# Current VLMS

Philipp Krähenbühl, UT Austin

### Herret

- VLM with arbitrary visual grounding
- Region input and output
  - Special box token
  - Natural language coordinates
- Spatial aware Visual Sampler (Region features)
  - Region -> k point features

Ferret: Refer and Ground Anything Anywhere at Any Granularity; You etal 2023



### Ferret

- Trained on
  - GRIT (Ground-and-Refer Instruction-Tuning Dataset)
  - Public datasets

Task	Three randomly chosen examples from many.
Referring-Object	What is the class of the object <location> within the image?         Classify object <location> in the image.         Identify the object <location> in the image.</location></location></location>
Referring-Relation	What does <object1> <location1> do to <object2> <location2> of the image?What is the physical relation between <object1> <location1> and <object2> <location2>?Can you figure out the geometric relation of the <object1> <location1> and <object2> <location2>?</location2></object2></location1></object1></location2></object2></location1></object1></location2></object2></location1></object1>
Referring-Region	Describe the region <location> in a short phrase.         What is in the region <location>? Describe in a phrase.         Capture in a phrase: what's near region <location> in the picture?</location></location></location>
REC.	Where is <object> in the image?         What are the coordinates for the given <object> in the image?         Given the image, could you please tell me where is <object></object></object></object>
Phrase Grounding	What are the locations of <objects>?         Could you provide me with the exact locations of <objects>?         Please indicate the positions of <objects> in the image?</objects></objects></objects>
Object Detection (O365)	Detect all objects among <class> in the image.         Perform object detection given the image within <class>.         Given the image and set <class>, identify all the objects that belong to the set.</class></class></class>
Grounded Captioning	What is this photo about? Use concise language.Describe the overall picture in just a few words.What do you see happening in this image? Provide the answer in short.
Object Hallucination	Is there a <object> in the image?         Are there <object> in the image?         Please tell me whether <object> exists in the image?</object></object></object>

Ferret: Refer and Ground Anything Anywhere at Any Granularity; You etal 2023

<b>Object:</b> Visual Genome RefCOCOs Flickr30k	Input What is the class of object [region0]? Input Where is the corgi?	Output It's a corgi. Output corgi [box0].	$\Leftrightarrow$	
<b>Relationship:</b> Visual Genome	Input Can you figure out the geometric relation of object [region1] and object [region2]?	Output Box on the floor.	$\Leftrightarrow$	box1
<b>Region:</b> Visual Genome	Describe the area [region3] in short phrase.	Output In short, a dog sit near a sofa.	$\Leftrightarrow$	
Detail + Reasoning: GPT Generated data	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.		region2
<b>Robustness:</b> Spatial Negative Mining	Input Is there a cat in the image?	No, but there is a dog [box0] in the image.	$\Leftrightarrow$	Бохо

### **Objects**

Object 0: chair at [0.596, 0.637, 0.698, 0.997].

### **Relationships**

Object 11 : frame  $\rightarrow$  with  $\rightarrow$  Object 9 : photo

### **Region Descriptions**

Region Description at [0.560, 0.466, 0.600, 0.529] : a white picture frame with a black and white photo on it.

### **Global Caption**

White chairs sit around a polished wood dining table while a sectional soft sits in the background.

### Conversation

...

### Question:

What is the object [0.212, 0.539, 0.718, 0.820] made of? ===

### Answer:

The dining table over there is crafted from a tree trunk.

### \_\_\_ Question:

Where would be a good place to have my dinner? Point out the locations in image.

### === Answer:

I suggest having dinner at the table [0.214, 0.541, 0.720, 0.997], possibly sitting on the white chair [0.264, 0.717, 0.524, 0.995]. Also, enjoy the wine [0.242, 0.644, 0.288, 0.682].

### Question:

In terms of comfort for resting, which object is better - the one [0.830, 0.431, 0.978, 0.596] or the one [0.264, 0.717, 0.524, 0.995]?

### ===

.....

===

Answer Among the two, the couch located [0.830, 0.431, 0.978, 0.596] would be more comfortable for resting. The object [0.264, 0.717, 0.524, 0.995] is a dining chair, which is typically less comfortable for extended relaxation.

### Ferret

- Good results on public datasets
  - Referring expression
  - Grounding
- Ferret-Bench
  - Referring description, Referring reasoning, Grounding in Conversation
  - 40 COCO images, LLaVA-style datagen

Ferret: Refer and Ground Anything Anywhere at Any Granularity; You etal 2023

Models	R	efCOC	0	R	efCOC	D+	RefC	OCOg	Flickr3	0k Entities
Widdels	val	testA	testB	val	testA	testB	val	test	val	test
MAttNet (Yu et al., 2018)	76.40	80.43	69.28	64.93	70.26	56.00	66.67	67.01	_	_
OFA-L (Wang et al., 2022b)	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	_	—
TransVG (Deng et al., 2021)	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	_	79.10
UNITER (Chen et al., 2020)	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	_	_
VILLA (Gan et al., 2020)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	_	_
UniTAB (Yang et al., 2022)	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	78.76	79.58
MDETR (Kamath et al., 2021)	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	82.3*	83.8*
Shikra-7B (Chen et al., 2023b)	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	75.84	76.54
Ferret-7B	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	80.39	82.21
Shikra-13B (Chen et al., 2023b)	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	77.41	78.44
Ferret-13B	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34	81.13	84.76

Referring Object Classification (LVIS) – box

Referring Object Classification (LVIS) – free form shape



### **Question:**

What is the purpose of the object [region0] on the bike?

### **Ground Truth:**

The object is a shock absorber on the bike. Its purpose is to absorb or dampen shock impulses. It does this by converting the kinetic energy of the shock into another form of energy which is then dissipated. In the context of the motorcycle, it is particularly useful in providing comfort and stability for the rider, especially when travelling over uneven or rough terrains.



### LLaVA:

A black and silver Kawasaki motorcycle which is likely used for transportation, commuting, leisure rides, or long-distance travel.

