

# Ferret-UI: A peek into multimodal mobile LLM understanding at Apple

Roger Lam  
[lamroger.com](http://lamroger.com)

# Ferret-UI is trained on mobile screenshots for reasoning

The original Ferret model is also from Apple published Oct 2023.

Ferret-UI uses a very similar architecture but trained on mobile screenshots, both iOS and Android. Published April 2024.

It's only June but feels like ages old. Multimodal LLMs papers are accelerating. See Meta's Chameleon released today (6/18)

**Elementary Tasks**

**Referring**  
widget classification, icon recognition, OCR

- Classify the UI type of the widget [point0].  
Button
- Identify the icon type of the widget [box1].  
launch
- What text is displayed in the element [scribble0]?  
Reminders  
Don't forget. Use reminders

**Grounding**  
find widget, find icon, find text, widget listing

- Where is the Open button?  
Open button [box0]
- Where is the launch icon?  
launch icon [box1]
- Where is the text "Reminders Don't forget. Use reminders"?  
Reminders, Don't forget. Use reminders [box4]
- Please list all the UI elements on this screen.  
UI widgets present in this screen include text displaying "Reminders, Don't forget. Use reminders" [box4], Open Button [box0], Launch icon [box1], ...

**Advanced Tasks**

**Conversation (perception + interaction)**

- How do I open the Reminders app?  
You can open the Reminders app by tapping on the Open button [box0].
- I want to share the Reminders app with my friend, what should I do?  
You can share the app by tapping on the launch icon [box1].
- How do I navigate to Games?  
You can navigate to the Games tab by tapping on the Games tab [box2].
- Can my 15-year-old daughter use this app?  
Yes, the app is suitable for all ages as indicated by the "4+" age rating [box3].

**Detailed Description**

- Please describe this screen in detail.  
The screen is for a reminders application. At the top left, there's a button labeled "Apple". The title of the screen, "Reminders", is located in the upper middle part of the screen. There's a text below the title that reads "Don't forget. Use reminders".  
The screen is divided into several sections. The first section, labeled "210K Ratings", is located just below the title. The second section, labeled "AGE", contains a text displaying "4+ Years Old".

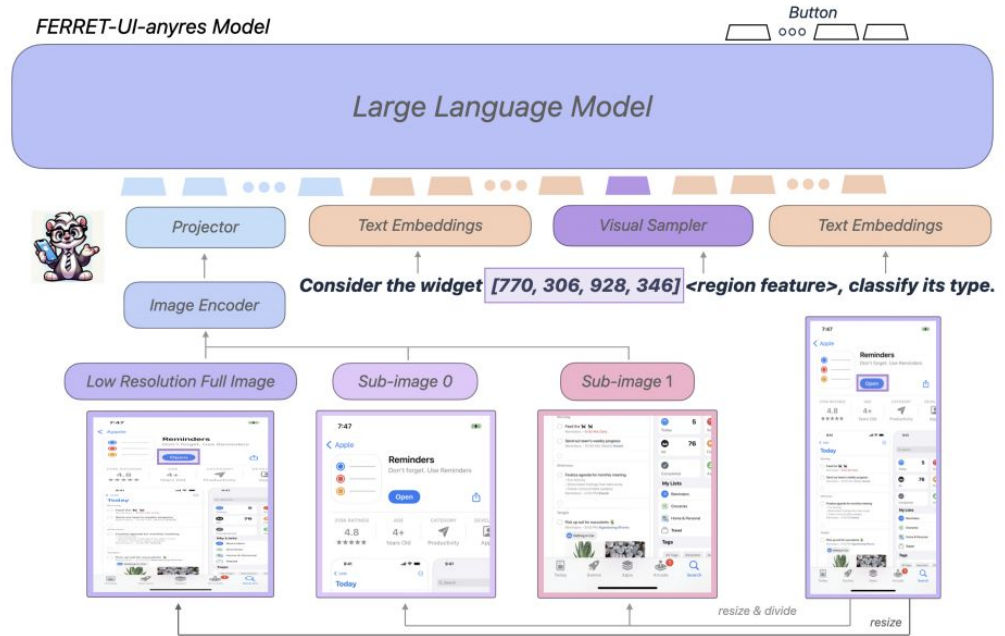
**Function Inference**

- What's the high-level function of the screen?  
The screen is the download page for a reminders application where you can set up reminders for various tasks such as sending out team's weekly progress, grocery shopping, traveling, and picking up kids.

