MP-GUI: Modality Perception with MLLMs for GUI Understanding

Supplementary Material

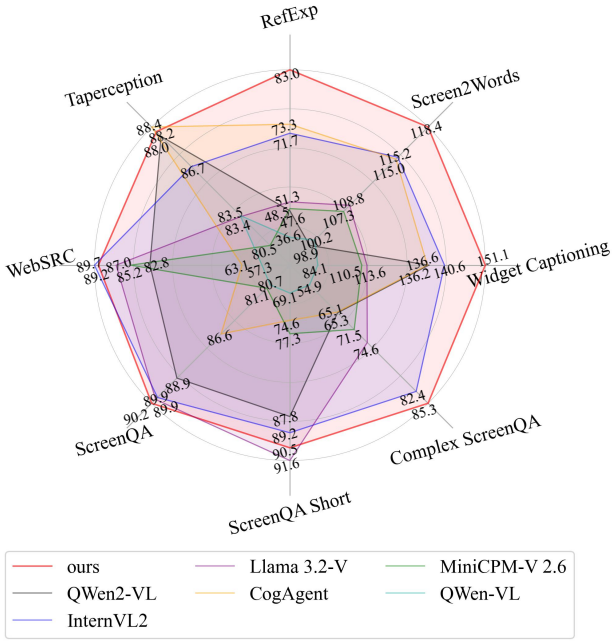


Figure 6. MP-GUI outperforms six open-source MLLMs in the GUI understanding benchmark.

Our codes and datasets are publicly available at https://github.com/BigTaige/MP-GUI.

A. Training Configurations

We report the detailed settings of MP-GUI during multi-step training and multi-task fine-tuning, as shown in Tab. 9. As introduced in Sec 3.2: *Step 1* represents Textual Perceiver training, *Step 2* represents Graphical Perceiver training, *Step 3* represents Spatial Perceiver training, and *Step 4* is Fusion Gate training.

B. Details of Evaluation Datasets

In this section, we describe the details of each task in the GUI understanding benchmark and the templates we used.

Widget Captioning (WC) [35]: It is a benchmark for automatically generating language description for the functionality of an object on the screen. The numbers of samples for the partitioned train/val/test are 14,878/1,292/1,265 respectively. The template we use is as follows, where ***bbox*** represents the coordinates area of the target and the *<image>* is a placeholder that will be replaced by image tokens:

Configurations	*Step 1*	*Step 2*	*Step 3*	*Step 4*	*MFT*
Training epochs	1				
Max dynamic patch	6				4
Training samples	160,031	187,657	200,000	93,419	107,373
Warmup ratio	0.03				
Warmup decay	0.01				
Global batch size	64				
Learning rate	1×10^{-5}				4×10^{-5}
Learning rate decay	Cosine schedule				
Optimizer	AdamW				
Adam ϵ	1×10^{-8}				
Adam β	$(0.9, 0.999)$				

Table 9. Training configuration details. ***MFT*** means multi-task fine-tuning.

The template for **Widget Captioning**

*<image>\n Describe the function within the selected area <box> [**bbox**] </box> of the image. answer with phrases rather than sentence.*

Taperception (TP) [52]: This benchmark is used to predict whether a given target element is clickable. It can be used to detect the accessibility of GUI elements on the screen. The numbers of samples for the partitioned train/val/test are 14,781/1,857/2,029. The template employed for this task is as follows:

The template for **Taperception**

*<image>\n Whether the graphic within the selected area <box> [**bbox**] </box> is clickable? If clickable, output 0. otherwise output 1.*

ScreenQA (QA) [23]: This is a benchmark for screen comprehension. It comprises UI elements and full-sentence answers as the ground truth. The objective of this dataset is to extract the OCR content from the screen in conjunction with the given question. The numbers of samples for the partitioned train/val/test are 68,951/8,614/8,419. The template used is as follows, where ***question*** represents the original question of the sample.

