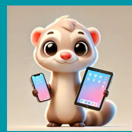
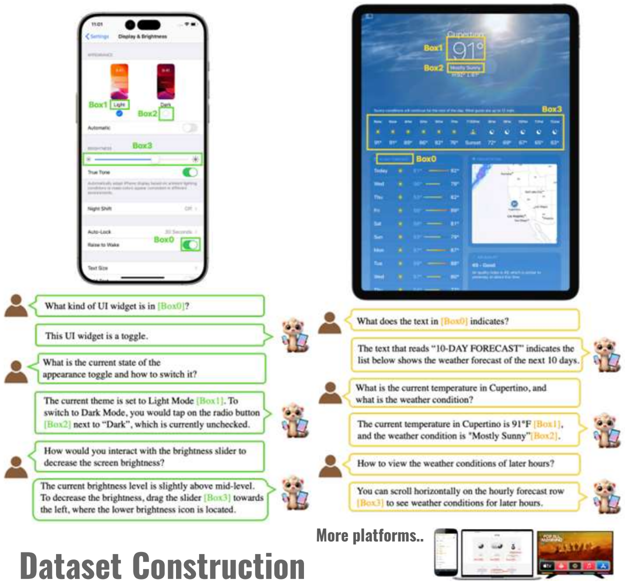


Ferret-UI 2: Mastering Universal User Interface Understanding Across Platforms

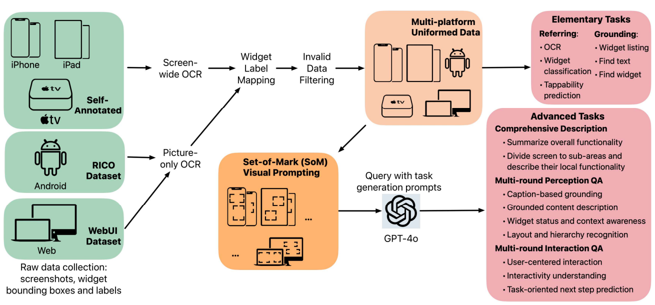
Zhangheng Li[2], Keen You[1], Haotian Zhang[1], Di Feng[1], Harsh Agrawal[1], Xiujun Li[1], Mohana Prasad Sathya Moorthy[1], Jeffrey Nichols[1], Yinfei Yang[1], Zhe Gan[1]
[1]Apple Inc.
[2]The University of Texas at Austin (work done during internship at Apple)

Introduction

Building a generalist model for user interface (UI) understanding is challenging due to various foundational issues, such as platform diversity, resolution variation, and data limitation. In this paper, we introduce Ferret-UI 2, a multimodal large language model (MLLM) designed for universal UI understanding across a wide range of platforms, including iPhone, Android, iPad, Webpage, and AppleTV. Building on the foundation of Ferret-UI, Ferret-UI 2 introduces three key innovations: support for multiple platform types, high-resolution perception through adaptive scaling, and advanced task training data generation powered by GPT-4o with set-of-mark visual prompting. These advancements enable Ferret-UI 2 to perform complex, user-centered interactions, making it highly versatile and adaptable for the expanding diversity of platform ecosystems.



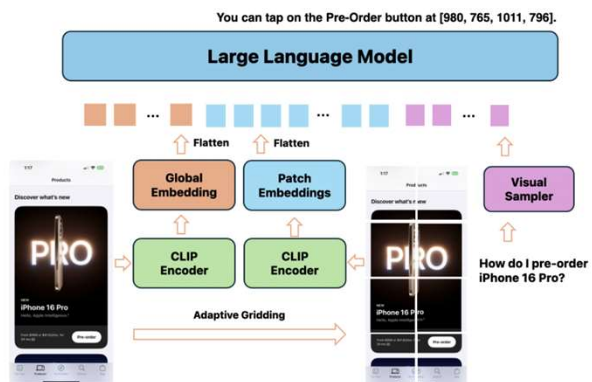
Dataset Construction



Task type: Multi-round Interaction QA
Conversations:
User:
<image>
How can I switch to dark mode on this screen?
Model:
You can switch to dark mode by tapping the "Dark" option [<WIDGET_34>] and then selecting the circle [<WIDGET_7>] underneath it.
User:
What does the "True Tone" option do?
Model:
The "True Tone" option [<WIDGET_21>] automatically adapts the iPhone display based on ambient lighting conditions to make colors appear consistent [<WIDGET_39> <WIDGET_38>].
User:
How can I adjust the brightness?
Model:
You can adjust the brightness using the slider [<WIDGET_4>] under the "BRIGHTNESS" section [<WIDGET_23>].
User:
Is there a way to toggle the automatic brightness?
Model:
Yes, you can toggle automatic brightness by switching the toggle [<WIDGET_15>] next to "Automatic" [<WIDGET_17>].
User:
Where do I turn off the "Raise to Wake" feature?
Model:
The "Raise to Wake" feature is currently on, you can tap on the toggle [<WIDGET_35>] to switch it off.

Example of set-of-mark visual prompting augmented screenshot (left) and one of its generated advanced task training examples (right).

