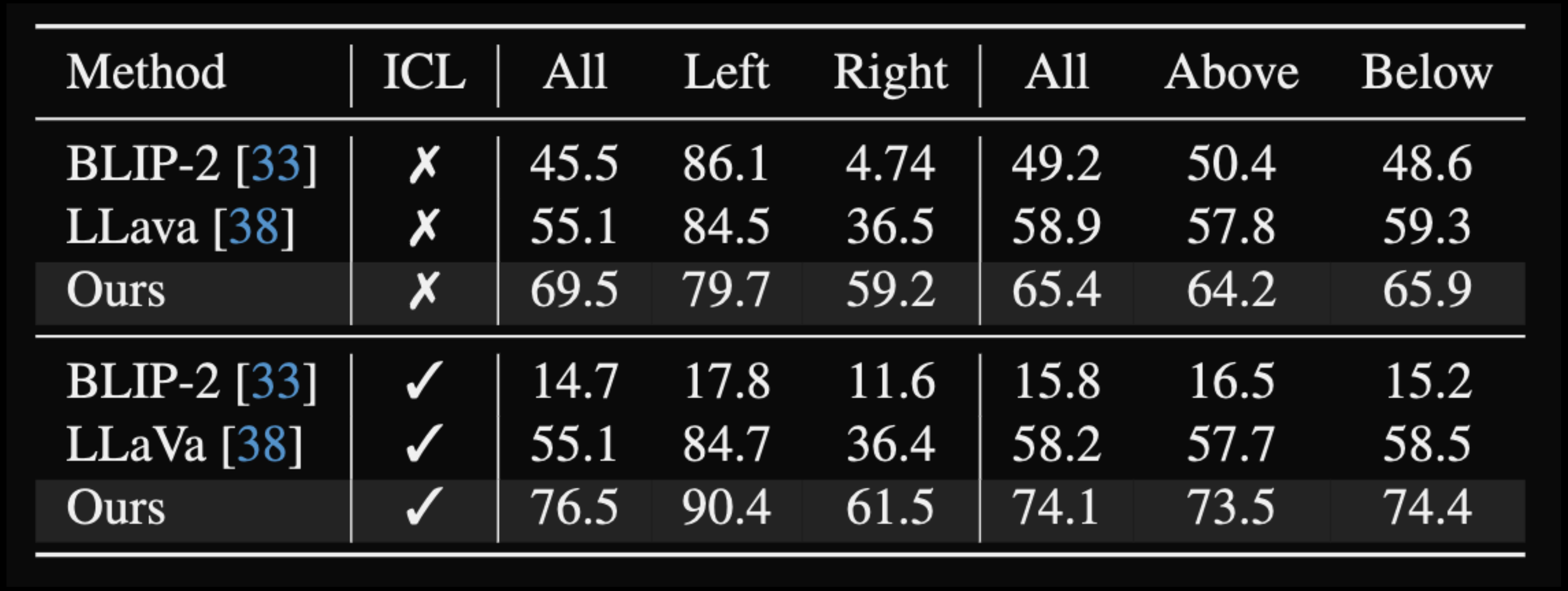
- Toy Exp: “Which side of image is object?”
    - High accuracy predicting left vs right correctly
    - High accuracy predicting top vs bottom correctly



Method	ICL	All	Left	Right	All	Above	Below
BLIP-2 [33]	✗	45.5	86.1	4.74	49.2	50.4	48.6
LLava [38]	✗	55.1	84.5	36.5	58.9	57.8	59.3
Ours	✗	69.5	79.7	59.2	65.4	64.2	65.9
BLIP-2 [33]	✓	14.7	17.8	11.6	15.8	16.5	15.2
LLaVa [38]	✓	55.1	84.7	36.4	58.2	57.7	58.5
Ours	✓	76.5	90.4	61.5	74.1	73.5	74.4

