

2024.04.24 / Lab Seminar



# Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs

**H** U M A N  
**C** E N T E R E D  
**C** O M P U T I N G  
**L A B** O R A T O R Y

인간중심컴퓨팅 연구실  
석사과정 송형우



# Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs

Keen You, Haotian Zhang, Eldon Schoop, Floris Weers,  
Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan

Apple

{k\_you,haotian\_zhang2,eldon,fweers,  
aswearngin,jwnichols,yinfeiy,zhe.gan}@apple.com

1. MLLMs (Multimodal Large Language Models)
  - UI 이해 및 상호작용 어려움
    - 일반적이지 않은 UI 스크린의 긴 종횡비 & 아이콘, 텍스트와 같은 소규모 객체
2. Ferret UI
  - 각 스크린을 2개의 하위 이미지로 나눔
    - 세로 스크린은 수평 분할, 가로 스크린은 수직 분할
  - Any Resolution 적용
    - 세부 사항 확대, 시각적 특성 강화
3. Dataset
  - For Elementary UI tasks
    - 아이콘 인식, 텍스트 찾기, 위젯 분류
    - Referring and grounding을 위한, 부분부분에 대한 annotation 진행
  - For Advanced Tasks
    - 상세한 설명, 인식/상호작용 대화, 기능 추론
4. Results
  - 위의 Tasks를 포괄하는 종합 벤치마크 설정
  - UI screen 이해 & Open-ended instructions: 모든 오픈 소스 LLM 능가
  - Elementary UI tasks는 GPT-4V 능가

- Contribution

- Refined Model
- Enhanced Data
- New Benchmark

## Use of MLLMs for UI tasks

기존: 단순화된 웹 및 모바일 화면 중점

but LLM 발달에 따른 변화  
→

ILuvUI, Spotlight: GPT 활용

단일 화면 UI, 화면 요약 및 위젯 상호 작용

MobileAgent, AppAgent: MLLM 활용

MobileAgent – 액션 생성을 위한 외부 탐지 모듈

AppAgent – 미리 정의된 액션을 위한, UI element ID와 화면 xml 파일 활용

## Ferret

# FERRET: REFER AND GROUND ANYTHING ANYWHERE AT ANY GRANULARITY

✉ **Haoxuan You**<sup>1†</sup>, **Haotian Zhang**<sup>2†</sup>, **Zhe Gan**<sup>2</sup>, **Xianzhi Du**<sup>2</sup>, **Bowen Zhang**<sup>2</sup>, **Zirui Wang**<sup>2</sup>,  
**Liangliang Cao**<sup>2</sup>, **Shih-Fu Chang**<sup>1</sup>, **Yinfei Yang**<sup>2</sup>