The template for **ScreenQA**

*<image>\n **question**?*

ScreenQA Short (QAS) [5]: It is a modified version of ScreenQA [23], having the same questions for the same screenshots, with answers autogenerated by PaLM 2-S [13] from original human-annotated data. The numbers of samples for the partitioned train/val/test are 68,951/8,614/8,419. The template acting on it is as follows:

The template for **ScreenQA Short**

<image>\n ***question****? Answer with numbers or phrases rather than sentence.*

Complex ScreenQA (CQA) [5]: An extension or substitute of ScreenQA Short [5], which incorporates more arduous questions, namely those related to counting, arithmetic, comparison, and non-answerable varieties, as well as screens possessing diverse aspect ratios, is employed to assess the model's proficiency in localizing, spatial perception and reasoning about screen elements, which needs multipart screen information. As the original data lacks details on data division, yet the author noted in the data card that CQA is founded on data synthesized by QAS [5], in this study, we partition the CQA data in line with the image index in QAS [5]. Finally, the numbers of samples for the partitioned train/val/test are 6,347/796/759. We maintain the template adopted in CQA consistent with that of QAS:

The template for **Complex ScreenQA**

<image>\n ***question****? Answer with numbers or phrases rather than sentence.*

WebSRC (WS) [12]: This is a web scenario question-answering benchmark, with the answers primarily centered around the OCR content within the page. The numbers of samples for the partitioned train/val/test are 307,315/4,558/4,558. We ensure that the template remains in line with that of QAS [5]:

The template for **WebSRC**

<image>\n ***question****? Answer with numbers or phrases rather than sentence.*

RefExp (RE) [7]: This is a task of generating the coordinates of the object referred to in the query, used to evaluate the model's accuracy in locating and identifying the position of specific objects within a given context. The numbers of samples for the partitioned train/val/test are 15,624/471/565. The template utilized on it is as follows, where ***reference*** represents the description of the target element:

The template for **RefExp**

<image>\n Please provide the bounding box coordinate of the region this sentence describes: <ref> ***reference****</ref>*

Screen2Words (S2W) [54]: This benchmark requires the model to be aware of the global and local information of the screen and use a concise text to summarize the content and function of the current screen. The numbers of samples for the partitioned train/val/test are 15,743/2,364/4,310. The template that we employed for this particular task is as follows:

The template for **Screen2Words**

<image>\n Use a phrase to describe the function of the page.

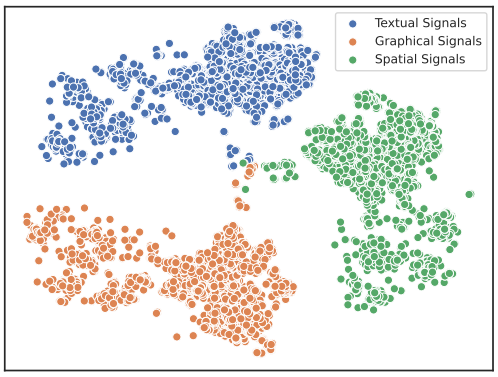
For all the above tasks, we format them into conversational QA pairs to adapt to the inference and training mode of MLLMs. To balance data distribution in multi-task fine-tuning, we sample only the first 10,000 samples from the QA [23] and QAS [5] datasets, and the first 20,000 samples from the WS [12] dataset.

C. Analysis of GUI Perceivers

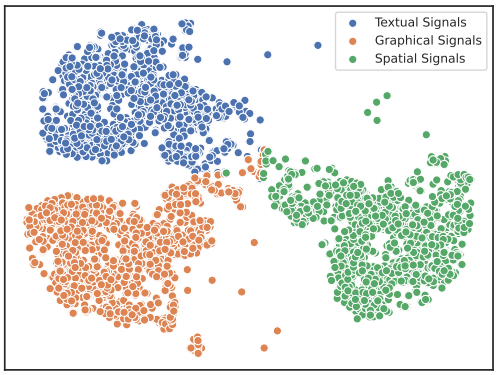
In this section, to confirm that different GUI Perceivers can extract specific GUI modality signals from the visual clues of the visual backbone, we analyze the distribution discrimination in feature space. Specifically, we use t-SNE (t-Distributed Stochastic Neighbor Embedding) to visualize the GUI modality signals generated by different perceivers on images of downstream tasks, and the results are shown in Fig. 7. It can be observed that the feature distributions are clearly distinguished into three groups, demonstrating that our method can extract different GUI modality information from the visual clues effectively.

D. More Comparisons for Grounding Results

Given that grounding ability serves as the foundation for MLLMs to attain more precise GUI understanding [15], in this section, we extend the evaluation metrics (Acc@IoU = 0.1) on the RefExp [7] benchmark. Specifically, we introduce Acc@IoU=0.3, Acc@IoU=0.5, Acc@IoU=0.7 and the Center Point Accuracy (Acc@CP) metrics to further assess the localization capabilities of diverse MLLMs. A larger IoU value (Acc@IoU=0.5/0.7) can quantify the degree of fit of the bounding box generated by the MLLM, and Acc@CP can reflect the model's ability to accurately click on the target area according to the instruction. The



(a) Results on Screen2Words [54].