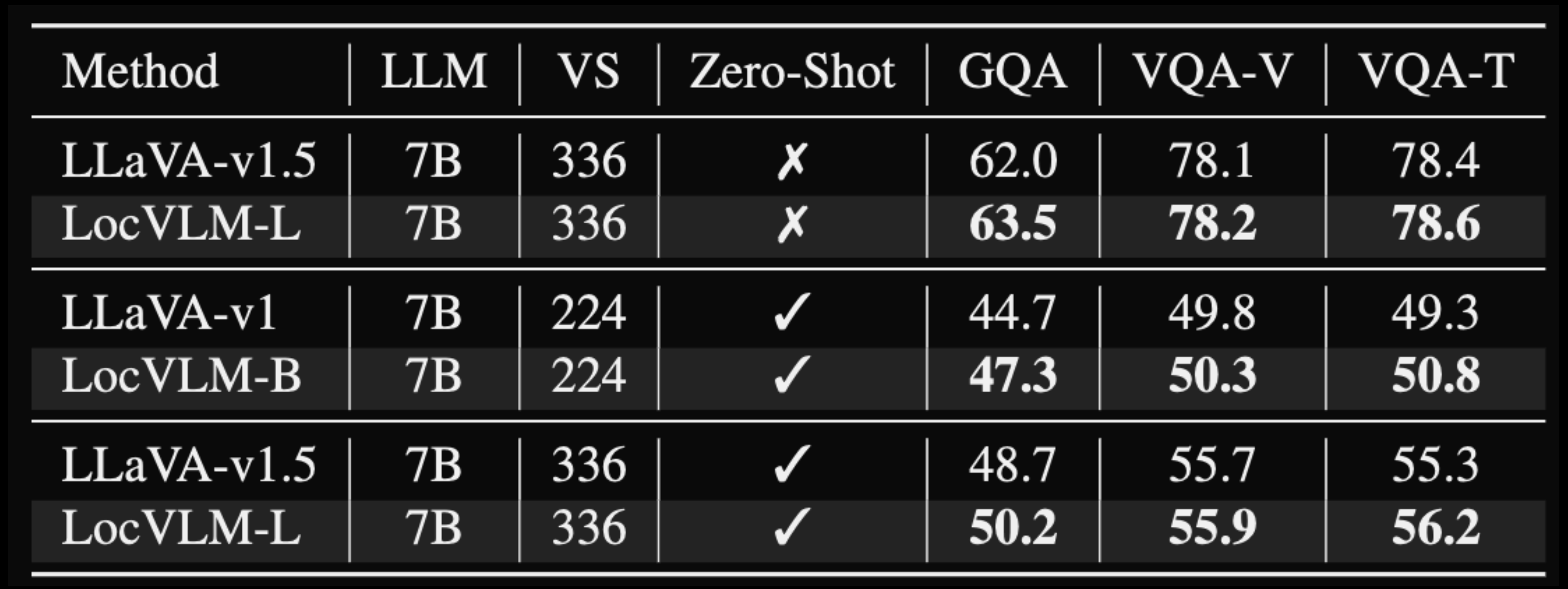
4.1. Improved VQA

- Toy Exp: “Which side of image is object?”
    - High accuracy predicting left vs right correctly
    - High accuracy predicting top vs bottom correctly



Method	ICL	All	Left	Right	All	Above	Below
BLIP-2 [33]	✗	45.5	86.1	4.74	49.2	50.4	48.6
LLava [38]	✗	55.1	84.5	36.5	58.9	57.8	59.3
Ours	✗	69.5	79.7	59.2	65.4	64.2	65.9
BLIP-2 [33]	✓	14.7	17.8	11.6	15.8	16.5	15.2
LLaVa [38]	✓	55.1	84.7	36.4	58.2	57.7	58.5
Ours	✓	76.5	90.4	61.5	74.1	73.5	74.4

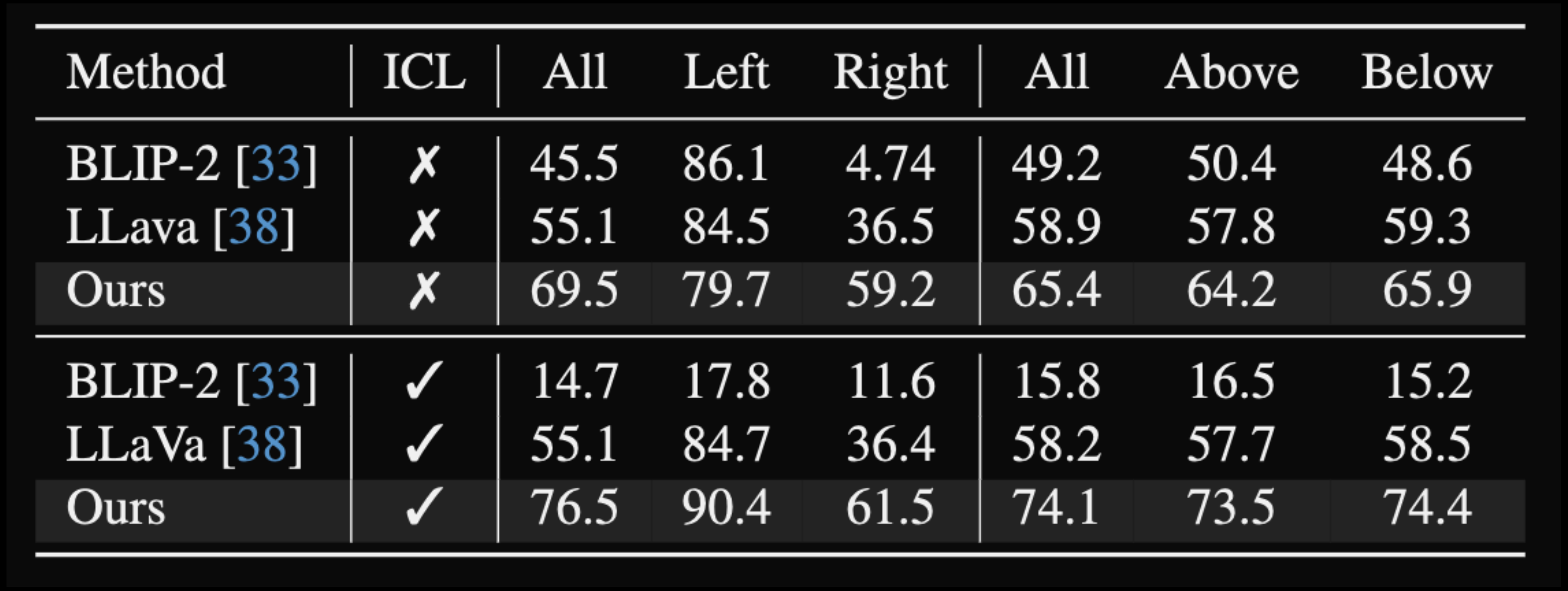
- Image and Video (frame-average) VQA
    - Improves over baselines



Method	LLM	VS	Zero-Shot	GQA	VQA-V	VQA-T
LLaVA-v1.5	7B	336	✗	62.0	78.1	78.4
LocVLM-L	7B	336	✗	**63.5**	**78.2**	**78.6**
LLaVA-v1	7B	224	✓	44.7	49.8	49.3
LocVLM-B	7B	224	✓	**47.3**	**50.3**	**50.8**
LLaVA-v1.5	7B	336	✓	48.7	55.7	55.3
LocVLM-L	7B	336	✓	**50.2**	**55.9**	**56.2**