### Kosmos-2:

The purpose of the object on the motorcycle is to store and protect the motorcycle's fuel tank. Shikra :

### The answer is leg.

Ferret (Ours):

The object is a shock absorber. This component is essential for the bike's stability, providing a smoother ride by soaking up the impact from bumps and uneven surfaces. It significantly improves the overall performance and comfort of the bike, enabling the rider to maintain control and stability even on rough terrains.

## Ferret UI

- VLM for UI understanding
- LLaVA 1.5-style multi-res (2 sub-images)
- Different datasets
  - RICO
  - Auto-labeled UI detector

Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs; You etal 2024



## Ferret UI

- Elementary tasks
  - Single round QA (grounded)
- Advanced tasks
  - Higher level reasoning





0.969] 0.969]



etections:	norma	lized co	ordinat	es
Text Shortcuts	[0.001,	0.068,	0.283,	0.110]
. Text Select +	[0.718,	0.070,	0.966,	0.108]
Text All Shortcuts	[0.019,	0.116,	0.584,	0.168]
TextField Search, microphone	[0.024,	0.168,	0.964,	0.227]
Text Starter Shortcuts	[0.030,	0.268,	0.548,	0.308]
Icon add	[0.891,	0.270,	0.973,	0.307]
Button more, Text Last Image	[0.042,	0.317,	0.489,	0.462]
Button more, Recognize Music	[0.516,	0.318,	0.959,	0.461]
Button more, copy, Make GIF	[0.041,	0.476,	0.489,	0.617]
Button layers, more, What's a, shortcut?	[0.518,	0.475,	0.959,	0.618]
0 Icon book	[0.030,	0.640,	0.126,	0.684]
1 Text Books >	[0.126,	0.643,	0.350,	0.684]
2 Button book, Read Current, Play Current	[0.000,	0.693,	1.000,	0.827]
3 Icon camera	[0.029,	0.845,	0.129,	0.891]
4 Text Camera >	[0.129,	0.849,	0.388,	0.889]
5 Tab Shortcuts	[0.003,	0.909,	0.327,	0.969]
6 Tab Automation	[0.327,	0.909,	0.680,	0.969]
7 Tab Gallery	[0.680,	0.909,	0.985,	0.969]
ared Prompt				
, are an AT visual accistance that can anal	vze mobi	la scra	one Vo	
reive information describing a screen when		T widoe	t detec	tion
represented using label text and boundin	a box co	ordinat		rated
represented using tabet, text, and boundin	ig box co	orutilat	es sepa	aceu

ou will ction arated by tab. Detections are separated with a newline. Bounding box coordinates are represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.

### Task Prompt

Using the provided texts and coordinates, design guestion and answer pairs simulating the interactions between a user and system. Conversations should focus on potential user actions on the screen (vs perception). When creating conversations that refer to a specific widget, append bounding box coordinates directly after mentioning the widget, not at the end of the sentence. Only generate questions and answers that can be answered with given information, do not use background information.

### One-shot example (For conversation tasks) User: TextField search [0.019, 0.050, 0.909, 0.117] Icon book [0.909, 0.067, 0.982, 0.109]

SegmentedControl FITNESS [0.017, 0.117, 0.278, 0.166] ..

Assistant: User: How do I search for a new product? System: You can use the TextField [0.019, 0.050, 0.909, 0.117] to search.

### Advanced tasks overview

- Bounding box coordinates in detections are normalized
- Concatenate shared prompt and task-specific prompt
- For conversation tasks, add one-shot example
- Provide detections, prompt, and example to GPT-4





**User:** I would like to create a new shortcut. *System:* You can click on the "Select +" Text [0.718, 0.070, 0.966, 0.108] to create a new shortcut.

**User:** How do I search for a specific shortcut? **System:** You can use the "Search" TextField [0.024, 0.168, 0.548, 0.308] at the top of the screen.

User: I want to read some books. System: Sure, you can go to Books by pressing the "book, Read Current, Play Current" button [0.000, 0.693, 1.000, 0.827] to get more about books.

## Ferret UI

Spotlight [30] Ferret [53] Ferret-UI-base Ferret-UI-anyres

GPT-4V [1]

- Mixed results
  - Public benchmarks (screen2word, widget captions, taperception [can you interact])
  - Elementary tasks (VQA / grounding)
  - Advanced Tasks (interactions)

Publ	ic Benchr	nark		Elementa	Advanced Tasks			
S2W	WiC	TaP	Ref-i	Ref-A	Grd-i	Grd-A	iPhone	Android
106.7 17.6 113.4 <b>115.6</b>	141.8 1.2 <b>142.0</b> 140.3	<b>88.4</b> 46.2 78.4 72.9	- 13.3 80.5 <b>82.4</b>	- 13.9 <b>82.4</b> <b>82.4</b>	- 8.6 79.4 <b>81.4</b>	- 12.9 83.5 <b>83.8</b>	-20.0 73.4 93.9	-20.7 80.5 71.7
34.8	23.5	47.6	61.3	37.7	70.3	4.7	114.3	128.2

### Ferret-v2

- Any resolution (LLaVA 1.5/1.6)
  - Original aspect ratio
  - Fixed max(width, height)
  - Learned encoder
- DINOv2 encoder
- High-res pre-training



(a). Qualitative Example

Referring:

Grounding:

"ABTO"

(b). Quantitative Result





Ferret-v2: An Improved Baseline for Referring and Grounding with Large Language; Zhang etal 2024

### Ferret-v2

• Strong results

Ferret-v2: An Improved Baseline for Referring and Grounding with Large Language; Zhang etal 2024