(b) Results on WebSRC [12].

Figure 7. Visualization results of different GUI modality signals processed by t-SNE.

formula of Acc@CP is defined as follows:

$$Acc@CP = \frac{\sum_{i=1}^{n} \mathbb{I}(pred_i, gt_i)}{n} \times 100\%, \qquad (3)$$

where $\mathbb{I}(pred, gt)$ means an indicator function, which is used to calculate whether the center point of the predicted coordinates $pred$ is located inside gt.

As shown in Tab. 10, although our MP-GUI (8B) achieves the second best result compared to CogAgent(18B) [22] at the Acc@IoU=0.7 metric, still shows advanced performance overall.

E. Spatial Relationship Prediction Examples

To strengthen the pure visual MLLMs in perceiving the spatial relationship among elements on the screen, we introduce the Spatial Perceiver and SRP training tasks for explicit modeling of the spatial relationship (refer to Sec. 3.1 and 3.2). In this part, we display more SRP data samples (see Fig. 8). The SRP dataset is constructed using the VH json files corresponding to the images in the public dataset [43].

	IoU=0.1	*IoU=0.3*	*IoU=0.5*	*IoU=0.7*	*Acc@CP*
Qwen-VL [8]	36.3	25.3	16.3	9.2	59.3
MiniCPM-V 2.6 [60]	48.5	26.2	11.0	2.5	66.5
Qwen2-VL [55]	47.6	36.2	27.7	12.2	86.5
Llama 3.2-V [3]	51.3	29.9	17.3	9.6	63.0
CogAgent [22]	73.3	68.0	58.8	**46.2**	83.9
InternVL2 [14]	71.7	52.9	35.7	17.9	74.9
MP-GUI (ours)	**83.0**	**74.3**	**60.0**	41.2	**87.4**

Table 10. Evaluation of baseline MLLMs on RefExp [7] benchmark using different metrics. IoU=0.1/0.3/0.5/0.7 are shorthand for Acc@IoU=0.1/0.3/0.5/0.7 respectively.

F. Prompts in Automated Pipeline

In this section, we present the prompts fed to Qwen2-VL (72B) [55] for generating Single Perceiver Enhanced Question Answering (SPE-QA) and Multi-Perceiver Enhanced Question Answering (MPE-QA) data, as introduced in Sec. 4.2. The framework of the data synthesis pipeline is shown in Fig. 9.

F.1. SPE-QA

The prompt for **SPE-QA**

```
Design some QA pairs based only
on the icons in the picture, only
on the text in the picture, only
on some relationships between
components and only on locations
of components (such as the
return icon is in the upper left
corner of the screen.), and give
questions and correct answers.
Please format the data as JSON
format such as 'question': ...,
'type': 'text' or 'icon' or
'relationship' or 'location',
'answer': ....
```