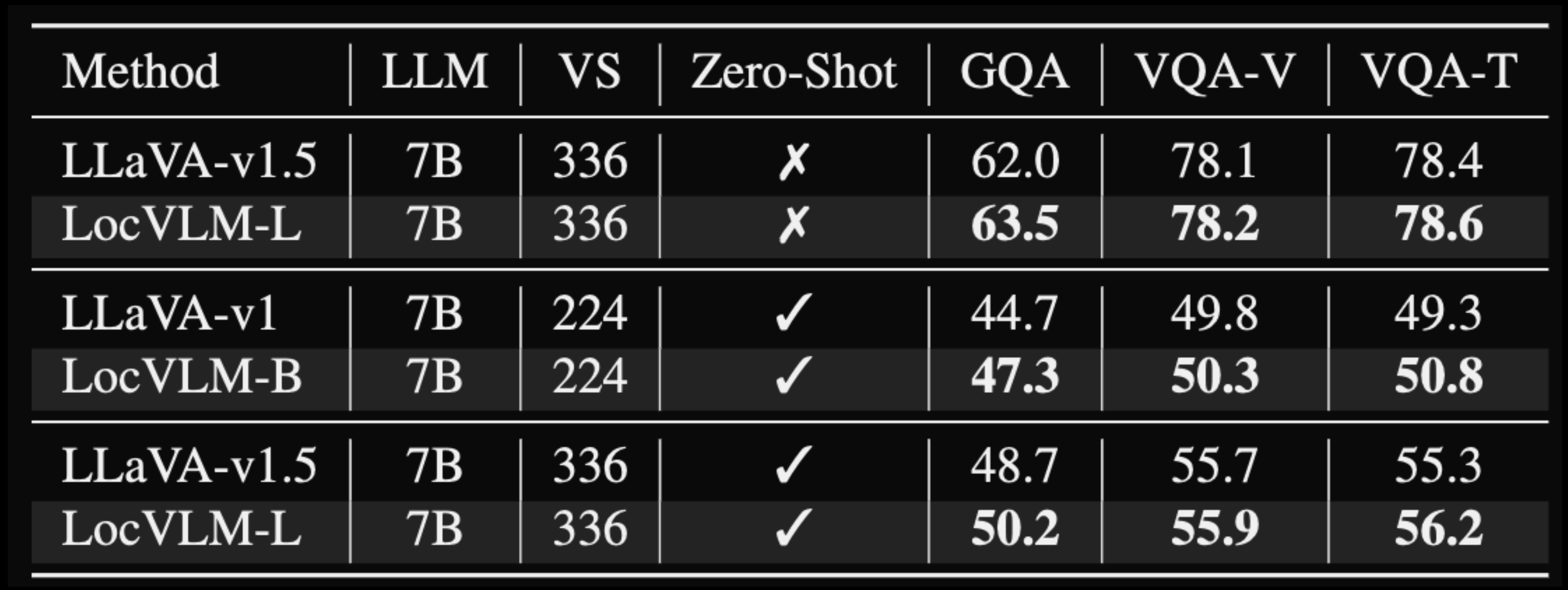
4.1. Improved VQA

- Toy Exp: “Which side of image is object?”
    - High accuracy predicting left vs right correctly
    - High accuracy predicting top vs bottom correctly

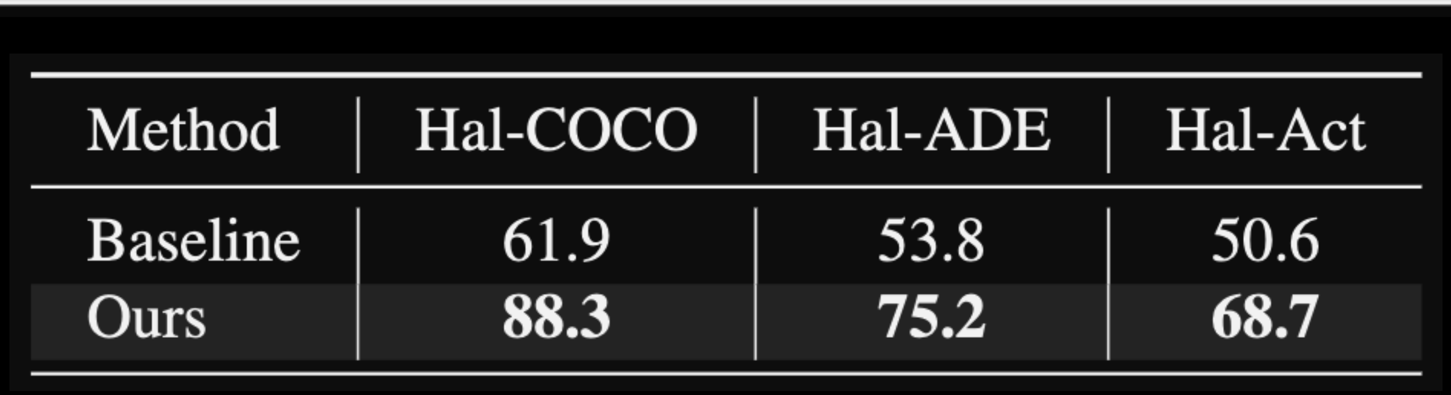


Method	ICL	All	Left	Right	All	Above	Below
BLIP-2 [33]	✗	45.5	86.1	4.74	49.2	50.4	48.6
LLava [38]	✗	55.1	84.5	36.5	58.9	57.8	59.3
Ours	✗	69.5	79.7	59.2	65.4	64.2	65.9
BLIP-2 [33]	✓	14.7	17.8	11.6	15.8	16.5	15.2
LLaVa [38]	✓	55.1	84.7	36.4	58.2	57.7	58.5
Ours	✓	76.5	90.4	61.5	74.1	73.5	74.4

- Image and Video (frame-average) VQA
    - Improves over baselines
    - Reduces object hallucination (including for unseen categories)