Models		RefCOCC	)	F	RefCOCO	+	RefC	OCOg	Flickr30	)k Entitie
Wodels	val	testA	testB	val	testA	testB	val	test	val	test
MAttNet (Yu et al., 2018)	76.40	80.43	69.28	64.93	70.26	56.00	66.67	67.01	–	_
OFA-L (Wang et al., 2022)	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	_	_
UNITER (Chen et al., 2020)	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	_	_
VILLA (Gan et al., 2020)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	_	_
UniTAB (Yang et al., 2022)	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	78.76	79.58
MDETR (Kamath et al., 2021)	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	82.3*	83.8*
G-DINO-L (Liu et al., 2023c)	90.56*	93.19*	88.24*	82.75*	88.95*	75.92*	86.13*	87.02*	_	_
Shikra-7B (Chen et al., 2023b)	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	75.84	76.54
MiniGPT-v2-7B (Chen et al., 2023a)	88.06	91.29	84.30	79.58	85.52	73.32	84.19	84.31	_	_
Qwen-VL-7B (Bai et al., 2023)	88.55	92.27	84.51	82.82	88.59	76.79	85.96	86.32	_	_
SPHINX-2k (Lin et al., 2023)	91.10	92.88	87.07	85.51	90.62	80.45	88.07	88.65	_	_
LLaVA-G (Zhang et al., 2023a)	89.16	_	-	81.68	-	_	84.82	-	83.03	83.62
VistaLLM (Pramanick et al., 2023)	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4	_	_
Ferret-7B (You et al., 2023)	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	80.39	82.21
Ferret-v2-7B (Ours)	92.79	94.68	88.69	87.35	92.75	79.3	89.42	89.27	85.52	85.83
Shikra-13B (Chen et al., 2023b)	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	77.41	78.44
Griffon v2 (Zhan et al., 2024)	89.6	91.8	86.5	81.9	85.5	76.2	85.9	86.0	_	84.8
CogVLM-Grounding-17B (Wang et al., 2023a)	92.76	94.75	88.99	88.68	92.91	83.39	89.75	90.79	_	_
Ferret-13B (You et al., 2023)	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34	81.13	84.70
Ferret-v2-13B (Ours)	92.64	94.95	88.86	87.39	92.05	81.36	89.43	89.99	85.33	86.25

Method	VQA <sup>v2</sup>	GQA	$VQA^{T}$	POPE	MME <sup>P</sup>	SEED	LLaVA <sup>C</sup>	LLaVA <sup>W</sup>	MM-Vet	Obj-Hal↓
BLIP-2-13B	41.0	41	42.5	85.3	1293.8	46.4	_	38.1	22.4	_
InstructBLIP-7B	_	49.2	50.1	_	_	53.4	_	60.9	26.2	_
IDEFICS-9B	50.9	38.4	25.9	_	_	_	_	_	_	_
Qwen-VL-7B	78.8*	59.3*	63.8	_	_	56.3	_	_	_	_
Qwen-VL-Chat-7B	78.2*	57.5*	61.5	_	1487.5	58.2	_	_	_	43.8/23.0
LLaVA-1.5-7B	78.5*	62.0*	58.2	85.9	1510.7	58.6	82.7	63.4	30.5	46.3/22.6
Ferret-v2-7B (Ours)	81.5*	<b>64.7</b> *	61.7	87.8	1510.3	58.7	89.1	67.7	34.9	23.8/14.7
Ferret-v2-7B (Ours)	81.5*	<b>64.7</b> * 49.5	<b>61.7</b> 50.7	<b>87.8</b> 78.9	<b>1510.3</b> 1212.8	<b>58.7</b>	<b>89.1</b>	<b>67.7</b> 58.2	<b>34.9</b> 25.6	23.8/14.7
Ferret-v2-7B (Ours) InstructBLIP-13B Shikra-13B	<b>81.5</b> *	<b>64.7</b> * 49.5 –	<b>61.7</b> 50.7	<b>87.8</b> 78.9 -	<b>1510.3</b> 1212.8 -	58.7 	<b>89.1</b> - -	<b>67.7</b> 58.2 -	<b>34.9</b> 25.6	23.8/14.7 _ _
Ferret-v2-7B (Ours) InstructBLIP-13B Shikra-13B IDEFICS-80B	<b>81.5</b> *	<b>64.7</b> * 49.5 - 45.2	<b>61.7</b> 50.7 - 30.9	<b>87.8</b> 78.9 	<b>1510.3</b> 1212.8 - -	58.7 - - -	<b>89.1</b> - -	<b>67.7</b> 58.2 –	<b>34.9</b> 25.6 –	23.8/14.7 - -
Ferret-v2-7B (Ours) InstructBLIP-13B Shikra-13B IDEFICS-80B LLaVA-1.5-13B	<b>81.5</b> * 77.4* 60.0 80.0*	<b>64.7</b> * 49.5 - 45.2 63.3*	<b>61.7</b> 50.7 - 30.9 61.3	<b>87.8</b> 78.9 - 85.9	<b>1510.3</b> 1212.8 - - 1531.3	<b>58.7</b> - - 61.6	<b>89.1</b> - - 83.4	<b>67.7</b> 58.2 - 70.7	<b>34.9</b> 25.6 - 35.4	23.8/14.7 - - -
Ferret-v2-7B (Ours) InstructBLIP-13B Shikra-13B IDEFICS-80B LLaVA-1.5-13B LLaVA-1.5-13B-HD	<b>81.5</b> * 77.4* 60.0 80.0* <b>81.8</b> *	<b>64.7</b> * 49.5 - 45.2 63.3* 64.7*	<b>61.7</b> 50.7 - 30.9 61.3 <b>62.5</b>	<b>87.8</b> 78.9 - 85.9 86.3	<b>1510.3</b> 1212.8 - - 1531.3 1500.1	<b>58.7</b> 61.6 <b>62.6</b>	<b>89.1</b> 83.4 -	<b>67.7</b> 58.2 - 70.7 <b>72.0</b>	<b>34.9</b> 25.6  35.4 <b>39.4</b>	23.8/14.7 - - - - -

