Code

Paper

Ferret: Refer and Ground Anything Anywhere at Any Granularity

Haoxuan You[1]*, Haotian Zhang[2]* ***[*equal contribution]***
Zhe Gan[2], Xianzhi Du[2], Bowen Zhang[2], Zirui Wang[2], Liangliang Cao[2], Shih-Fu Chang[2], Yinfei Yang[2]
[1]Columbia University & [2]Apple AI/ML

Is GPT-4V/Bard perfect?

Observation: They are powerful in understanding global image semantic, but struggle with **Spatial&Regional Understanding**.

Why is **MLLM + Spatial Understanding** important?

Why is MLLM + Spatial Understanding important?

New Functions:

1. Users to refer to specific regions/objects and ask model's help.
2. Model to localize/ground particular objects in response for better helping users.

Better Model:

1. Less Hallucination
2. More Trustworthy
3. Open-Vocabulary Concept Grounding

New Applications:

1. Phone/VR/AR Assistant
2. Robotics
3. Medical Assistant
4. …

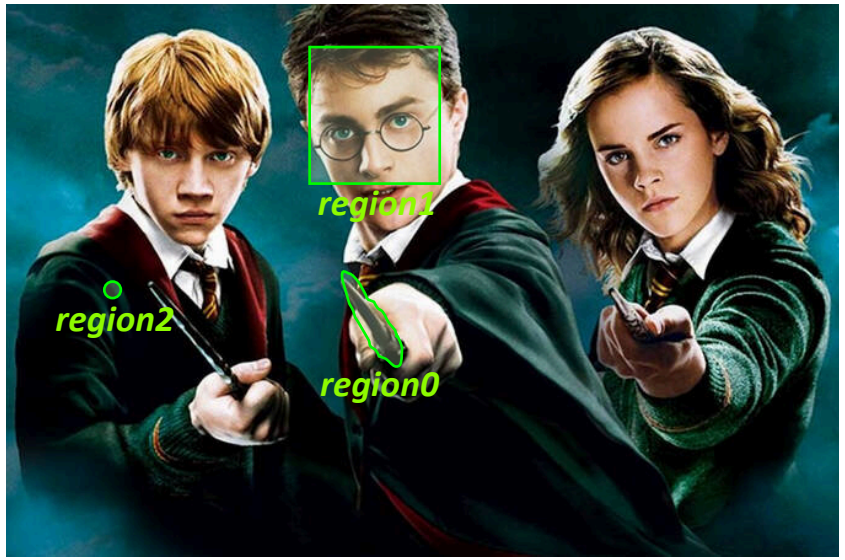
Building Ferret, a MLLM w/ Strong Spatial Understand

- Problem Definition
- Model Structure
- Data Collection
- Evaluation (Ferret-Bench) and Ablation

Problem Definition

Spatial Understanding can be reflected in two types of tasks:

1. Referring
   Input: Image + Text Instruction + Region
   Model is required to understand the referred regions and respond to the instruction.



*Q: What is in **region0**? What is it used for?*

*Q: Which movie characters are in **region1** and **region2**? And what is their relationship?*

Problem Definition

Spatial Understanding can be reflected in two types of tasks:

2. Grounding

    Output: Text Response + Region

    Model is required to localize the objects in image when mentioning them in response

Q: How to make a sandwich with available ingredients in the image? And where are they?

Region Definition

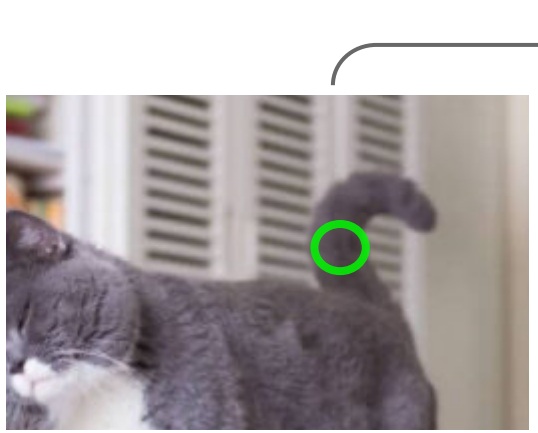
Region in Input:
- Point
- Box
- Free-Form Shape: Scribble, Segmentation Mask, …