# Incorporates higher resolution sub-images for more detail

Uses pre-trained visual encoder and a decoder-only language model.

They pass in two higher resolution sub-images along with the lower resolution full image and call the model “Ferret-UI-anyres”. Works better than just one full image.



# Used relatively small number of training images

26k Android and 84k iPhone training images

Used Android RICO dataset and iPhone AMP dataset

Cool that these datasets are open

Platform	Resolution	Train	Test
Android	2560×1440	26,527	3,080
	1792×828	74,953	8,297
	828×1792	4,225	461
iPhone	2436×1125	5,420	635
	1125×2436	87	17

(a) Number of screens by resolution.

Task	iPhone	Android
screen2words	-	78k
widget captions	-	109k
taperception	-	14k
elementary tasks	40k×7	40k×7
advanced tasks	10k×4	10k×4

(b) Number of samples per training task.

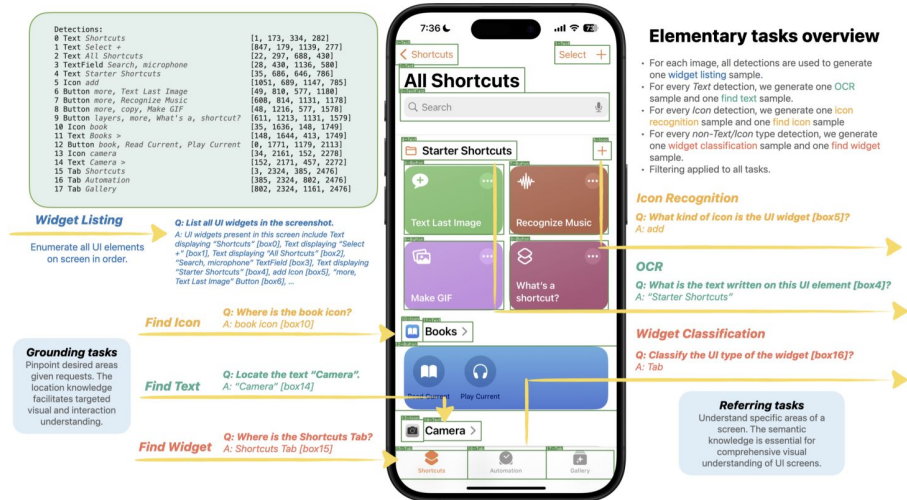
**Table 1:** Mobile UI screen and training data statistics.

# Tasks are split between Elementary and Advanced

Elementary tasks are like find the icon.

Authors detail exactly how they generated data.

They used GPT 3.5 Turbo to generate variants of the task question.

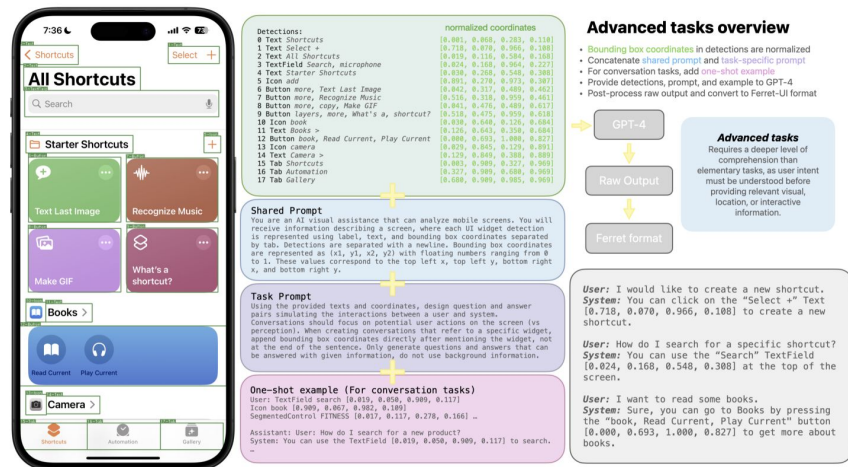


**Fig. 3: Elementary task data generation overview.** A UI detector outputs all detected elements, with each element's *type*, *text*, and *bounding boxes*. These detections are used to create training samples for elementary tasks. For *grounding tasks*, we use all element detections to create one sample for widget listing whereas the remaining tasks focus on one element at a time. We separate the elements into *icons*, *text*, and *non-icon/text widgets*. For each type, we create one referring and one grounding sample.

# Advanced uses stronger techniques to generate data

Advanced tasks are more open-ended:  
“I want to read a book”

This time normalized coordinates, uses multiple prompts and one-shot example to GPT-4 and then postprocess to Ferret-UI format.