# CubeIIM

dataset	images	$labels_{2D}$	$labels_{3I}$	o captions	# QAs s	tage 1 s	tage 2	Reference Image		Random point→2D BoxOut	iput
LLaVA data [35]	80K	$\checkmark$	X		158K	1	0.5		ding	containing the 2D point [x', y'].	[x <sub>1</sub> , y <sub>1</sub> , x <sub>2</sub> , y <sub>2</sub> ]
refCOCO/ $+/g$ [67]	67K	$\checkmark$	×	$\checkmark$	154K	1	0.5		2D roun	Text→2D CenterOut	tput
GRIT (subset) [43]	4M	$\checkmark$	×	$\checkmark$	$6.9 \mathrm{M}$	1	0.3		G	Provide the 2D center coordinate of the region this sentence describe: Forklift truck in front of the garage.	[x <sub>c</sub> , y <sub>c</sub> ]
AS (filtered) $[56]$	$3.7 \mathrm{M}$	$\checkmark$	×	$\checkmark$	13.2M	1	0.5				
COCO [33]	118K	$\checkmark$	×	X	860K	1	0.5		ing	_Text→3D BoxOut	tput
Objects $365 \ [47]$	600K	$\checkmark$	×	X	$25.4\mathrm{M}$	0.3	0.2		3D Bund	Provide 3D box coordinates of the region this sentence [x, y,	, z, w, h, l, r <sub>1</sub> , r <sub>2</sub> , r <sub>3</sub> ]
SUN-RGBD $[49]$	5K	$\checkmark$	$\checkmark$	X	41K	$1^{\star}$	5		Gro	describes: Forklift truck in front of the garage.	
Hypersim $[45]$	$67 \mathrm{K}$	$\checkmark$	$\checkmark$	X	2M	$1^{\star}$	5			_2D Box→3D Box	tout
ArkitScenes [6]	53K	$\checkmark$	$\checkmark$	X	420K	1*	5		o B	Provide 3D bounding box coordinates of the	zwhlr.rorol
Objectron $[2]$	37K	$\checkmark$	$\checkmark$	X	43K	1*	5	Object Caption Object 3D Box	o 3[ onir	following region in 2D: $[x_1, y_1, x_2, y_2]$ .	2, 0, 1, 1, 1, 1, 2, 3]
KITTI [22]	$4\mathrm{K}$	$\checkmark$	$\checkmark$	X	25K	1*	5	Forklift truck in front	2D t Reas	2D Center→DepthOut	tput
NuScenes [8]	40K	$\checkmark$	$\checkmark$	X	$1.1\mathrm{M}$	1*	2	of the garage		Provide depth of the object in the following region in	z
Lyft $[25]$	105K	$\checkmark$	$\checkmark$	X	723K	0	2	Object 2D Box		2D. [x <sub>1</sub> , y <sub>1</sub> , x <sub>2</sub> , y <sub>2</sub> ].	
Argoverse2 [61]	79K	$\checkmark$	$\checkmark$	X	915K	0	4			_3D Box→Text	tout
Waymo [51]	680K	$\checkmark$	$\checkmark$	×	$5.1\mathrm{M}$	0	0.4		БĽ	Provide a short caption of the 3D bounding box [x, y, Fork	klift truck in front
Total	$9.6 \mathrm{M}$	$\checkmark$	$\checkmark$	$\checkmark$	$40.9 \mathrm{M}$	0.87	0.52	Object 2D Center —	tionii	z, w, h, l, r <sub>1</sub> , r <sub>2</sub> , r <sub>3</sub> ] of th	ne garage
								[x <sub>c</sub> , y <sub>c</sub> ]	Cap	2D Box $\rightarrow$ Text Provide a short caption of the 2D bounding box [x <sub>1</sub> , y <sub>1</sub> , x <sub>2</sub> , y <sub>2</sub> ]. Out <sub>1</sub> Fork of the	put klift truck in front he garage

- Convert 2D and 3D vision annotations to text
  - Standardized / normalized
- Simple DINOv2 + Vicuna 7B model
- Large scale training
  - 2D + 3D detection data, grounding data, captioning, etc
- Visual Chain-of-Thought
- Specialist inputs

Language-Image Models with 3D Understanding, Cho et al 2024



### i. Visual Chain-of-Thought Prompting

### Text Prompt

Provide the 2D bounding box of the region this sentence describes: Pull up behind the guy wearing a white shirt. He is my uncle.

![](_page_9_Picture_15.jpeg)

Output (x1, y1, x2, y2) = (0.5, 0.5, 0.6, 0.75)

Text Prompt Provide the 3D bounding box of the 2D box (0.5, 0.5, 0.6, 0.75) that this sentence describes: Pull up behind the guy wearing a white shirt. He is my uncle

> Output (x, y, z, w, l, h, yaw) = (0.14, 0.65, 10.22, 0.89,1.85, 0.79, -1.70)

![](_page_9_Picture_19.jpeg)

### ii. Specialist Model Prompting

![](_page_9_Picture_21.jpeg)

Given the 3D bounding boxes detected by the specialized LiDAR-based detection models:  $(b_1, b_2, b_3, b_4, b_5)$ , provide the 3D bounding box of the region that this sentence describes: Stop in front of the man in the yellow hat.