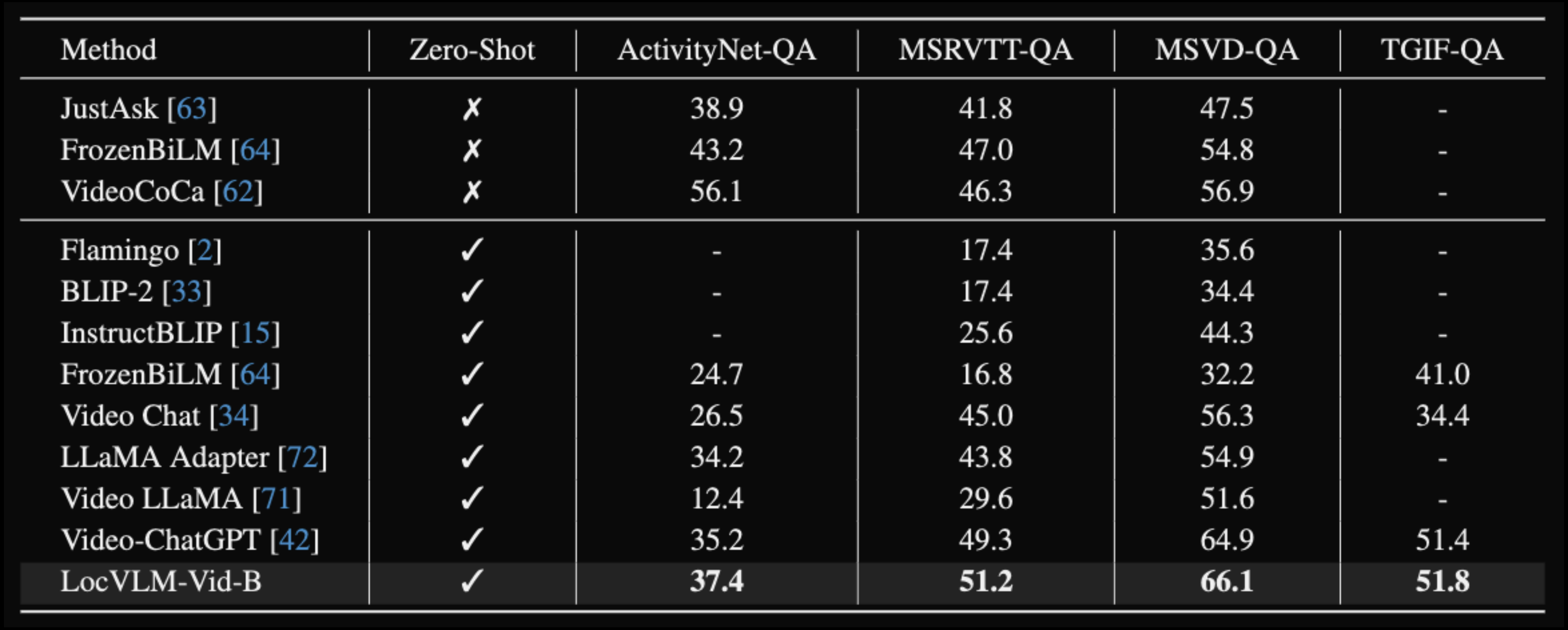
Method	LLM	VS	Zero-Shot	GQA	VQA-V	VQA-T
LLaVA-v1.5	7B	336	✗	62.0	78.1	78.4
LocVLM-L	7B	336	✗	**63.5**	**78.2**	**78.6**
LLaVA-v1	7B	224	✓	44.7	49.8	49.3
LocVLM-B	7B	224	✓	**47.3**	**50.3**	**50.8**
LLaVA-v1.5	7B	336	✓	48.7	55.7	55.3
LocVLM-L	7B	336	✓	**50.2**	**55.9**	**56.2**



Method	Hal-COCO	Hal-ADE	Hal-Act
Baseline	61.9	53.8	50.6
Ours	**88.3**	**75.2**	**68.7**

4.1. Improved VQA

Video Domain: adding our learned LLM to a video baseline

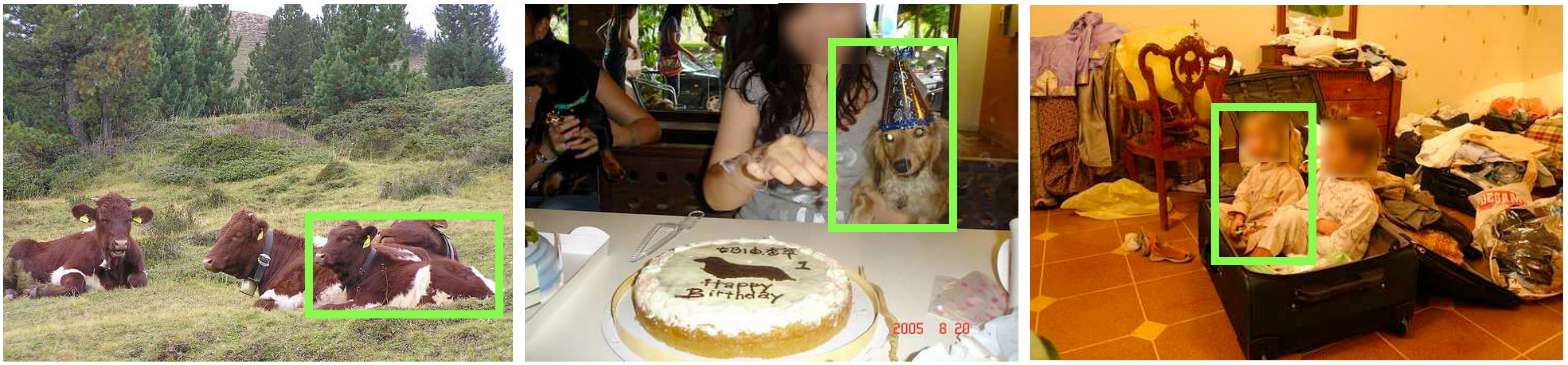


Method	Zero-Shot	ActivityNet-QA	MSRVTT-QA	MSVD-QA	TGIF-QA
JustAsk [63]	✗	38.9	41.8	47.5	-
FrozenBiLM [64]	✗	43.2	47.0	54.8	-
VideoCoCa [62]	✗	56.1	46.3	56.9	-
Flamingo [2]	✓	-	17.4	35.6	-
BLIP-2 [33]	✓	-	17.4	34.4	-
InstructBLIP [15]	✓	-	25.6	44.3	-
FrozenBiLM [64]	✓	24.7	16.8	32.2	41.0
Video Chat [34]	✓	26.5	45.0	56.3	34.4
LLaMA Adapter [72]	✓	34.2	43.8	54.9	-
Video LLaMA [71]	✓	12.4	29.6	51.6	-
Video-ChatGPT [42]	✓	35.2	49.3	64.9	51.4
LocVLM-Vid-B	✓	**37.4**	**51.2**	**66.1**	**51.8**