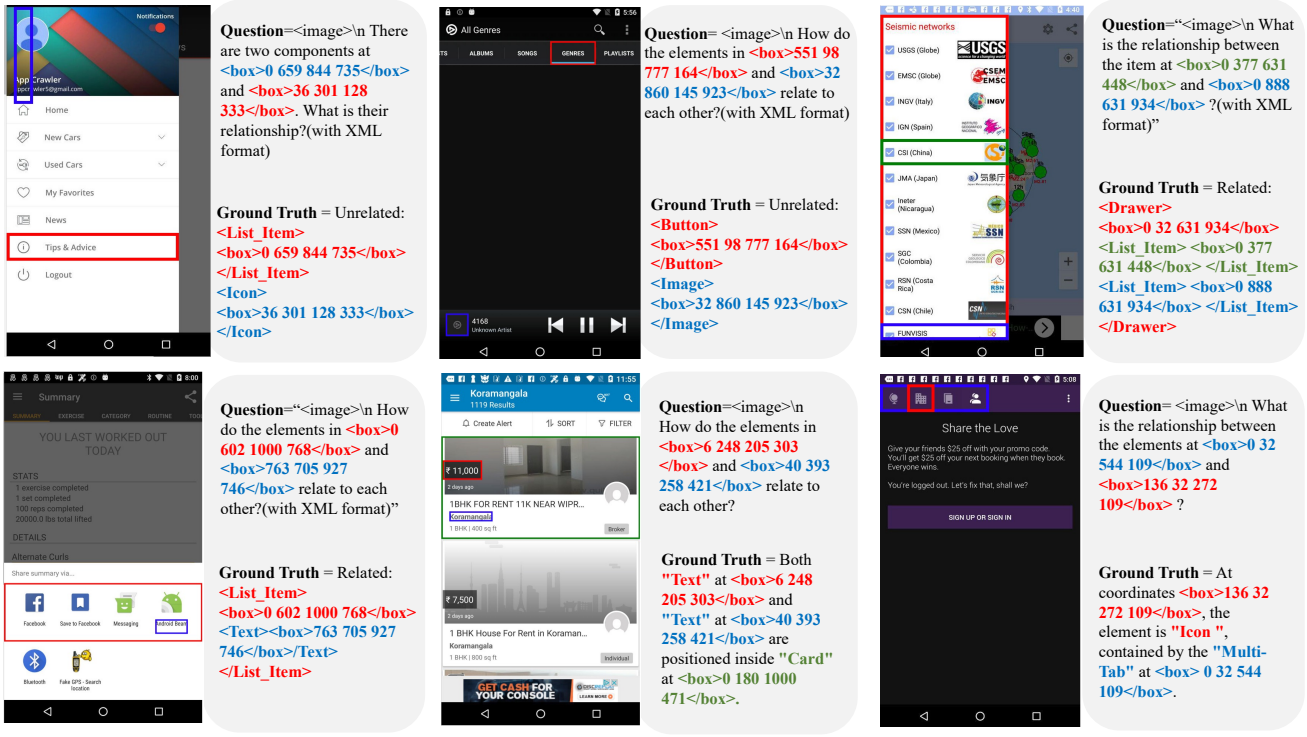


Figure 8. Examples of our SRP data.

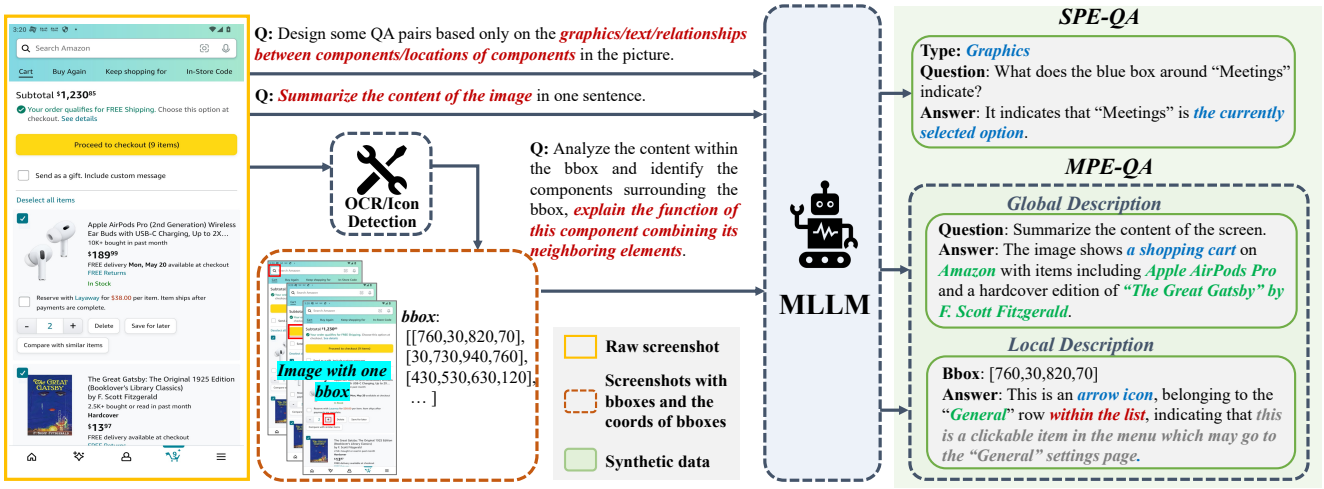


Figure 9. The pipeline for synthetic data generation. We categorize the data into: SPE-QA (Single Perceiver Enhanced Question Answering) and MPE-QA (Multi-Perceiver Enhanced Question Answering).