} Hybrid Region Representation

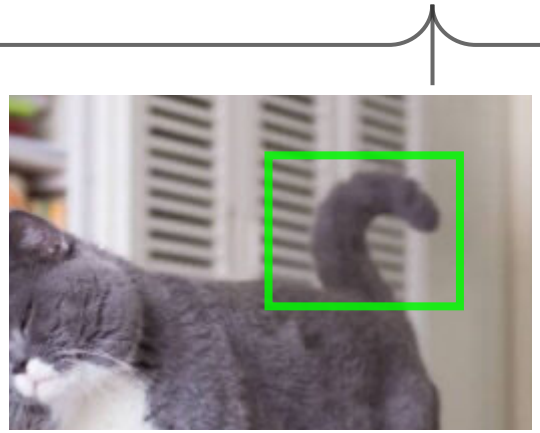
Region in Output:
- Box → Coordinates
- Free-Form Shape

Hybrid Region Representation

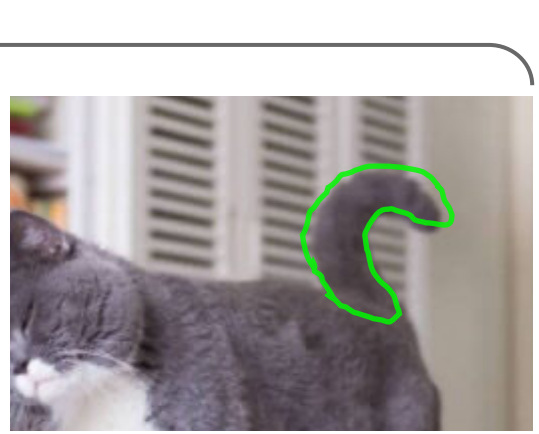
Region Name + Discrete Coordinate + Continuous Feature



Point



Box



Free-form Shape
(Sketch, Scribble, polygons)

Hybrid Region Representation

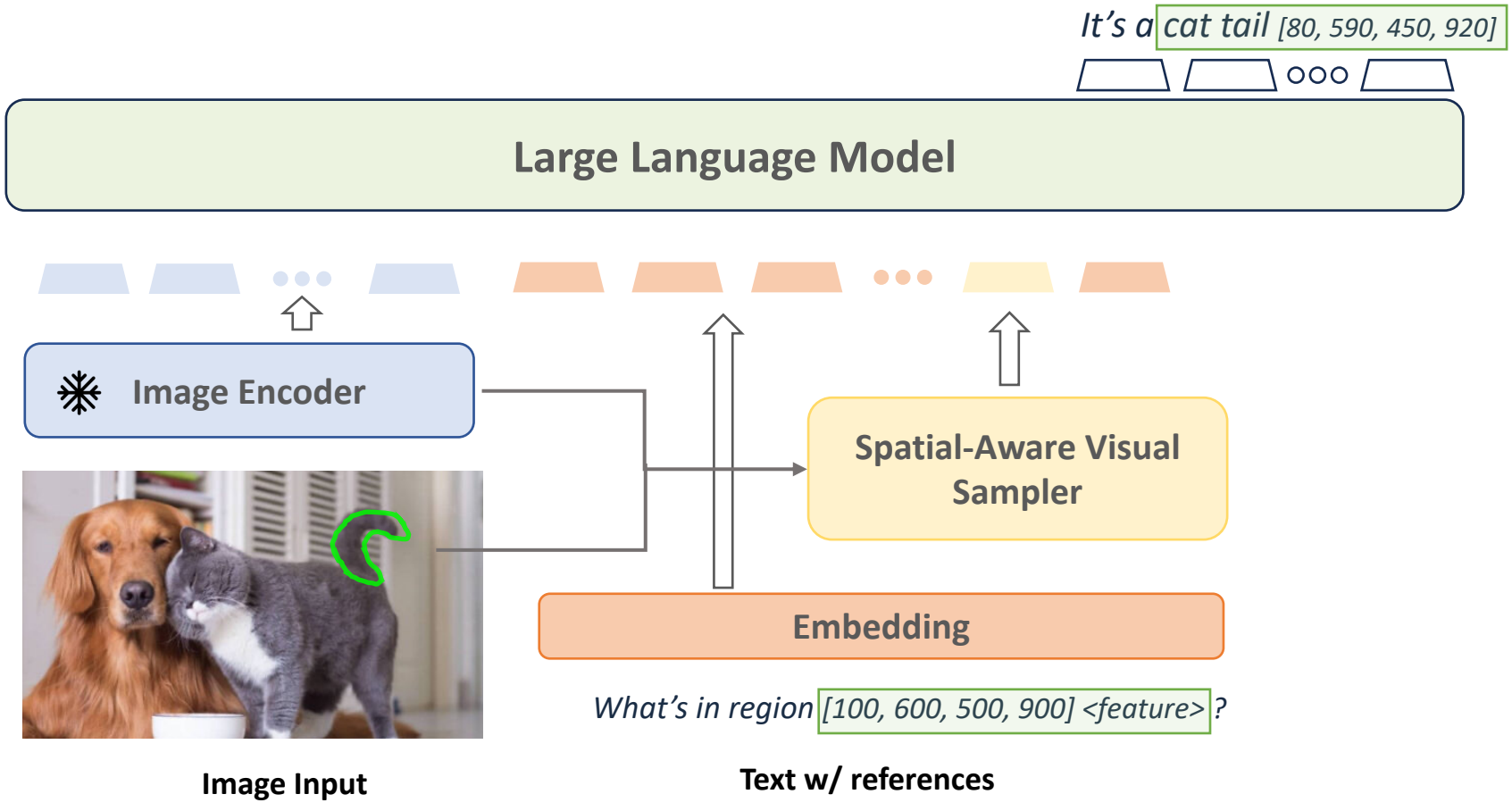
- **Discrete Coordinates**
  - Point: [x, y] (center point)
    Box and Free-form Shape: [x1, y1, x2, y2] (top-left and bottom-right points)
  - Tokenize them by LLM tokenizer.