![](_page_9_Picture_23.jpeg)

![](_page_9_Picture_24.jpeg)

## CubeLLM

- Good generalization
  - Where do I do X
  - Spatial understanding for free

Language-Image Models with 3D Understanding, Cho etal 2024

![](_page_10_Picture_5.jpeg)

## CubeLLM

- Good results
  - 2D and 3D grounding
  - 2D referring expression
  - VQA

Language-Image Models with 3D Understanding, Cho etal 2024

Madala	C:	R	efCO	CO	Re	efCOC	O+	RefC	OCOg	
Models	Size	val	testA	testB	val	testA	testB	val	test	Avg.
Specialist										
MAttNet [66]		76.4	80.4	69.3	64.9	70.3	56.0	66.7	67.0	68.9
OFA-L [55]		80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6	72.7
TransVG $[16]$		81.0	82.7	78.4	64.8	70.7	56.9	68.7	67.7	71.4
UNITER [11]		81.4	87.0	74.2	75.9	81.5	66.7	74.0	68.7	76.2
VILLA [20]		82.4	87.5	74.8	76.2	81.5	66.8	76.2	76.7	77.8
UniTAB [63]		86.3	88.8	80.6	78.7	83.2	69.5	80.0	80.0	80.6
MDETR $[27]$		86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9	81.8
Generalist										
LLaVA-1.5 [34]	$7\mathrm{B}$	75.6	82.1	66.9	65.5	76.2	53.9	68.9	69.1	69.8
VisionLLM-H [57]	$7\mathrm{B}$	86.7	-	-	-	-	-	-	-	-
Shikra [10]	$7\mathrm{B}$	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2	82.9
Ferret [65]	$7\mathrm{B}$	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8	83.9
MiniGPT-v2 [9]	$7\mathrm{B}$	88.7	91.7	85.3	80.0	85.1	74.5	84.4	84.7	83.8
LLaVA-G $[68]$	$7\mathrm{B}$	89.2	-	-	81.7	-	-	84.8	-	-
Qwen-VL $[5]$	$7\mathrm{B}$	88.6	92.3	84.5	82.8	88.6	76.8	86.0	86.3	85.7
Cube-LLM	7B	90.9	92.6	87.9	83.9	89.2	77.4	86.6	87.2	87.0

		~	~			
Method	Input	$\begin{array}{c} 2D\\ AP_{0.5} \end{array}$	BI AP <sub>A</sub>	$EV$ $AP_B$	3 AP <sub>A</sub>	D AP <sub>B</sub>
2D Specialist						
Talk2Car-2D [18]	$\mathbf{C}$	50.5	-	-	-	-
VL-Bert [50]	$\mathbf{C}$	63.1	-	-	-	-
Vilbert [36]	$\mathbf{C}$	68.9	-	-	-	-
CMRT [38]	$\mathbf{C}$	69.1	-	-	-	-
Stacked VLBert [14]	$\mathbf{C}$	71.0	-	-	-	-
FA [17]	С	73.5	-	-	-	-
Cube-LLM (zero-shot)	С	46.3	32.0	19.5	22.3	9.8
Cube-LLM	С	79.2	46.3	30.1	34.7	18.2
3D Specialist						
Talk2Car-3D [18]	L + C	-	30.6	24.4	27.9	19.1
MSSG [12]	L + C	-	50.1	35.7	45.4	23.7
Cube-LLM $^{\dagger}$	L + C	76.3	71.4	61.2	64.1	39.8
Т		0				

Talk2Car

Method	Pro-train Data	BI	EV	3	D
	1 IC-train Data	$AP_A^{BEV}$	$AP_B^{BEV}$	$AP_A^{3D}$	$AP_B^{3D}$
LLaVA-1.5 [34]	LLaVA data	33.2	16.3	21.7	7.7
Cube-LLM	LLaVA data	39.6	21.7	25.8	10.5
Cube-LLM	LV3D $(2D)$	50.5	31.2	32.5	17.3
Cube-LLM	LV3D	66.0	52.1	56.2	<b>40.5</b>

driveLM

Size	$VQA^{v2}$ [23]	GQA [26]	VizWiz [24]	$SQA^{I}$ [37]	POPE [32]
13B	41.0	41.0	19.6	61.0	85.3
$7\mathrm{B}$	-	49.2	34.5	60.5	-
13B	-	49.5	33.4	63.1	78.9
9B	50.9	38.4	35.5	-	-
13B	77.4	-	-	-	-
$7\mathrm{B}$	<b>78.8</b>	59.3	35.2	67.1	-
7B	78.2	57.5	38.9	68.2	-
7B	-	60.1	53.6	-	-
7B	78.5	62.0	50.0	66.8	85.9
7B	78.3	62.4	51.0	69.2	87.1
	Size 13B 7B 13B 9B 13B 7B 7B 7B 7B 7B 7B	Size VQA <sup>v2</sup> [23] 13B 41.0 7B - 13B - 9B 50.9 13B 77.4 7B 78.8 7B 78.2 7B - 7B 78.5 7B 78.3	Size $VQA^{v2}$ [23] GQA [26]13B41.07B-49.213B-9B50.938.413B77.4-7B <b>78.8</b> 59.37B78.257.57B-60.17B78.362.4	Size VQA <sup>v2</sup> [23] GQA [26] VizWiz [24]13B41.041.019.67B-49.234.513B-49.533.49B50.938.435.513B77.47B <b>78.8</b> 59.335.27B78.257.538.97B-60.1 <b>53.6</b> 7B78.562.050.07B78.3 <b>62.4</b> 51.0	Size VQA $^{v2}$ [23] GQA[26] VizWiz[24] SQA <sup>I</sup> [37]13B41.041.019.661.07B-49.234.560.513B-49.533.463.19B50.938.435.5-13B77.47B <b>78.8</b> 59.335.267.17B78.257.538.968.27B-60.1 <b>53.6</b> -7B78.562.050.066.87B78.3 <b>62.4</b> 51.0 <b>69.2</b>

# SpatialVLM

- Creating a large spatial reasoning dataset
  - Semantic filtering (CLIP)
    - 🙂 An iPhone photo of an indoor scene
    - 😕 A close of shot of a single object
    - ...
  - Region Captioning: RPN+NPS+FlexCap
    - Ambiguity removal: Remove close captions in CLIP-embedding space, or append qualifier
  - 3D context / canonicalization -> 3D objects
    - Human-alignment = rounding
  - 43 Question templates for pairs of objects
- Mixed results

SpatialVLM: Endowing Vision-Language Models with Surger Control of the Second S

![](_page_12_Figure_13.jpeg)