F.2. MPE-QA

The prompt for Global Description

```
Generate a summary of the screen
in one sentence.  Do not focus on
specifically naming the various
UI elements, but instead, focus on
the content.
```

The prompt for **Local Description**

```
Describe this image. You will
receive a screenshot of a GUI that
includes a bounding box (bbox)
with specified coordinates. Your
task is to analyze the content
within the bbox and identify the
component to which it belongs by
looking for surrounding component
boundaries. Please provide a
detailed description that includes
the following:

    1.Identify the content
    inside the bbox (text or
    graphic element).

    2.Look for the component
    boundary surrounding
    the bbox and describe
    the overall component it
    belongs to.

    3.Explain the function of
    this component and any
    other relevant elements
    it contains.

    4.If there are no
    surrounding component
    boundaries, state that
    there are no related
    components nearby.

Output Example (response with just
one sentence):

    "This is an icon of a
    house, belonging to a
    button component that
    describes the home page;
    it also includes another
    house icon as part of this
    component."

    "This is an arrow icon,
    belonging to the 'General'
    row within the list,
    indicating that this is
    a clickable item in the
    menu which may go to the
    'General' page."

    "This is a standalone
    button labeled 'Submit',
    and there are no related
    components nearby."
```

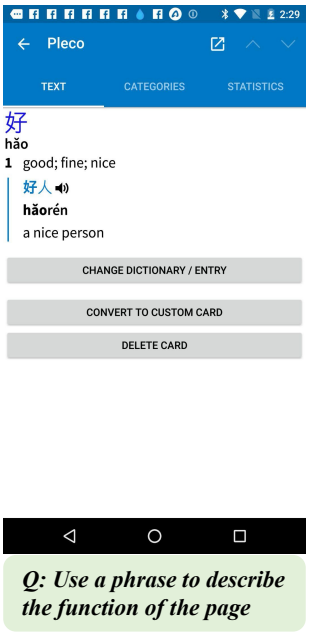


Figure 10. A comparison on Screen2Words [54].

```
Now the coordinate of bbox I'd
like you to analyze is [bbox]
```

G. More Qualitative Analysis

In this section, we show more qualitative results of our MP-GUI with other MLLMs on downstream tasks.

Screen2Words. As shown in Fig. 10, MP-GUI is capable of taking into account the overall layout and determining that the page belongs to the language learning app. In contrast, all other methods are distracted by the sizable translation portion in the middle of the screen.

Widget Captioning. As depicted in Fig. 11, under the guidance of the novel Local Description task (see Sec. 4.2), our MP-GUI is more inclined to summarize the graphics by combining the spatial context information. In the first example, MP-GUI can summarize the high-level function of *"play episode 489"* by combining the text on the left of the button, instead of only focusing on the graphical element *"play"*. Meanwhile, our method is also capable of differentiating the core content within the target area, as demonstrated in the third example. Furthermore, due to the excellent grounding ability, MP-GUI is able to precisely comprehend the coordinates in the input question and provide accurate answers, rather than misidentifying the location as *"dashcam support"* (in Example 2) or *"continue"* (in Example 4).

ScreenQA Short. In the scenarios presented in Fig. 12, we observe that MP-GUI exhibits favorable OCR and com-