- **Continuous Visual Features**.
  - Introduce a Visual Sampler module to extract and summarize visual features of **referred regions (point -> circle)** into a single feature vector.

- Examples of data:
  Input:
  What is in region [100, 600, 500, 900] <feature>?
  Output:
  It's a box of egg [100, 600, 500, 900].

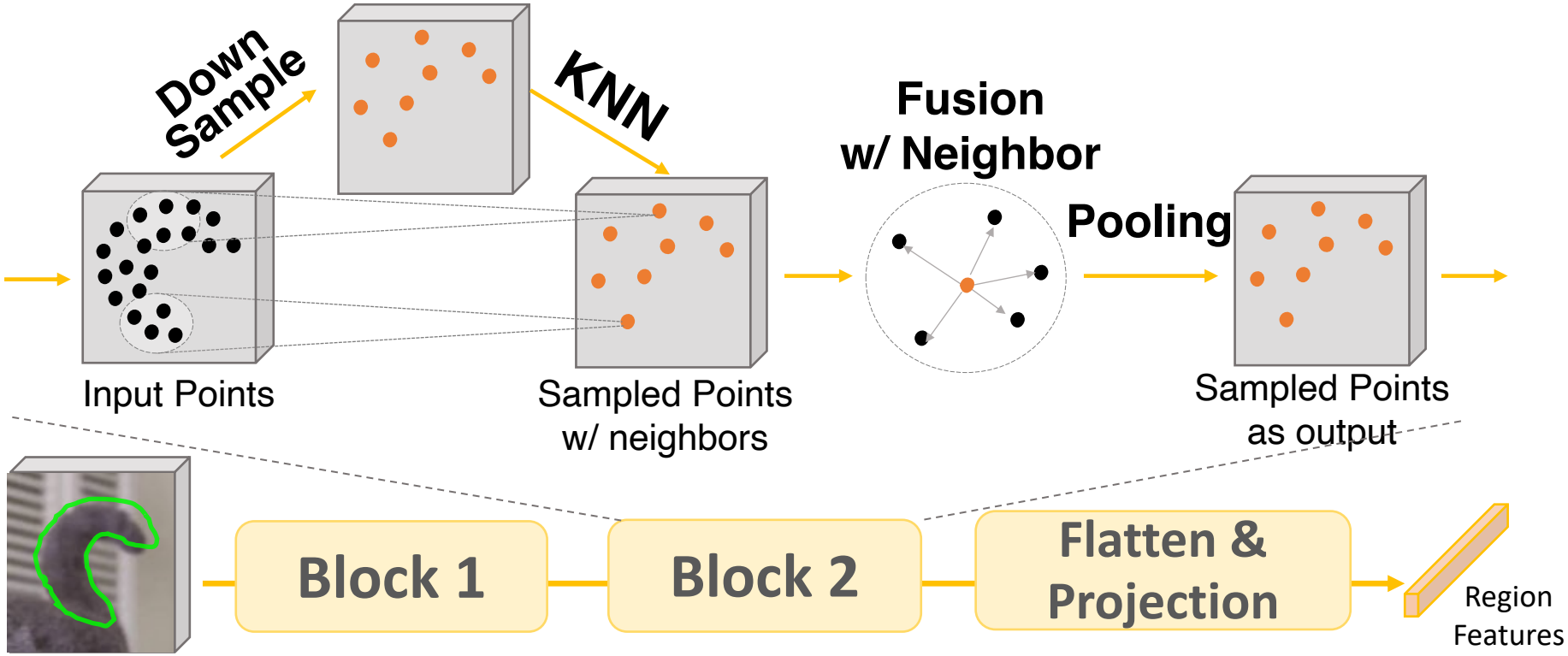
Ferret Model Structure

- Model:
  - Image Encoder: CLIP-ViT-L/14
  - LLM: Vicuna-V1.3
  - Proposed Spatial-Aware Visual Sampler
- Optimization:
  - Next Token Prediction.
  - Fix Image Encoder, Update Others.



Spatial-aware Visual Sampler

- Sample 512 points inside the region from feature maps.
- Go through 2 blocks. Inside each one,
  - Down-sample the number of points.
  - Find K-Nearest Neighbors
  - Fuse neighbor features and then pooling
- Flatten final 32 points and linearly project their features to LLM's embedding space.