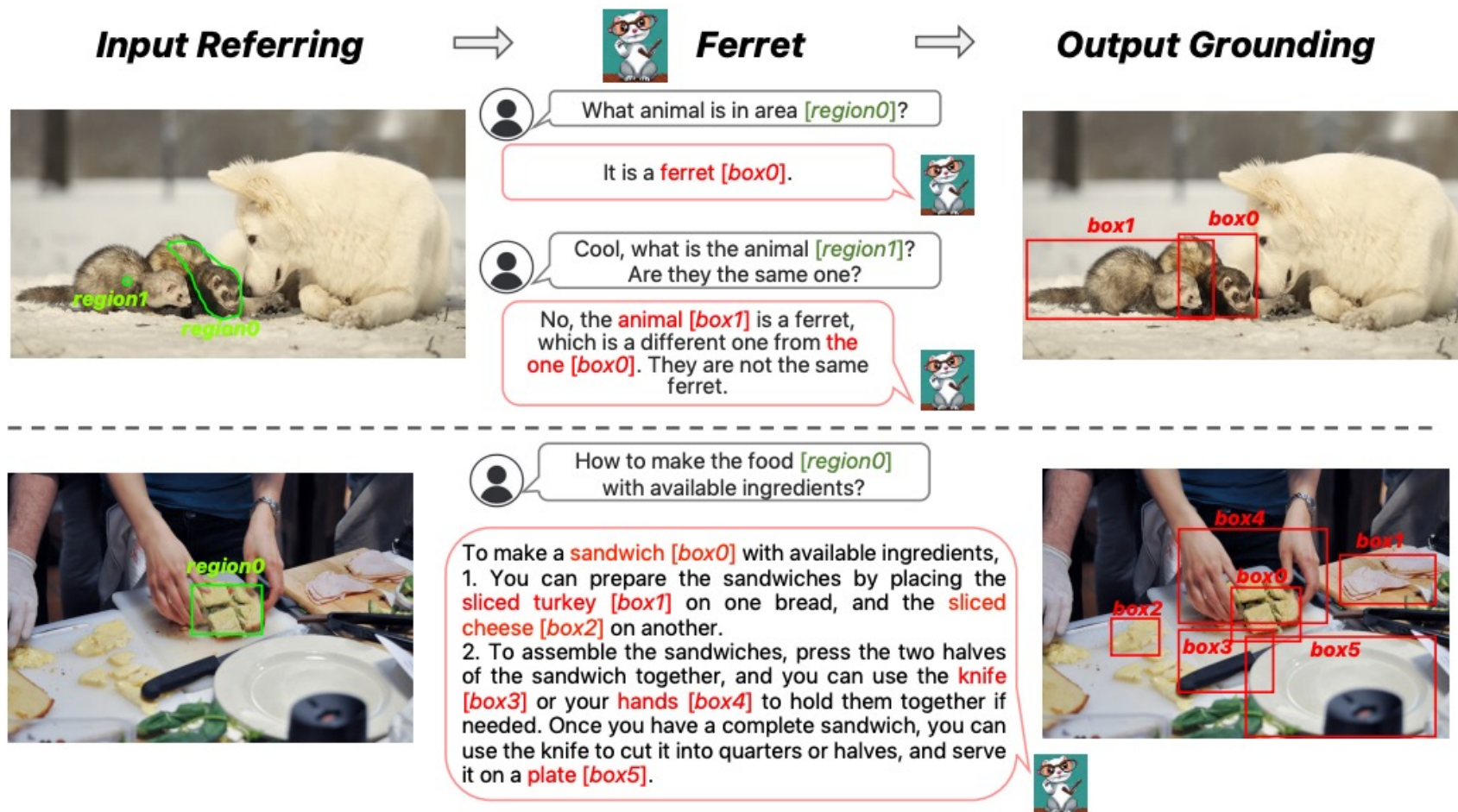
<sup>1</sup>Columbia University, <sup>2</sup>Apple AI/ML

haoxuan.you@cs.columbia.edu, {haotian-zhang2, zhe.gan, yinfeiy}@apple.com

2023년 10월 submit

- Referring과 Grounding을 강화한 모델

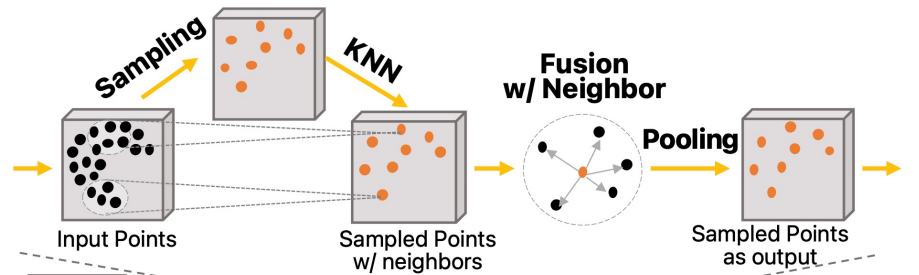
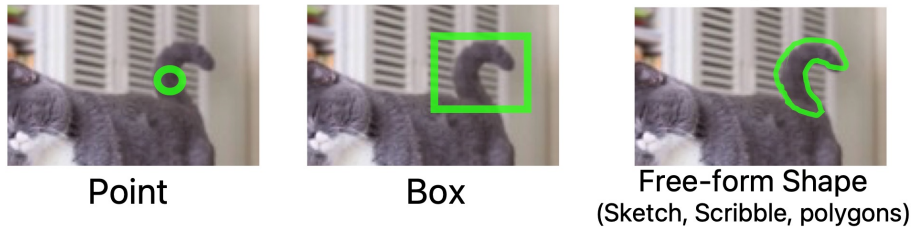
# Ferret