4.2. Novel Skills

Contextual Description of Regions

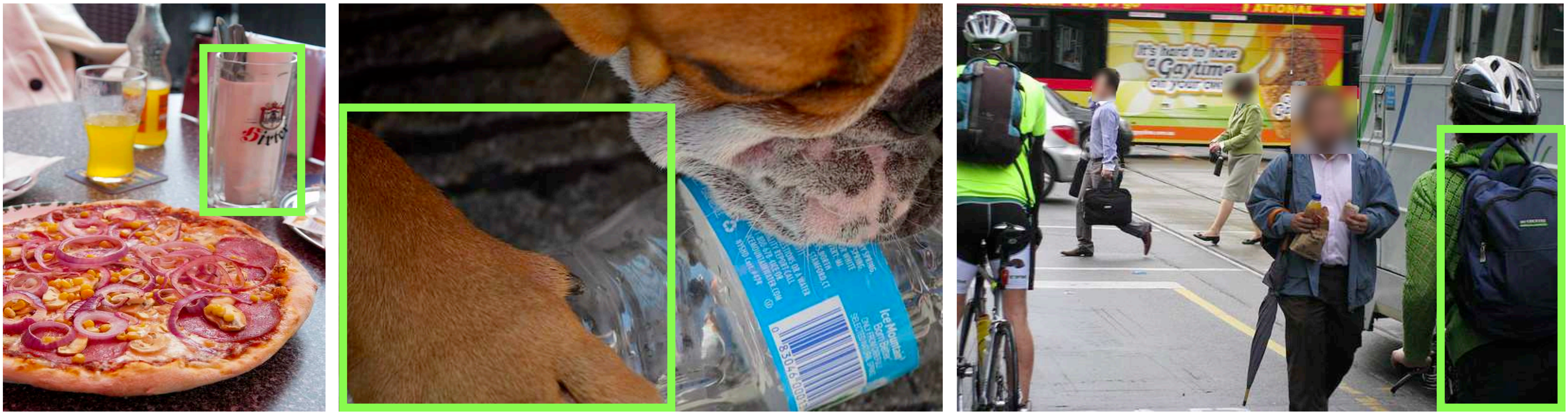
Prompt: Describe the region described by (x1,y1, x2, y2) bounding box.



There is a cow that is lying down on a grassy hillside, surrounded by other cows and trees.

There is a dog that is standing next to a cake on a table, looking at the cake with a candle on it.

There is a person that is a young girl, sitting in a suitcase filled with clothes and other items.



There is a cup that is a tall glass, placed on a table next to a pizza.

There is a dog that is a brown and white dog, and it is standing next to a bottle of water, possibly drinking from it.

There is a backpack that is green and is positioned on the back of a person, likely a man, who is walking down the street.

4.2. Novel Skills

Contextual Description of Regions

Prompt: Describe the region described by (x1,y1, x2, y2) bounding box.

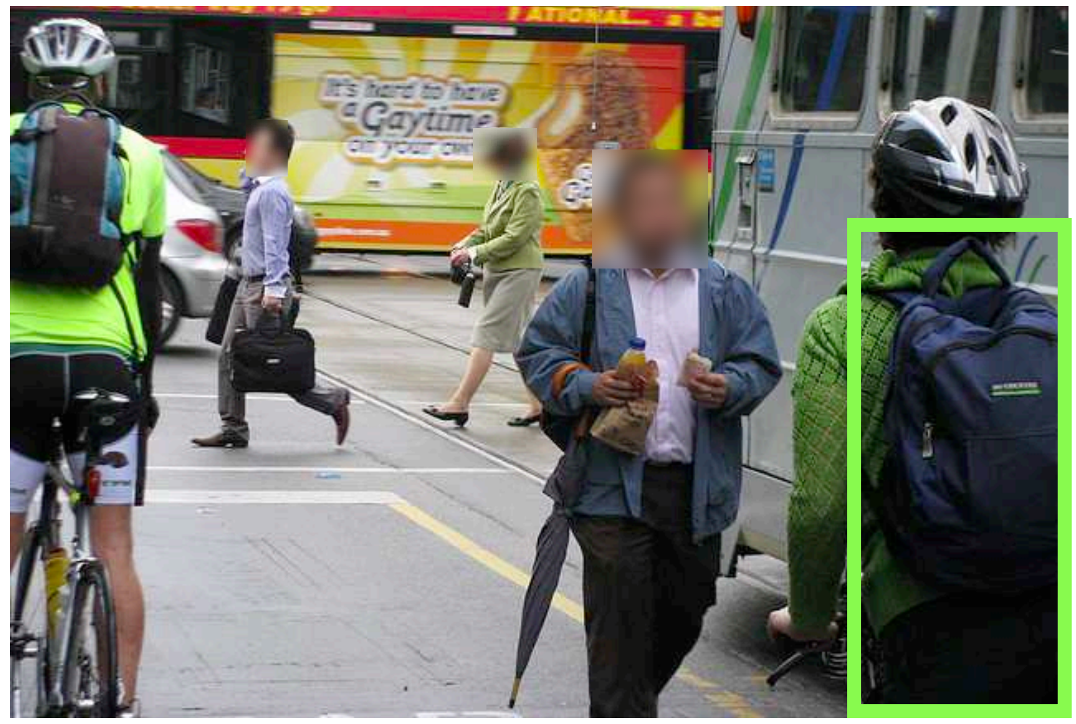
There is a cow that is lying down on a grassy hillside, surrounded by other cows and trees.