Ferret: a MLLM w/ Spatial Understanding

GRIT: A Ground-and-Refer Instruction-Tuning Dataset

- Hierarchical; Unified Format, Instruction-Following; Robust.
- Dataset Size: 1.1M

Object:

Visual Genome
RefCOCOs
Flickr30k

Input What is the class of object [region0]?

Output It's a corgi.

Input Where is the corgi?

Output corgi [box0].

Relationship:

Visual Genome

Input Can you figure out the geometric relation of object [region1] and object [region2]?

Output Box on the floor.

Region:

Visual Genome

Input Describe the area [region3] in short phrase.

Output In short, a dog sit near a sofa.

Detail + Reasoning:

GPT Generated data

Input Do you know when the object [region4] was invented?

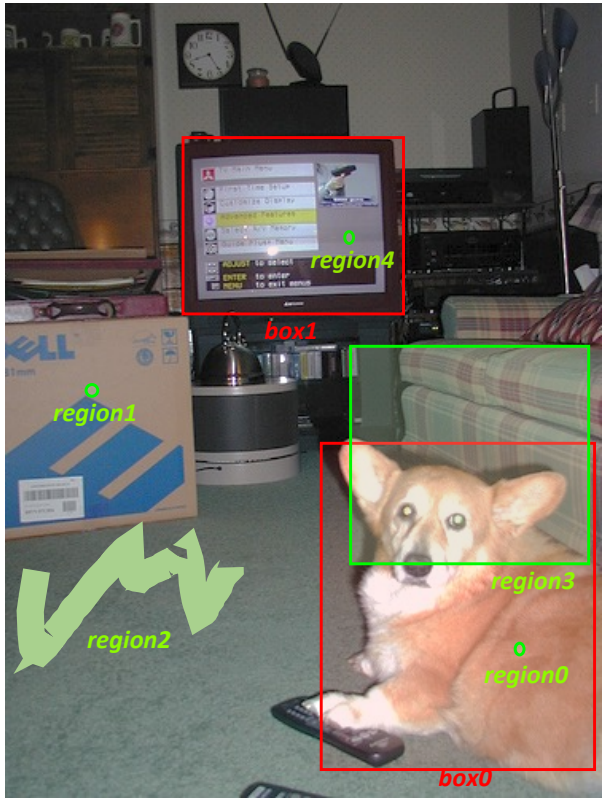
Output The invention of TV [box1] is credit to John Logie Baird. He invented in the early 1920s.