Is the plate positioned on the left or right side of The plate is to the the green and yellow riaht

between the microwave oven and the kitchen counter?	They are 1 me
How wide is the plate?	20 centimeters

20 centimeters

### Qualitative Q&A

Method	GPT-4V	LLaVA-1.5	InstructBLIP	PaLI	PaLM-E	PaLM 2-E	Ours
Accuracy	68.0%	71.3%	60.4%	60.7%	50.2%	50.4%	<b>75.2</b> %

### Qualitative Q&A

	GPT-4V	LLaVA-1.5	InstructBLIP	PaLI	PaLM-E	PaLM 2-E	Ours
Output numbers %	1.0%	20.9%	26.0%	52.0%	83.2%	88.8%	99.0%
In range [50, 200]%	0.0%	13.0%	7.9%	5.3%	23.7%	33.9%	37.2%

![](_page_12_Picture_20.jpeg)

nankinʻ

![](_page_12_Picture_21.jpeg)

### MM1

- LLaVA analysis paper
- Lot's of ablations
  - Good reference for design choices
- Mixed results

MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training; McKinzie etal 2024

![](_page_13_Figure_6.jpeg)

Model	VQA <sup>v2</sup>	<sup>2</sup> VQA <sup>1</sup>	SQA <sup>I</sup>	MMMU	MathV	/ MME <sup>P</sup>	$\mathbf{MME}^{\mathbf{C}}$	MMB	SEED	POPE	E LLaVA <sup>W</sup>	/ MM-Vet	Model Shot		Captioning			Visua	Visual Question Answ		
3B Model Comparison													Model	COCO NoCaps TextCaps			s VQAv2	VQAv2 TextVQA VizWiz			
MobileVLM [20] LLaVA-Phi [135]	71.4	$47.5 \\ 48.6$	$61.0 \\ 68.4$		_	$1288.9 \\ 1335.1$	_	$59.6 \\ 59.8$	-/- -/-	$84.9 \\ 85.0$	_	-28.9	MM1-3B Model Ce	ompari	sons	-					
Imp-v1 [99] TinyLLaVA [133] Bunny [42]	79.45 79.9 79.8	59.38 59.1	69.96 69.1 70.9	$\begin{vmatrix} -/-\\ -/-\\ 38 \ 2/33 \ 0 \end{vmatrix}$	_	1434.0 1464.9 1488.8	- - 280-3	66.49 66.9 68.6	_ _/_ 62.5./_	88.02 86.4 86.8	75.8	$33.1 \\ 32.0 \\ -$	Flamingo-3B [3]	$\begin{array}{c} 0^{\dagger} \\ 8 \end{array}$	73.0 90.6	_	_	$49.2 \\ 55.4$	$30.1 \\ 32.4$	$\begin{array}{c} 28.9\\ 38.4 \end{array}$	
Gemini Nano-2 [106] MM1-3B-Chat	67.5 82.0	65.9 71.9	- 69.4	32.6/- 33.9/33.7	30.6 32.0	- 1482.5	209.3 - 279.3	- 67.8	- 63.0/68.8	- 8 87.4	72.1	43.7	MM1-3B	0 8	73.5 <b>114.6</b>	55.6 <b>104.7</b>	63.3 88.8	46.2 63.6	29.4 <b>44.6</b>	15.6 <b>46.4</b>	
MMI-5B-MOE-Chat 82.5 72.9 76.1 [38.6/35.7 32.6 1469.4 303.1 70.8 63.9/69.4 87.6 76.8 42.2										42.2	MM1-7B Model Comparisons										
InstructBLIP-7B [24] Qwen-VL-Chat-7B [5]	78.2	$50.1 \\ 61.5$	60.5 68.2	-/- 35.9/32.9	25.3	-1487.5	360.7	$36.0 \\ 60.6$	53.4/- 58.2/65.4	-	60.9	26.2	IDEFICS-9B [58]	$0^{\dagger}$ 8	46.0* 97.0*	$36.8 \\ 86.8$	$25.4 \\ 63.2$	$50.9 \\ 56.4$	$25.9 \\ 27.5$	$\begin{array}{c} 35.5\\ 40.4 \end{array}$	
LLaVA-1.5-7B [74] ShareGPT4V-7B [15]	78.5 80.6	58.2 60.4	66.8 68.4	-/- -/-	_	1510.7 1567.4 1528.2	$316.1 \\ 376.4$	64.3 68.8	58.6/66.1 _/_	85.9	63.4 72.6	31.1 - 21.5	Flamingo-9B [3]	$\begin{array}{c} \\ 0^{\dagger} \\ 8 \end{array}$	79.4 99.0			51.8 58.0	31.8 33.6	28.8 39.4	
VILA-7B [71] SPHINX-Intern2 [36]	79.0 79.9 75.5	58.7 64.4 -	$68.2 \\ 70.4$	_/_ _/_ _/_	$^{-}_{-}$	1528.2 1531.3 1260.4	 294.6		60.0/- 61.1/- 68.8/-	80.0 85.5 86.9	$69.7 \\ 57.6$	$31.5 \\ 34.9 \\ 36.5$	Emu2-14B [105]	$\begin{array}{c} - & - & - \\ & 0^{\dagger} \\ & 8 \end{array}$				52.9 59.0		34.4 43.9	
LLaVA-NeXT-7B [75] MM1-7B-Chat MM1-7B-MoE-Chat	81.8 82.8 83.4	64.9 72.8 73.8	70.1 72.6 74.4	35.8/- 37.0/35.6 40.9/37.9	34.6 35.9 40.9	1519 1529.3 1597.4	332 328.9 394.6	67.4 72.3 72.7	-/70.2 64.0/69.9 65.5/70.9	86.53 86.6 87.8	81.6 81.5 84.7	43.9 42.1 45.2	MM1-7B	0 8	76.3 116.3	61.0 <b>106.6</b>	64.2 88.2	47.8 63.6	28.8 46.3	15.6 <b>45.3</b>	
30B Model Comparison								MM1-30B Model (	Compar	isons											
Emu2-Chat-37B [105] CogVLM-30B [114] LLaVA-NeXT-34B [75] MM1-30B-Chat	84.9 83.4 83.7 83.7	66.6 68.1 69.5 73.5	- 81.8 81.0	$\begin{vmatrix} 36.3/34.1\\ 32.1/30.1\\ 51.1/44.7\\ 44.7/40.3 \end{vmatrix}$	- 46.5 39.4 <sup>†</sup>	- 1631 1637.6	- $397$ $431.4$	-79.3	62.8/- - -/75.9 65.9/72.1	- 87.73 87.6	- $89.6$ $89.3$	$     48.5 \\     56.8 \\     57.4 \\     48.7 $	IDEFICS-80B [58]	$\begin{array}{c} 0^{\dagger} \\ 8 \\ 16 \end{array}$	91.8* 114.3* 116.6*	65.0 105.7 107.0	56.8 77.6 81.4	60.0 64.8 65.4	30.9 35.7 36.3	$\begin{array}{r} 36.0\\ 46.1\\ 48.3\end{array}$	
Gemini Pro [106] Gemini Ultra [106] GPT4V [1]	71.2 77.8 77.2	74.6 82.3 78.0		$\begin{vmatrix} 47.9 / - \\ 59.4 / - \\ 56.8 / 55.7 \end{vmatrix}$	$45.2 \\ 53.0 \\ 49.9$		436.79 - 517.14	73.6 - 75.8	-/70.7 - 67.3/69.1			$64.3 \\ - \\ 67.6$	Flamingo-80B [3]	$egin{array}{c} 0^\dagger \ 8 \ 16 \end{array}$	84.3 108.8 110.5		-	56.3 65.6 66.8	35.0 37.3 37.6	31.6 44.8 48.4	
														0	_	_	_	33.3	26.2	40.4	