# Ferret

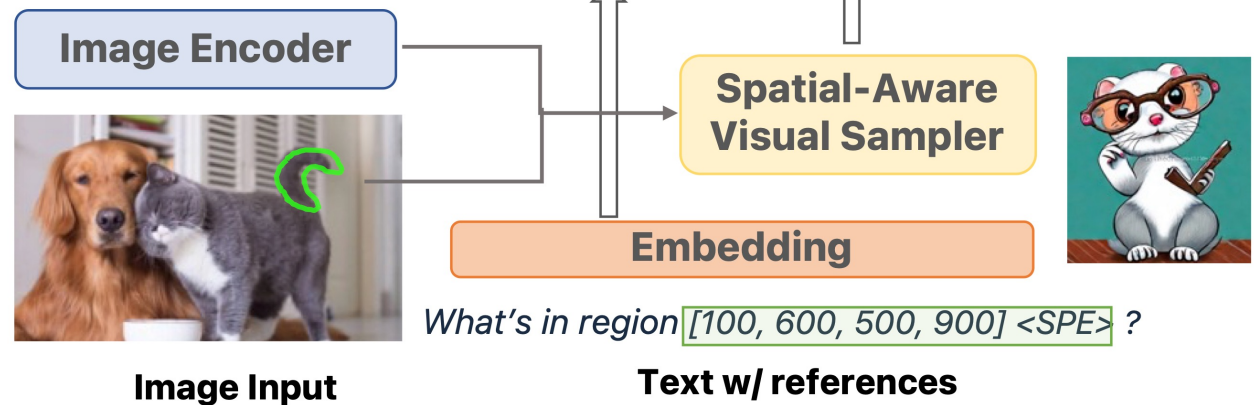
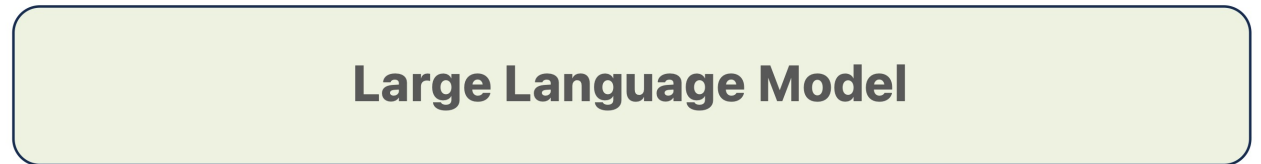
## Hybrid Region Representation