Robustness:

Spatial Negative Mining

Input Is there a cat in the image?

Output No, but there is a dog [box0] in the image.



Data Collection – GPT-Assisted Instructions

Two types of generated data: Single-round and Multi-round Conversations.

- Source Data:
  - Overlap of VG dataset and MSCOCO
  - Visual Contexts:
    Object, Relationships, Region Descriptions from VG annotation.
    Global Captions from MSCOCO annotation.

- Details of GPT few-shot prompting
  - Each few-shot examples has its visual context and human written conversation.
  - 3 few-shot examples for each type of data.
  - Generate-then-Revise:
    - Use ChatGPT to generate new conversations
    - Use GPT4 to revise the generated conversations.

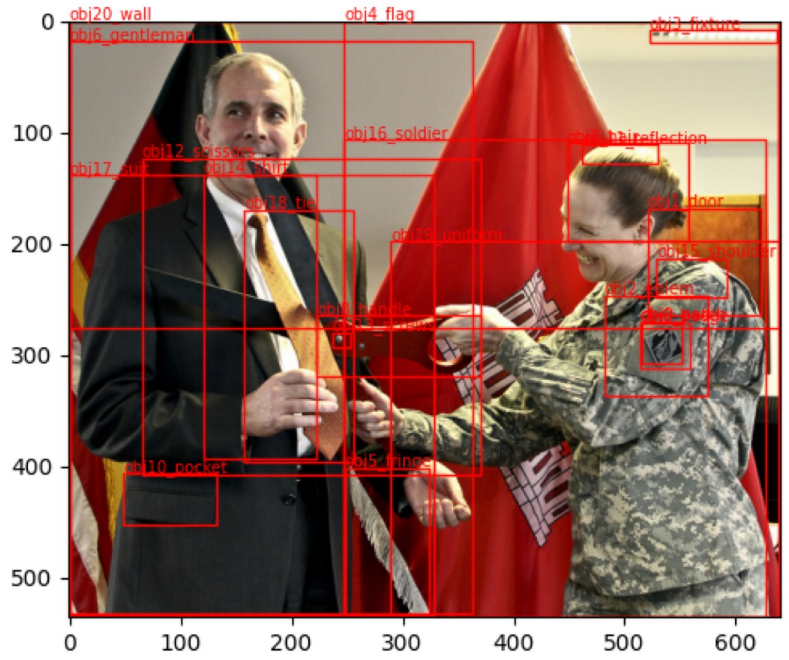
- Additionally, we apply GLIPv2 to LLaVA's instruction data to ground the objects in their responses. And append the location after the corresponding objects.

Total Data: 34k + 158k

Ferret: a MLLM w/ Spatial Understanding

Data Collection – GPT-Assisted Instructions

A few-shot Example:



Context:

```
Objects:
Object 0 : badge at [0.802, 0.505, 0.872, 0.581].
Object 1 : door at [0.814, 0.313, 0.972, 0.493].
Object 2 : eblem at [0.752, 0.459, 0.898, 0.629].
Object 3 : fixture at [0.816, 0.014, 0.996, 0.036].
Object 4 : flag at [0.386, 0.000, 0.998, 0.998].
Object 5 : fringe at [0.388, 0.749, 0.506, 1.000].
Object 6 : gentleman at [0.000, 0.033, 0.568, 0.993].
Object 7 : hair at [0.700, 0.203, 0.870, 0.368].
Object 8 : handle at [0.348, 0.493, 0.578, 0.596].
Object 9 : patch at [0.804, 0.502, 0.862, 0.574].
Object 10 : pocket at [0.076, 0.758, 0.208, 0.844].
Object 11 : reflection at [0.722, 0.206, 0.828, 0.239].
Object 12 : scissors at [0.102, 0.230, 0.578, 0.761].
Object 13 : screws at [0.370, 0.522, 0.396, 0.545].
Object 14 : shirt at [0.188, 0.256, 0.348, 0.734].
Object 15 : shoulder at [0.826, 0.397, 0.926, 0.464].
Object 16 : soldier at [0.386, 0.199, 0.980, 1.000].
Object 17 : suit at [0.000, 0.258, 0.512, 0.995].
Object 18 : tie at [0.246, 0.316, 0.400, 0.739].
Object 19 : uniform at [0.452, 0.368, 1.000, 0.998].
Object 20 : wall at [0.000, 0.000, 0.998, 0.514].
Relationships:
object 5 : fringe -> on -> object 4 : flag.
object 2 : eblem -> on -> object 9 : patch.
object 11 : reflection -> on -> object 7 : hair.
object 0 : badge -> on -> object 15 : shoulder.
object 10 : pocket -> on -> object 17 : suit.
Region Description:
Region Description at [0.038, 0.029, 0.508, 0.995] : A Grinning gentleman about to get his tie cut.
Region Description at [0.864, 0.297, 0.980, 0.426] : Possible wooden door in the background.
Global Caption:
a female in military uniform cutting a businessman's neck tie
A woman in a soldier's uniform pretends to cut the tie of a man using an enormous pair of scissors.
A funny picture of a woman holding a large pair of scissors attempting to cut a mans tie.
A woman soldier holding up a pair of giant scissors to someone's tie.
A man in a suit and a woman in military clothes.
```