Model Architecture



Algorithm 1: Adaptive N-gridding

Require: Original resolution: w × h, grid size: 336 × 336, size limit N
Ensure: Optimal gridding size N_w and N_h $(N_w, N_h \in \mathbb{N}^+)$

```
1:  N_{w_best}, N_{h_best} ← 0, Δ_best ← ∞, N_{w_0} ← w/336, N_{h_0} ← h/336
2:  for N_w = 1 to N do                                  ▷ Traverse all grid configuarations
3:      for N_h = 1 to N − N_w do
4:          Δ_aspect ← sqrt(|N_w/N_h − N_{w_0}/N_{h_0}| |N_h/N_w − N_{h_0}/N_{w_0}|)   ▷ Get aspect ratio change
5:          Δ_pixel ← |N_w×N_h − N_{w_0}×N_{h_0}| / (N_{w_0}×N_{h_0})                 ▷ Get relative pixel change for resizing
6:          if Δ_best > Δ_aspect × Δ_pixel then
7:              (N_{w_best}, N_{h_best}) ← (N_w, N_h)
8:              Δ_best ← Δ_aspect × Δ_pixel
9:          end if
10:     end for
11: end for
12: return (N_{w_best}, N_{h_best})
```

We introduce adaptive N-gridding to decide optimial gridding config under constrained budget (Grid number limit N).

Experiments

Results on our constructed benchmarks for elementary and advanced tasks, as well as the GUIDE benchmark (Chawla et al., 2024). Results on elementary and advanced tasks are averaged over all platforms, including iPhone, Android, iPad, Webpage, and AppleTV. Each platform includes 6 elementary tasks and 3 advanced tasks. SeeClick model (Cheng et al., 2024) trained on their original data is compared. (†) In tasks that require referring, GPT-4o is equipped with set-of-mark (SoM) prompting by adding a red rectangular box to screenshots for the referred widget. Note that SoM visual prompting is not used for Ferret-UI and Ferret-UI 2.

Model	Backbone	Elementary		Advanced		GUIDE Bench	
		Refer	Ground	GPT-4o Score	Multi-IoU	BertScore	IoU
Ferret-UI	Vicuna-13B	64.15	57.22	45.81	18.75	41.15	26.91
Ferret-UI 2	Gemma-2B	75.20	78.13	80.25	40.51	83.71	51.13
	Llama3-8B	80.28	**82.79**	**89.73**	41.15	**91.37**	**55.78**
	Vicuna-13B	**81.34**	81.31	86.25	**41.71**	88.81	54.71
SeeClick (Cheng et al., 2024)	QWen-VL-9.6B	51.58	62.82	67.49	21.56	54.70	39.51
GPT-4o	-	56.47	12.14	77.73	7.06	75.31	9.64
GPT-4o + SoM-Prompt†	-	**87.91**	-	84.33	7.36	-	-

Zero-shot performance of Ferret-UI 2 on the GUI-World benchmark (Chen et al., 2024a).

Model	GPT-4 Score			
	iOS	Android	Webpage	Average
MiniGPT4Video (Ataallah et al., 2024)	1.501	1.342	1.521	1.455
VideoChat2 (Li et al., 2024a)	2.169	2.119	2.221	2.170
Chat-Univi (Jin et al., 2024)	2.337	2.390	2.349	2.359
GUI-Vid (Chen et al., 2024a)	2.773	2.572	2.957	2.767
QWen-VL-MAX (Bai et al., 2023)	2.779	2.309	2.656	2.580
SeeClick Cheng et al. (2024)	2.614	2.650	2.848	2.704
Ferret-UI (You et al., 2024)	2.713	2.791	2.411	2.638
Ferret-UI 2	**2.881**	**2.954**	**3.013**	**2.948**
Gemini-Pro 1.5 (Reid et al., 2024)	3.213	3.220	3.452	3.295
GPT-4o	3.558	3.561	3.740	3.619

Ablation

Zero-shot cross-platform transfer results of Ferret-UI 2 demonstrate the original domain gaps between different platforms.

Training	Test - Referring					Test - Grounding				
	iPhone	iPad	AppleTV	Web	Android	iPhone	iPad	AppleTV	Web	Android
iPhone	86.3	68.1	31.2	45.3	71.2	84.1	65.2	43.1	51.7	63.1
iPad	67.5	80.2	40.7	51.5	63.3	64.5	82.1	32.1	38.5	53.8
AppleTV	29.1	45.1	79.3	54.2	36.4	33.7	41.2	81.6	52.1	29.7
Web	59.2	57.4	41.2	85.5	41.7	54.0	51.2	46.5	87.5	45.9
Android	72.5	60.7	35.7	51.2	86.2	66.7	48.9	29.7	44.1	83.9

Comparison of two gridding strategy: AnyRes from LLaVA v.s. Adaptive N-gridding

Training Data	Model	iPhone v1		iPhone v2	
		GPT-4o Score	Multi-IoU	GPT-4o Score	Multi-IoU
iPhone v1	Ferret-UI-anyRes	91.3	36.89	68.3	27.13
	Ferret-UI 2	93.7 (+2.4)	37.12 (+0.23)	70.2 (+1.9)	28.21 (+1.08)
iPhone v2	Ferret-UI-anyRes	86.2	35.89	85.97	39.81
	Ferret-UI 2	88.1 (+1.9)	36.43 (+0.54)	89.7 (+3.73)	41.73 (+1.92)

Conclusion

With multi-platform support, highresolution image encoding, and improved data generation, Ferret-UI 2 demonstrates strong zero-shot transferability across platforms, establishing it as a solid foundation for universal UI understanding.