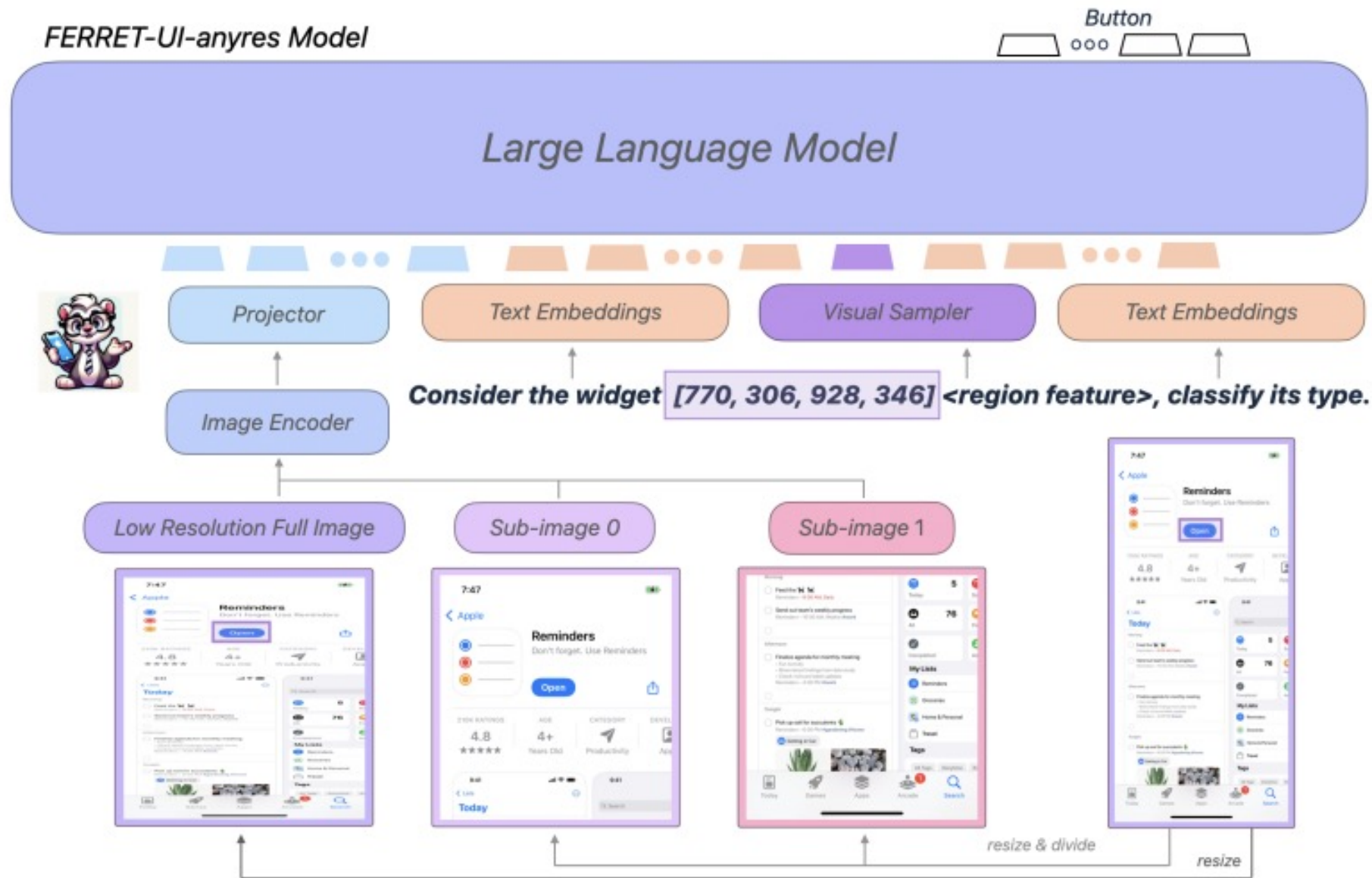
Region Name + [Coordinates] + <feature>



## Ferret Model

It's a cat tail [80, 590, 450, 920]





Platform	Resolution	Train	Test
Android	2560×1440	26,527	3,080
iPhone	1792×828	74,953	8,297
	828×1792	4,225	461
	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.**Spotlight Tasks**

- Screen2words
- Widget Captions
- Taperception

**Elementary Tasks**

- Referring Tasks: OCR, 아이콘 인식, 위젯 분류
- Grounding Tasks: 텍스트 찾기, 아이콘 찾기, 위젯 찾기, 위젯 목록화