- Custom image tokenizer
  - From Make-a-scene (VQ-GAN)
  - 512x512 images
    - 32x32 tokens
  - 8k codebook
- Single auto-regressive model
  - LLama2++
  - 4k context length

Chameleon: Mixed-Modal Early-Fusion Foundation Models; Chameleon Team 2024

![](_page_14_Figure_10.jpeg)

![](_page_14_Figure_11.jpeg)

(b) Mixed-Modal Generation

- Training Stage 1
  - 2.9T text-only tokens
  - 1.5T text-image (caption) tokens
  - 0.4T interleaved text-image tokens
- Training Stage 2
  - Stage 1 data (50% weight)
  - "Higher quality datasets"
- 0.8M 4M GPU hours
- Alignment
  - LLama2, CodeLLama, own data (sourced from external vendors)
  - Various safety data

Chameleon: Mixed-Modal Early-Fusion Foundation Models; Chameleon Team 2024

Prompt: Can you write a poem using rhyming couplets?Response: Sure, here's a short poem using rhyming couplets:

The sun sets in the sky, A beautiful sight to the eye. The stars come out to play, A magical sight to behold.

The wind whispers through the trees,

A gentle breeze that brings ease. The birds sing their sweet melodies, A soothing sound that calms the soul.

The world is full of wonder, A place of beauty and thunder. So let us cherish every moment, And bask in the beauty all around us.

**(a)** Text

![](_page_15_Picture_19.jpeg)

**Prompt:** Renaissance square of Stepanakert **Response:** 

![](_page_15_Picture_21.jpeg)

**(b)** Code

(c) Image Generation

![](_page_15_Picture_24.jpeg)

**Prompt:** Which elements contribute to this flower arrangement's distinctiveness, and what sets its surroundings apart?

**Response:** The image features an autumnal-themed floral arrangement in a classically designed vase. The arrangement stands out with its striking combination of red maple leaves, purple flowers, and pink-red roses. The vase's rosettes give it an antique design. In the background, blurry green vegetation is visible, creating a contrast to the autumnal theme of the bouquet. The red maple leaves, both in the arrangement and scattered around the vase, evoke autumn, while the green vegetation suggests a more summer-like setting. The outdoor setting is bright and sunny, adding to the contrast between the arrangement and its settings.

(d) Visual Chat

Gyumri is an urban municipal community and the second-largest city in Armenia, serving as the administrative center of Shirak Province in the northwestern part of the country.

![](_page_15_Picture_29.jpeg)

Archaeological excavations conducted throughout the Soviet period have shown that the area of modern-day Gyumri has been populated since at least the third millennium BC.

![](_page_15_Picture_31.jpeg)

(e) Interleaved Text/Image Generation

- Stability: Divergence is mid-to-late training
  - Vision and language tokens compete through norm growth

![](_page_16_Figure_3.jpeg)

Chameleon: Mixed-Modal Early-Fusion Foundation Models; Chameleon Team 2024

![](_page_16_Figure_5.jpeg)

![](_page_16_Figure_6.jpeg)

![](_page_16_Figure_7.jpeg)

![](_page_16_Picture_8.jpeg)

0k

- Human evaluation
  - Similar or better than GPT-4V and Gemini
  - XX+ use DALL-E image generation outputs

![](_page_17_Figure_5.jpeg)

**Prompt:** I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2–3 sentences.

![](_page_17_Picture_7.jpeg)

**Response:** Absolutely! Here are three birds with some unique features that you might find interesting:

Keel-Billed Toucan The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird: **<img>** 

**Puffin** Puffins are small, aquatic birds that are found in the northern hemisphere. They have distinctive black and white feathers, as well as a brightly colored beak that is used for attracting mates and defending their territory. Puffins are known for their distinctive "puffin" call, which sounds like a loud, guttural chuckle. They feed on fish, which they catch by diving into the water from the air. Puffins are colonial birds, and they breed in large colonies on rocky cliffs and islands. Here is an image of this adorable bird: **<img>** 

![](_page_17_Picture_11.jpeg)

![](_page_17_Picture_12.jpeg)

Golden Pheasant The Golden Pheasant is a brightly colored bird native to the forests of China. It is a member of the pheasant family and is known for its striking plumage, which features a golden-yellow body, red face, and green tail feathers. The Golden Pheasant is a ground-dwelling bird, and it feeds on a diet of insects, seeds, and fruits. It is a popular game bird, and is often hunted for its meat and feathers. In captivity, the Golden Pheasant is sometimes kept as an ornamental bird, due to its striking appearance. Here is an image of this beautiful bird: **<img>** 

I hope these images and descriptions have helped to brighten your day! If you have any other questions or requests, please don't hesitate to ask.