There is a dog that is standing next to a cake on a table, looking at the cake with a candle on it.

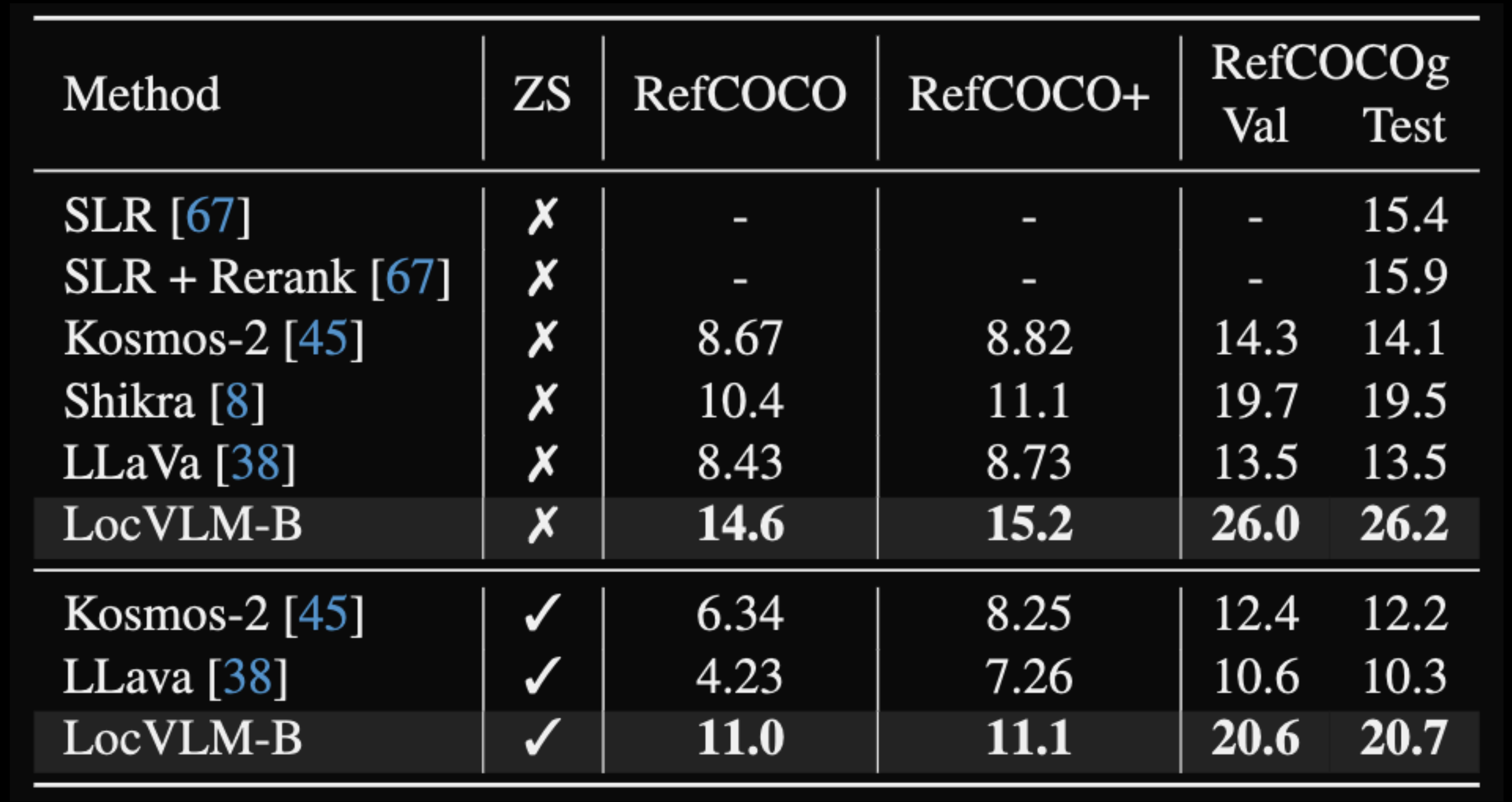
There is a person that is a young girl, sitting in a suitcase filled with clothes and other items.

There is a cup that is a tall glass, placed on a table next to a pizza.

There is a dog that is a brown and white dog, and it is standing next to a bottle of water, possibly drinking from it.



There is a backpack that is green and is positioned on the back of a person, likely a man, who is walking down the street.