**Advanced Tasks**

- 구체적인 설명
- 인식/상호작용 대화
- 기능 추론

**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?

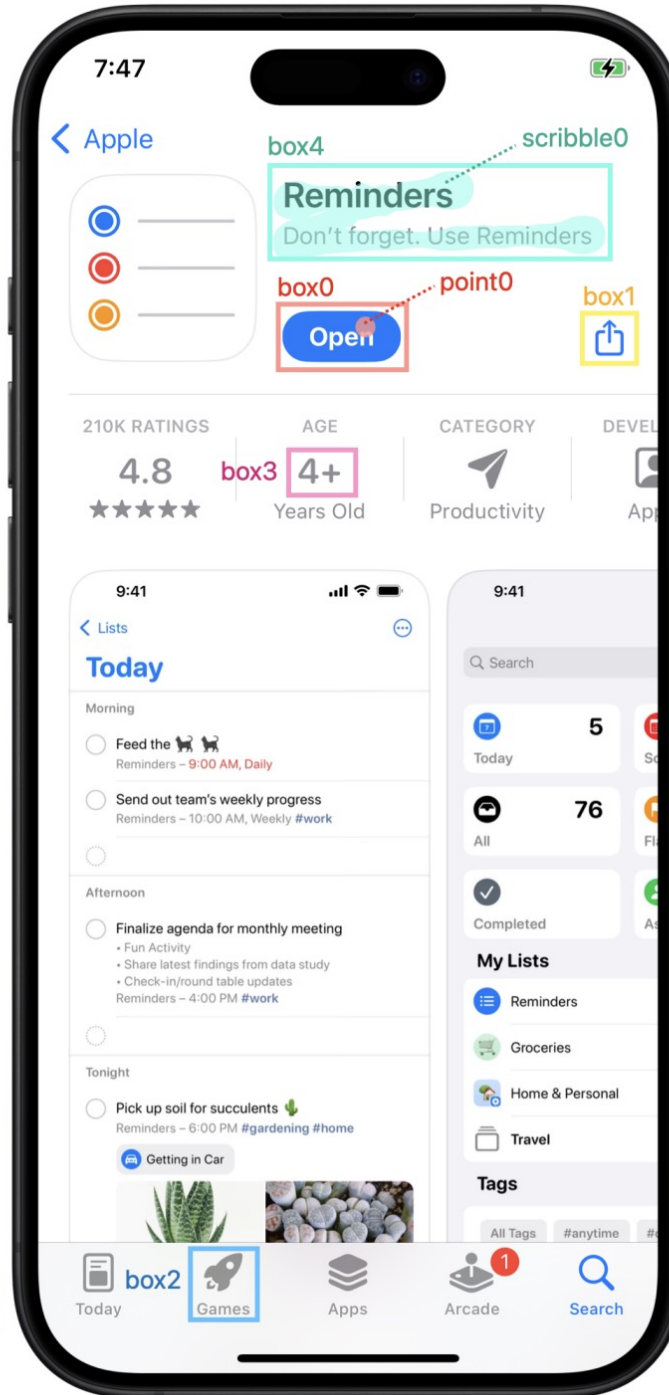
launch icon [box1]

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+", is located just below the title. The second section, labeled "AGE", contains a text displaying "4+ Years Old" ...