**Fig. 4: Advanced task data generation overview.** We first normalize bounding box coordinates from the detection outputs, then we send the detections, prompts, and optional one-shot example to GPT-4. For detailed description and function inference, we pair the generated response with a pre-selection of prompts to train Ferret-UI. For conversation tasks, we directly transform GPT-4 output to multi-turn conversations.

# A shared prompt, task prompt, and one-shot example

They shared the prompts too.

## Shared Prompt

You are an AI visual assistance that can analyze mobile screens. You will receive information describing a screen, where each UI widget detection is represented using label, text, and bounding box coordinates separated 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 question 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.

# Pretty good? Cool that they benchmarked against GPT-4V

anyres performed better than base

Spotlight used a lot more training data. 80M web screenshots. 2.69M mobile screenshots.

GPT-4V performed a lot better in Advanced Tasks but they used GPT-4 to score and believe it's biased toward GPT-4 answers.

Pretty fast training too btw.

	Public Benchmark			Elementary Tasks				Advanced Tasks	
	S2W	WiC	TaP	Ref-i	Ref-A	Grd-i	Grd-A	iPhone	Android
Spotlight [30]	106.7	141.8	<b>88.4</b>	-	-	-	-	-	-
Ferret [53]	17.6	1.2	46.2	13.3	13.9	8.6	12.9	20.0	20.7
Ferret-UI-base	113.4	<b>142.0</b>	78.4	80.5	<b>82.4</b>	79.4	83.5	73.4	80.5
Ferret-UI-anyres	<b>115.6</b>	140.3	72.9	<b>82.4</b>	<b>82.4</b>	<b>81.4</b>	<b>83.8</b>	93.9	71.7
GPT-4V [1]	34.8	23.5	47.6	61.3	37.7	70.3	4.7	<b>114.3</b>	<b>128.2</b>

**Table 2:** Results of Ferret-UI and baseline models. *S2W*: screen2words, *WiC*: widget captions, *TaP*: taperception. We report the CIDEr score for S2W and WiC and F1 for TaP. For elementary and advanced tasks, we report the averaged performance of corresponding tasks. “i”: iPhone, “A”: Android, “Ref”: Referring, “Grd”: Grounding.

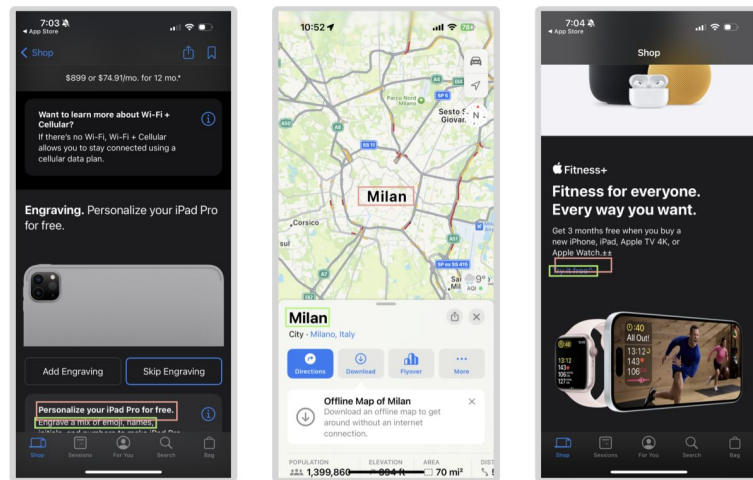
and the training objective is the same as in Ferret. In total, our training mixture has 250K samples. Ferret-UI-base takes 1 day to train while Ferret-UI-anyres takes about 3 days on 8 A100 GPUs.



# Challenges are still there and nice that they ack them

Some tricky cases like selecting neighboring text, when text with the same content occurs on the screen, and bounding boxes are off.

Expected and hope the tech gets better over time or systems build around the uncertainty



**Fig. 8: Find Text Analysis.** *Left:* a neighboring text is mis-identified as the target. *Middle:* multiple occurrences of the same text. *Right:* predicted boxes not precise.

# Thanks!

Ferret-UI: <https://arxiv.org/pdf/2404.05719>

All the Meta drops today:

<https://about.fb.com/news/2024/06/releasing-new-ai-research-models-to-accelerate-innovation-at-scale/>

Original Ferret: <https://arxiv.org/abs/2310.07704>

Take away: If you're interested, you can do it too! Very exciting that they were able to do this with small amounts of data. And maybe some version of this is in their action automation.