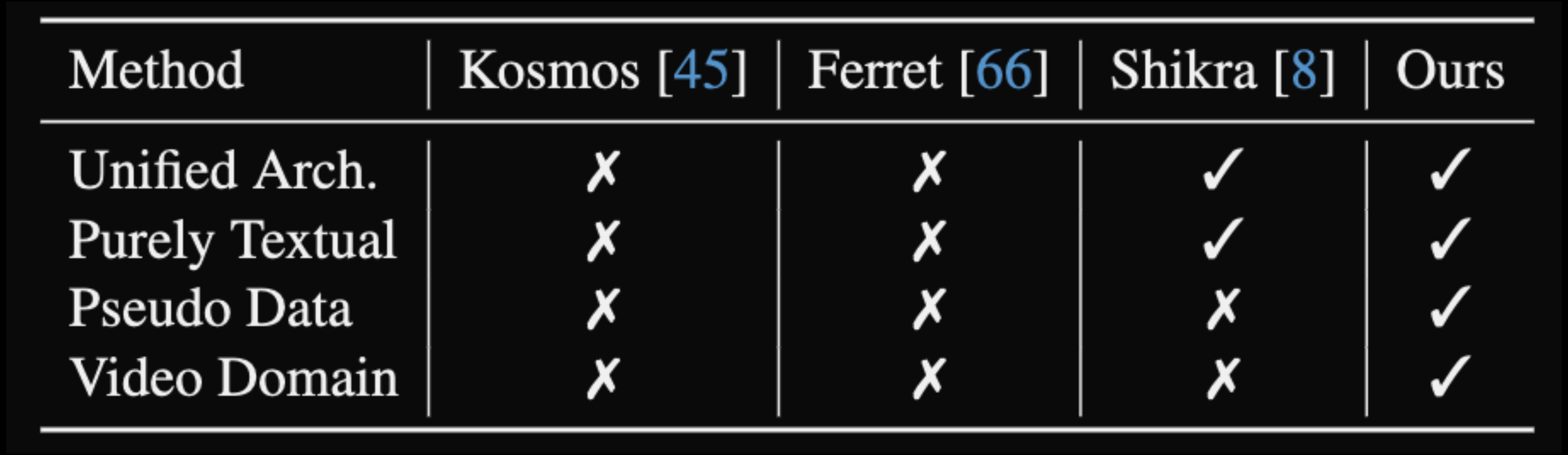


Method	ZS	RefCOCO	RefCOCO+	RefCOCOg Val	RefCOCOg Test
SLR [67]	✗	-	-	-	15.4
SLR + Rerank [67]	✗	-	-	-	15.9
Kosmos-2 [45]	✗	8.67	8.82	14.3	14.1
Shikra [8]	✗	10.4	11.1	19.7	19.5
LLaVa [38]	✗	8.43	8.73	13.5	13.5
LocVLM-B	✗	**14.6**	**15.2**	**26.0**	**26.2**
Kosmos-2 [45]	✓	6.34	8.25	12.4	12.2
LLava [38]	✓	4.23	7.26	10.6	10.3
LocVLM-B	✓	**11.0**	**11.1**	**20.6**	**20.7**

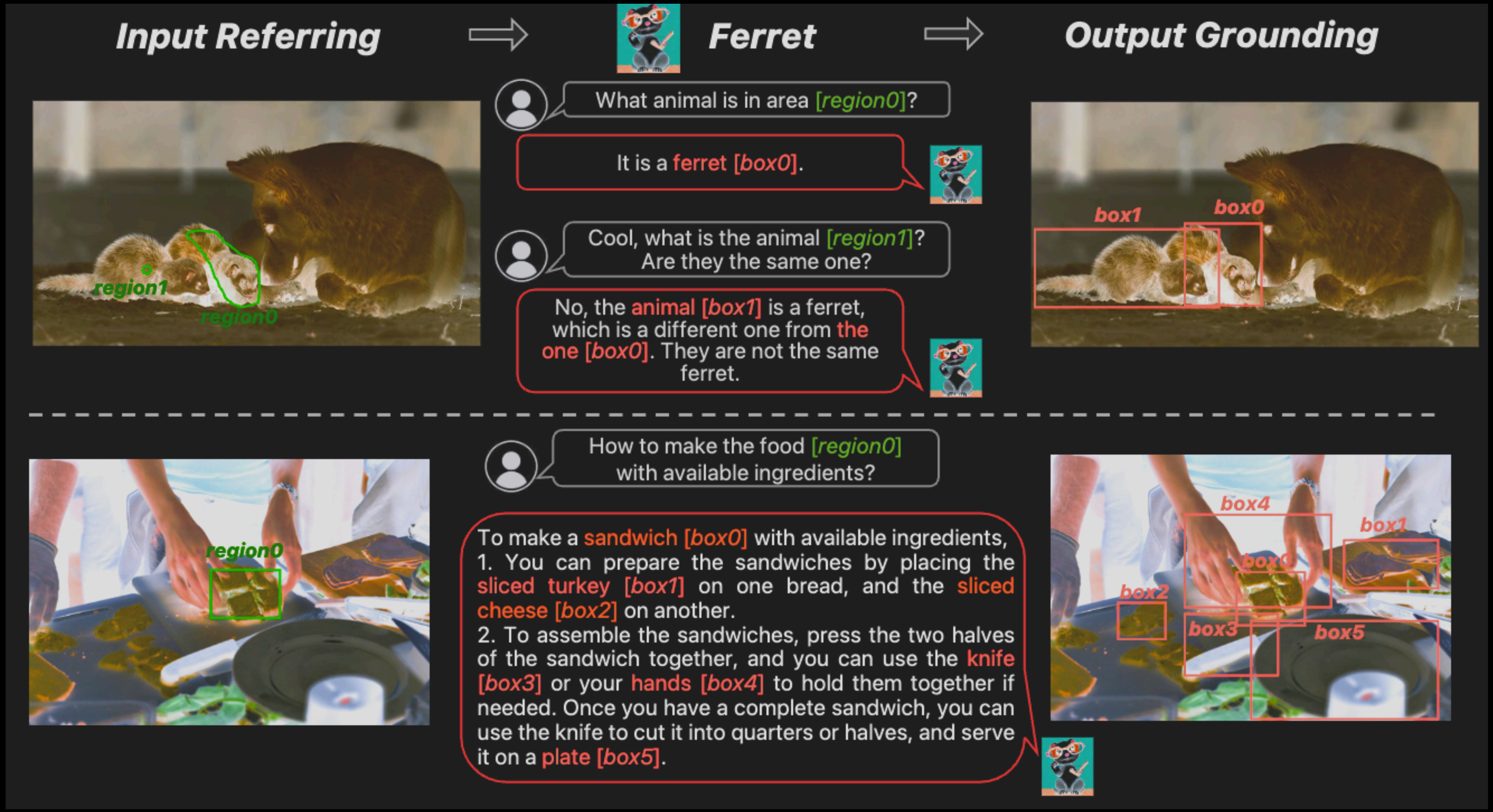
Region Description Task: Evaluation for the reverse of referring object detection. Given a bounding box, generate a region description using contextual information as well. The METEOR scores (text similarity) is calculated against GT human-written captions for each object region.