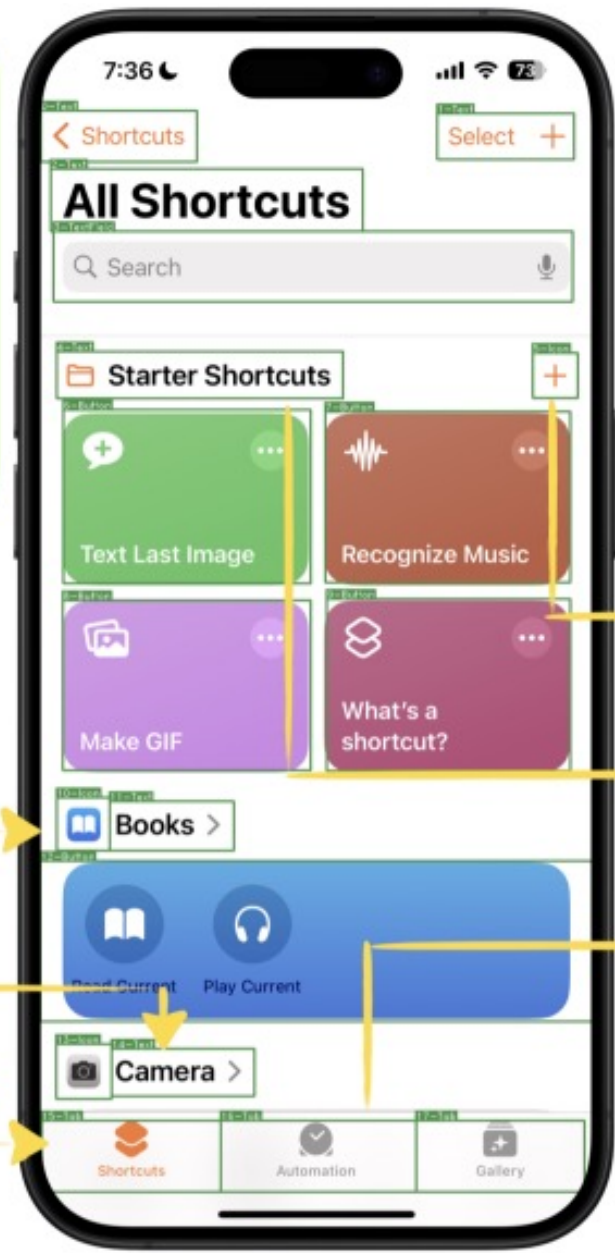
**Function Inference**

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.

Detections:

0	Text Shortcuts	[1, 173, 334, 282]
1	Text Select +	[847, 179, 1139, 277]
2	Text All Shortcuts	[22, 297, 688, 430]
3	TextField Search, microphone	[28, 430, 1136, 580]
4	Text Starter Shortcuts	[35, 686, 646, 786]
5	Icon add	[1051, 689, 1147, 785]
6	Button more, Text Last Image	[49, 810, 577, 1180]
7	Button more, Recognize Music	[608, 814, 1131, 1178]
8	Button more, copy, Make GIF	[48, 1216, 577, 1578]
9	Button layers, more, What's a, shortcut?	[611, 1213, 1131, 1579]
10	Icon book	[35, 1636, 148, 1749]
11	Text Books >	[148, 1644, 413, 1749]
12	Button book, Read Current, Play Current	[0, 1771, 1179, 2113]
13	Icon camera	[34, 2161, 152, 2278]
14	Text Camera >	[152, 2171, 457, 2272]
15	Tab Shortcuts	[3, 2324, 385, 2476]
16	Tab Automation	[385, 2324, 802, 2476]
17	Tab Gallery	[802, 2324, 1161, 2476]



## Elementary tasks overview

- For each image, all detections are used to generate one **widget listing** sample.
- For every *Text* detection, we generate one **OCR** sample and one **find text** sample.
- For every *Icon* detection, we generate one **icon recognition** sample and one **find icon** sample
- For every *non-Text/Icon* type detection, we generate one **widget classification** sample and one **find widget** sample.
- Filtering applied to all tasks.

### Icon Recognition

**Q:** What kind of icon is the UI widget [box5]?  
**A:** add

### OCR

**Q:** What is the text written on this UI element [box4]?  
**A:** "Starter Shortcuts"

### Widget Classification

**Q:** Classify the UI type of the widget [box16]?  
**A:** Tab

**Referring tasks**  
 Understand specific areas of a screen. The semantic knowledge is essential for comprehensive visual understanding of UI screens.

### Widget Listing

Enumerate all UI elements on screen in order.

**Q:** List all UI widgets in the screenshot.  
**A:** UI widgets present in this screen include Text displaying "Shortcuts" [box0], Text displaying "Select +" [box1], Text displaying "All Shortcuts" [box2], "Search, microphone" TextField [box3], Text displaying "Starter Shortcuts" [box4], add Icon [box5], "more, Text Last Image" Button [box6], ...