![](_page_17_Picture_15.jpeg)

**Prompt: <img>** While chameleons are masters of camouflage in the natural world, let's see how they might fare in a man-made setting.

1. Analyze the colors, patterns, and textures in this image and identify the most challenging spot for a chameleon to blend in.

2. Then generate an image of a chameleon with the optimal pattern to camouflage itself.

3. Then, rate and explain how successful each attempt is on a scale from 1-10.

**Response:** The image is of a stop sign in a forested area, with some sunlight filtering through the trees The sign has a red octagonal shape with a white border and the word "STOP" written in white letters In the background, there are green leaves and trees.

Challenging Spot for a Chameleon: The stop sign would be the most challenging spot for a chameleon to blend in due to its bright red color and bold white letters, which would stand out in the forested environment.

Optimal Pattern for Camouflage: To camouflage itself on the stop sign, a chameleon would need to have a pattern that would break up its outline and make it less conspicuous. A pattern with shades of brown and red would help the chameleon blend in more effectively.

![](_page_17_Picture_23.jpeg)

Additionally, a mottled or dappled pattern would help to scatter light and further reduce the chameleon's visibility. <img>

Rating and Explanation: A rating of 2/10 would be given for the chameleon's ability to blend in on the stop sign. While the chameleon's green and brown pattern helps to break up its outline, the bright red color of the stop sign will still make it stand out in the forested environment.

## Pali-Gemma

- StageO: Unimodal pretraining use existing off-theshelf components.
- Stage1: Multimodal pretraining with prefix
  - caption {lang} OC

answer en {question} question {lang} {English answer} detect {thing} ; {thing} ; ... segment {thing} ; {thing} ; ... caption <ymin><xmin><ymax><xmax>

- Stage2: Resolution increase 224 x 224 -> 448 x 448 -> 896 x 896
- Stage3: Transfer

PaliGemma: A versatile 3B VLM for transfer, Beyer etal 2024

![](_page_18_Figure_8.jpeg)

## Molmo and PixMo

### PixMo

### Captions

![](_page_19_Picture_3.jpeg)

![](_page_19_Figure_4.jpeg)

### AskModelAnything

![](_page_19_Picture_6.jpeg)

![](_page_19_Picture_7.jpeg)

![](_page_19_Figure_8.jpeg)

### Pointing

![](_page_19_Picture_10.jpeg)

### Synthetic

![](_page_19_Picture_12.jpeg)

### Molmo

### Fine-grained Understanding

![](_page_19_Picture_15.jpeg)

![](_page_19_Picture_16.jpeg)

The image captures a bustling scene at the Louvre Museum in Paris, France. Dominating the foreground is the iconic Louvre Pyramid, a large glass and metal structure that stands prominently in the center of the courtyard. The pyramid's intricate design allows light to filter through, illuminating the area beneath. Surrounding the pyramid, the courtyard is paved with stone and features a series of fountains, each with water jets shooting upwards, creating a dynamic and lively atmosphere. In the background, the grand Louvre Palace stretches across the frame. This historic building, with its ornate architecture, showcases a blend of yellow and brown hues, accented by numerous pillars and ...

### **User Interaction**

Where is this bus going?

![](_page_19_Picture_20.jpeg)

The bus is heading to Highland Park. This is clearly visible on the destination sign at the front of the bus, which displays "71B HIGHLAND PARK"

![](_page_19_Picture_22.jpeg)

No, you cannot park here at 2:00 PM on November 15th. The parking sign clearly shows that from 1 PM to 11 PM on November 15th, there is no parking allowed in this area...

### **Pointing and Counting**

![](_page_19_Picture_25.jpeg)

### **Visual Skills**

![](_page_19_Picture_27.jpeg)

## Molmo and PixMo

- New open dataset
- New pointing data
- Point outputs and in-context inputs

Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, Deitke etal 2024

![](_page_20_Figure_5.jpeg)

![](_page_20_Picture_6.jpeg)

## Vision language models Trend 1

### • Data matters

• As important, or more important than model design

### PixMo Captions

![](_page_21_Picture_4.jpeg)

![](_page_21_Figure_5.jpeg)

### AskModelAnything

![](_page_21_Picture_7.jpeg)

### Pointing

![](_page_21_Picture_9.jpeg)

### Synthetic

![](_page_21_Figure_11.jpeg)

![](_page_21_Picture_12.jpeg)

## Vision language models Trend 2

### Towards zero-shot anything models

- A single model for many tasks: OCR, detection, pointing, generation as different prompts / prefixes
- Push towards in-context learning and broadly useful models

![](_page_22_Figure_4.jpeg)

![](_page_22_Figure_5.jpeg)

## References

- Ferret: Refer and Ground Anything Anywhere at Any Granularity, You etal. 2023
- Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs, You etal. 2024
- Ferret-v2: An Improved Baseline for Referring and Grounding with Large Language, Zhang etal. 2024
- Language-Image Models with 3D Understanding, Cho etal. 2024
- SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities, Chen etal. 2024
- MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, McKinzie etal. 2024
- Chameleon: Mixed-Modal Early-Fusion Foundation Models, Team etal. 2024
- PaliGemma: A versatile 3B VLM for transfer, Beyer etal. 2024
- Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, Deitke etal. 2024