Contemporary Works

- Several recent works explore similar ideas for VQA
- Overlap but also some distinctions



Method	Kosmos [45]	Ferret [66]	Shikra [8]	Ours
Unified Arch.	✗	✗	✓	✓
Purely Textual	✗	✗	✓	✓
Pseudo Data	✗	✗	✗	✓
Video Domain	✗	✗	✗	✓



What animal is in area [region0]?

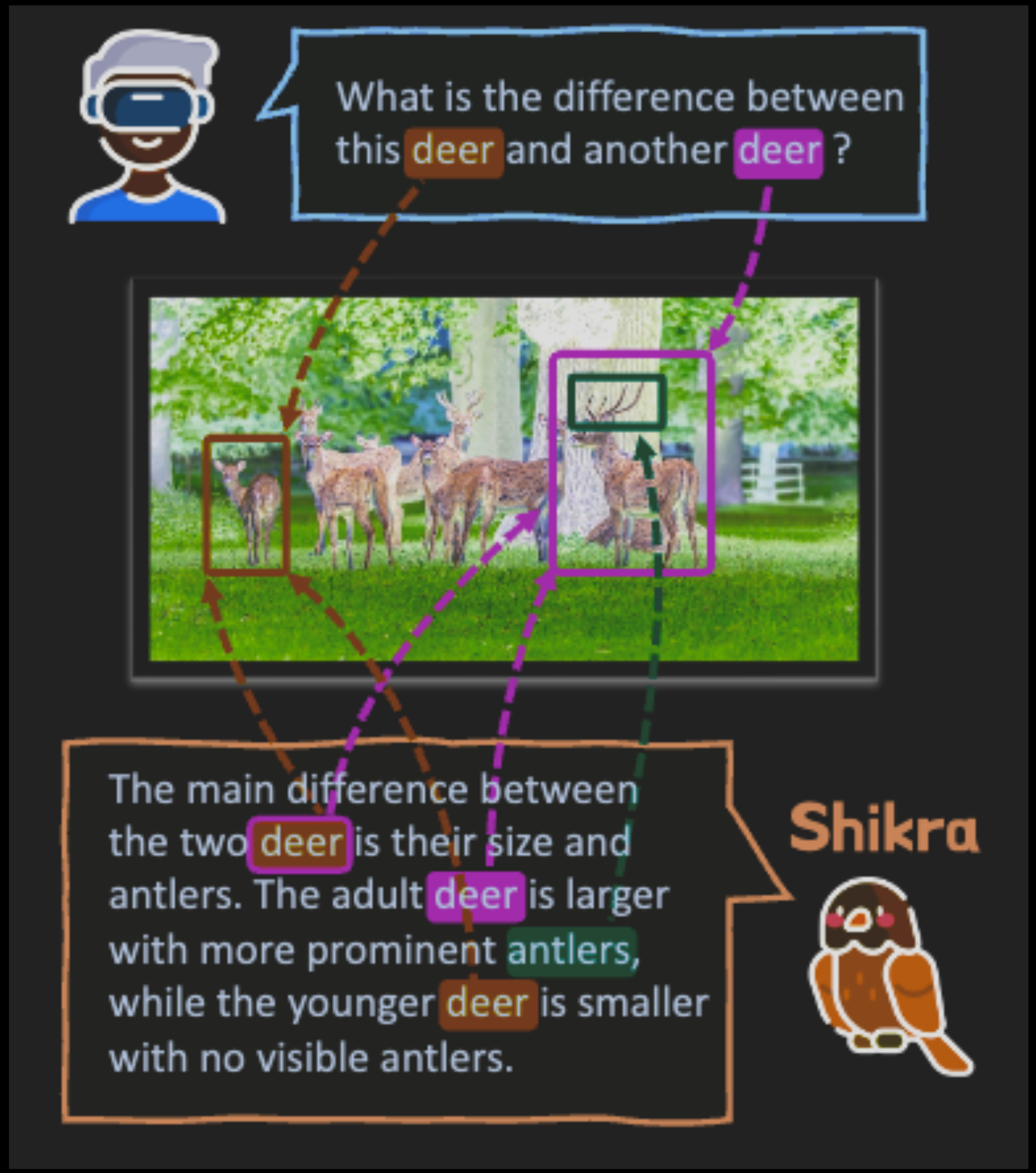
It is a ferret [box0].

Cool, what is the animal [region1]? Are they the same one?

No, the animal [box1] is a ferret, which is a different one from the one [box0]. They are not the same ferret.

How to make the food [region0] with available ingredients?

To make a sandwich [box0] with available ingredients,
1. You can prepare the sandwiches by placing the sliced turkey [box1] on one bread, and the sliced cheese [box2] on another.
2. To assemble the sandwiches, press the two halves of the sandwich together, and you can use the knife [box3] or your hands [box4] to hold them together if needed. Once you have a complete sandwich, you can use the knife to cut it into quarters or halves, and serve it on a plate [box5].



What is the difference between this deer and another deer ?

The main difference between the two deer is their size and antlers. The adult deer is larger with more prominent antlers, while the younger deer is smaller with no visible antlers.

[8] Chen, Ke et al. “Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic.” ArXiv abs/2306.15195 (2023)
[66] You, Haoxuan et al. “Ferret: Refer and Ground Anything Anywhere at Any Granularity.” ICLR 2024
[45] Peng, Zhiliang et al. “Kosmos-2: Grounding Multimodal Large Language Models to the World.” ICLR 2024

5. Discussion

- Can we modify visual LLMs to understand image-space coordinates as text?
  Yes! Performs on par with alternate approaches

- Does this improve general VQA?
  Yes. Better spatial awareness (on selected settings)
  + Reduced object hallucination.

- Any new abilities of these models?
  Contextual descriptions for object regions