### Find Icon

**Q:** Where is the book icon?  
**A:** book icon [box10]

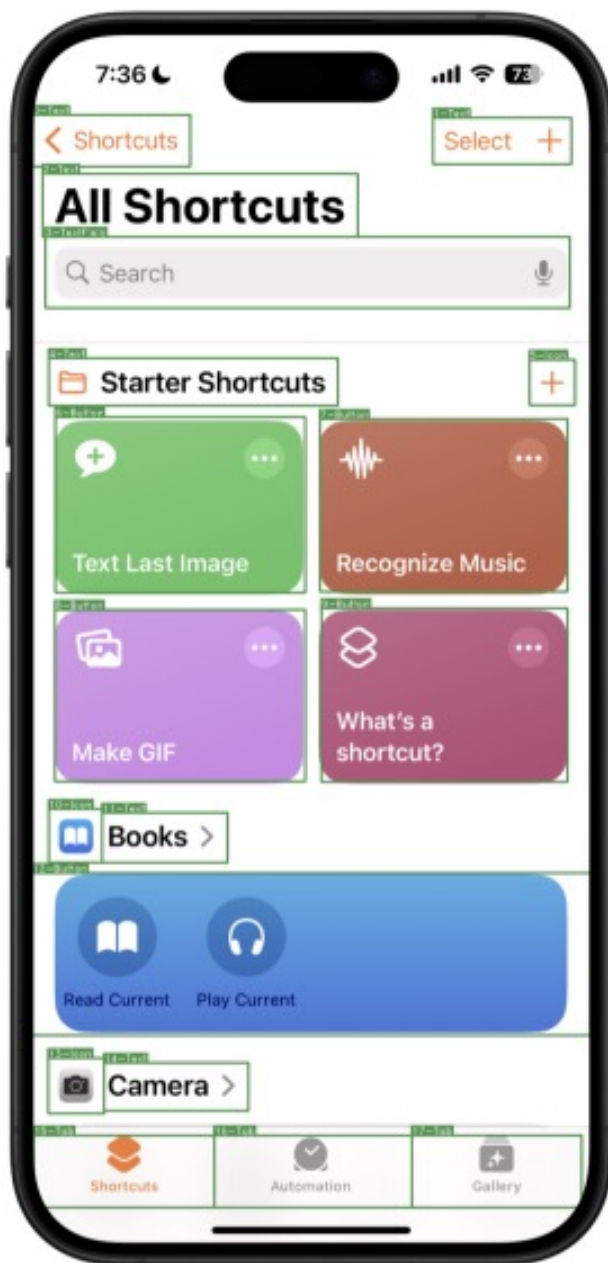
### Find Text

**Q:** Locate the text "Camera".  
**A:** "Camera" [box14]

### Find Widget

**Q:** Where is the Shortcuts Tab?  
**A:** Shortcuts Tab [box15]

**Grounding tasks**  
 Pinpoint desired areas given requests. The location knowledge facilitates targeted visual and interaction understanding.



Detections:	normalized coordinates
0 Text Shortcuts	[0.001, 0.068, 0.283, 0.110]
1 Text Select +	[0.718, 0.070, 0.966, 0.108]
2 Text All Shortcuts	[0.019, 0.116, 0.584, 0.168]
3 TextField Search, microphone	[0.024, 0.168, 0.964, 0.227]
4 Text Starter Shortcuts	[0.030, 0.268, 0.548, 0.308]
5 Icon add	[0.891, 0.270, 0.973, 0.307]
6 Button more, Text Last Image	[0.042, 0.317, 0.489, 0.462]
7 Button more, Recognize Music	[0.516, 0.318, 0.959, 0.461]
8 Button more, copy, Make GIF	[0.041, 0.476, 0.489, 0.617]
9 Button layers, more, What's a, shortcut?	[0.518, 0.475, 0.959, 0.618]
10 Icon book	[0.030, 0.640, 0.126, 0.684]
11 Text Books >	[0.126, 0.643, 0.350, 0.684]
12 Button book, Read Current, Play Current	[0.000, 0.693, 1.000, 0.827]
13 Icon camera	[0.029, 0.845, 0.129, 0.891]
14 Text Camera >	[0.129, 0.849, 0.388, 0.889]
15 Tab Shortcuts	[0.003, 0.909, 0.327, 0.969]
16 Tab Automation	[0.327, 0.909, 0.680, 0.969]
17 Tab Gallery	[0.680, 0.909, 0.985, 0.969]

### 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.

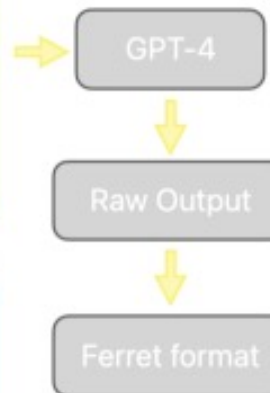
### One-shot example (For conversation tasks)

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
- Post-process raw output and convert to Ferret-UI format



**Advanced tasks**  
 Requires a deeper level of comprehension than elementary tasks, as user intent must be understood before providing relevant visual, location, or interactive information.