Annotation:

Question:
What is unusual about the region in [0.102, 0.230, 0.578, 0.761]?
===
Answer:
In the specified region, there is an intriguing sight: a pair of oversized scissors positioned near a man's tie, accompanied by a woman in military attire. The two of them share a hearty laugh, creating an unexpected contrast. The atmosphere appears celebratory, resembling a ribbon-cutting ceremony with flags in the background. However, rather than cutting a ribbon, the scissors are directed toward the man's tie, adding an unexpected twist to the situation.

Evaluation:

- Conventional Abilities:
  - Referring
  - Grounding
  - LLaVA

- Ferret-Bench: A new benchmark to evaluate tasks jointly requiring referring/grounding, semantics, knowledge, and reasoning.

- Object Hallucination

Evaluation – Ferret-Bench

A new benchmark to evaluate tasks jointly requiring referring/grounding, semantics, knowledge, and reasoning.

- Three New Tasks:
  - **Referring Description**: Describe a referred region based on its interaction with surrounding objects.
  - **Referring Reasoning**: Reason on top of one or more referred regions correctly
  - **Grounding in Conversation**: Reason correctly and accurately ground/localize the objects/regions necessary for the reasoning.
- Each task have 40 annotations, generated by GPT-4 in a similar way to preparing GPT-Assisted Data in GRIT, verified by human.
- Evaluated by GPT-4.

Ferret: a MLLM w/ Spatial Understanding

Evaluation – Ferret-Bench

Quantitative Comparison:

	Ferret-Bench			
	Referring Description	Referring Reasoning	Grounding in Conversation	Avg.
LLaVA[8]	41.4	31.7	28.8	34.0
Kosmos-2	51.8	33.7	48.4	44.6
Shikra-7B	46.0	41.6	50.1	45.9
Ferret-7B	68.7	67.3	57.5	64.5
Ferret-13B	**70.6**	**68.7**	**59.7**	**66.3**

An Example:

Question:
What is the purpose of the object [region0]?

Ground Truth:
The object is a bottle. Bottles are typically used to hold and store different types of liquids. In this context, it might be used to store a beverage for the meal.

LLaVA:
A knife and a fork, indicating that it is ready for a meal.

Kosmos-2 :
The purpose of the object is to attract the birds to the table.

Shikra :
The answer is keep birds away.

Ferret (Ours) :
The object is a bottle, generally used to store and dispense liquids like water, juice, or other beverages.

Ablation – Mutual Benefit of Referring and Grounding

Ablation study on the mutual benefit of grounding data and referring data.

Model	Referring (LVIS)		Grounding
	Point	Box	Flickr30k
Ferret	**67.9**	**79.4**	**80.4**
w/o Grounding data	65.4	75.6	×
w/o Referring data	×	×	79.8

Ablation – Spatial-aware Visual Sampler

Ablation study on the effectiveness of the proposed spatial-aware visual sampler.

Module	Referring (LVIS)		
	Point	Box	Free-form
Spatial-aware Visual Sampler	**67.9**	**79.4**	**69.8**
Visual Sampler in SEEM	67.1	77.2	68.9