**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.

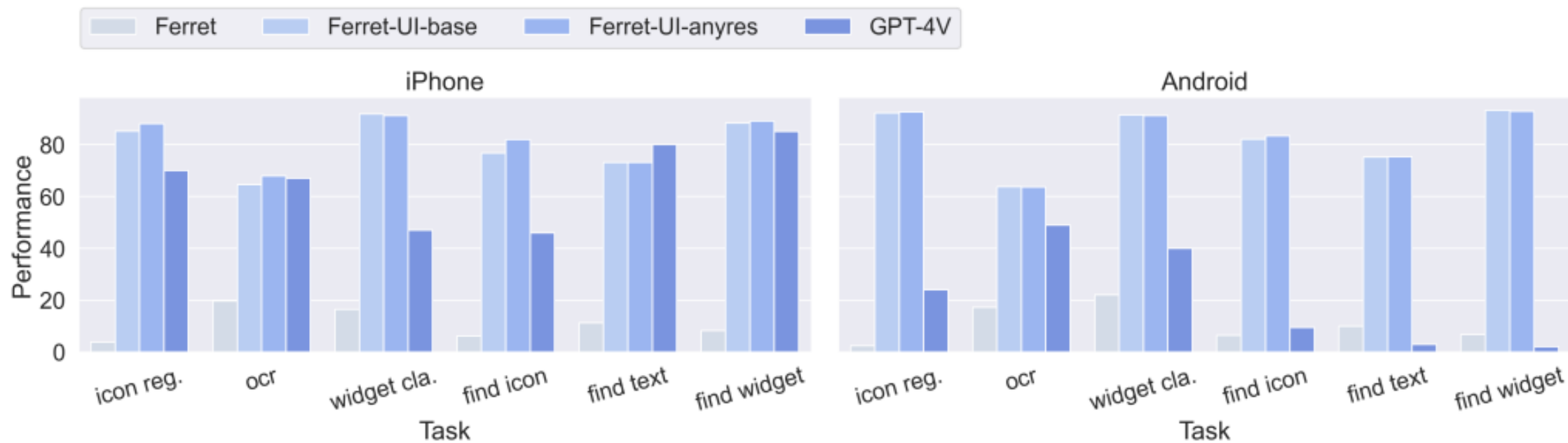
**User:** How do I search for a specific shortcut?  
**System:** You can use the "Search" TextField [0.024, 0.168, 0.548, 0.308] at the top of the screen.

**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.

## Overall Results

	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>

## Elementary Task Performance



## Advanced Task Performance

	iPhone					Android				
	DetDes	ConvP	ConvI	FuncIn	<b>Avg</b>	DetDes	ConvP	ConvI	FuncIn	<b>Avg</b>
Ferret [53]	2.5	34.7	23.7	19.1	20.0	2.0	33.9	24.9	21.9	20.7
Fuyu [6]	5.0	24.6	18.8	35.7	21.0	2.0	20.8	44.5	36.1	25.9
CogAgent [20]	53.1	59.7	74.8	71.9	64.9	28.0	58.5	90.1	<b>90.5</b>	66.8
Ferret-UI-base	64.5	75.0	77.5	76.5	73.4	90.8	72.8	79.3	79.2	80.5
Ferret-UI-anyres	<b>97.4</b>	92.1	91.1	<b>95.2</b>	93.9	86.4	70.3	50.2	77.3	70.1
GPT-4V [1]	66.8	<b>105.6</b>	<b>198.5</b>	86.3	<b>114.3</b>	<b>126.6</b>	<b>109.4</b>	<b>188.6</b>	88.3	<b>128.2</b>

## Ablation Study

### 1. Any Resolution

1. 2% 정도의 성능 향상을 볼 수 있음

### 2. Elementary Task Training Data

1. Elementary를 학습하고나서 Advanced를 하니, 3~9% 정도의 성능 향상

## Contribution

1. Anyres를 통합하고, referring과 grouning을 UI에 초점을 맞춰 실행할 수 있는 MLLM 최초 개발
2. UI screen data와 task별 dataset 생성
3. 여러 task를 아우르는 새로운 Benchmark 제안

행복한 하루 되세요