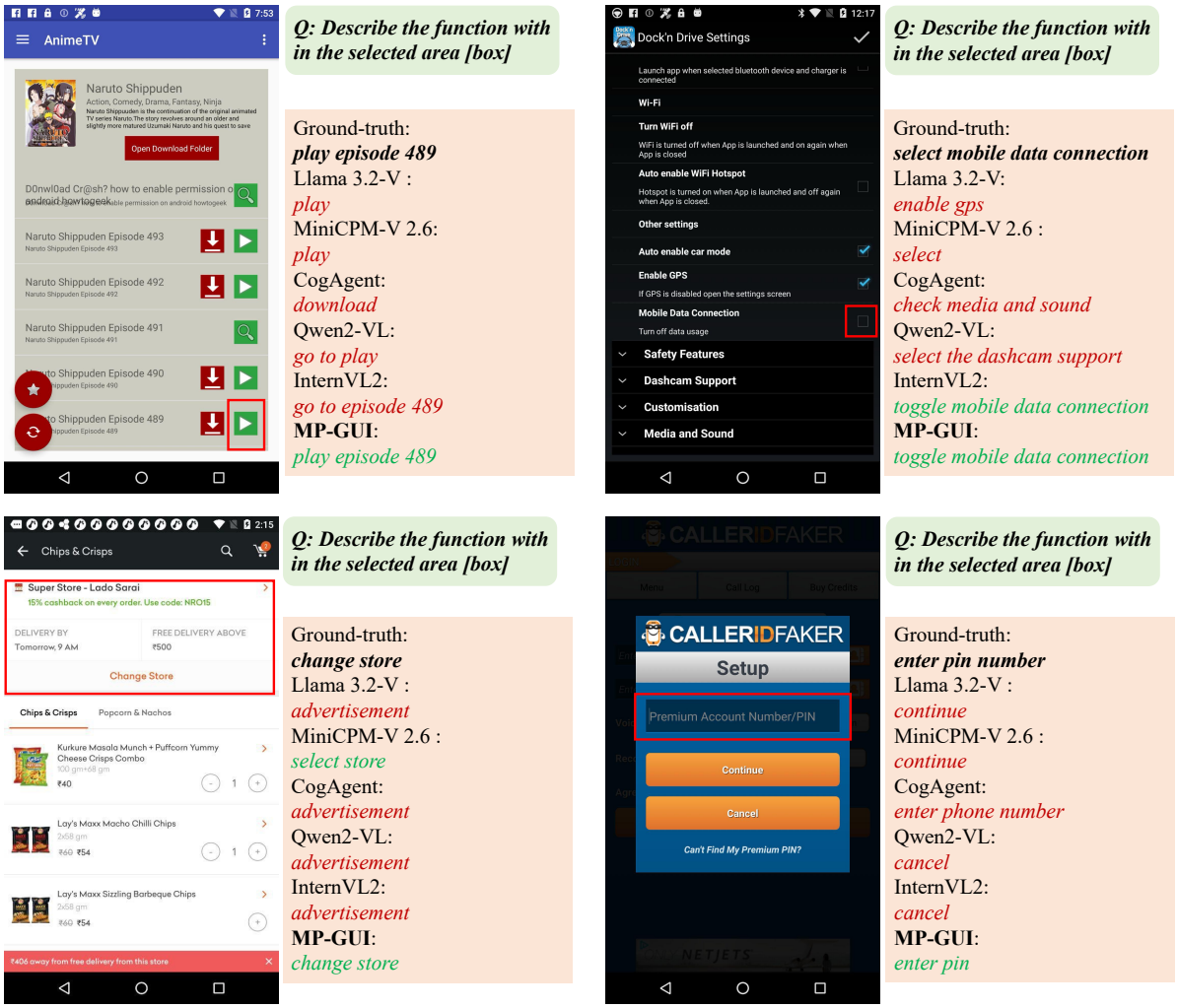
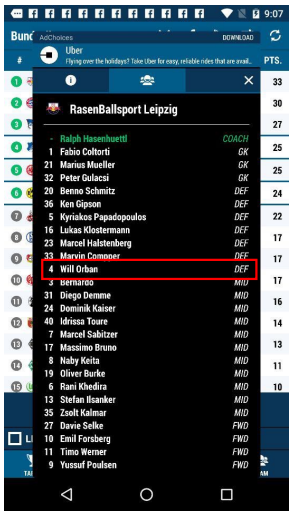


Figure 11. Comparisons on Widget Captioning [35].

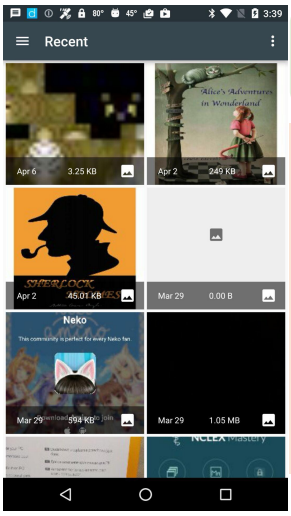
prehension capabilities. The Graphical Perceiver boosts the model's capacity to center on smaller areas. In contrast, Llama3.2-V(11B) [3], Qwen2-VL(7B) [55], and InternVL2(8B) [14] are influenced by the sizable '*12:30*' in the middle of the screen (as seen in Example 4). It is noteworthy that even when the question is unanswerable, as shown in the second example, our method still functions robustly.

Complex ScreenQA. The Spatial Perceiver enhances the awareness of spatial relationships between GUI elements on the screen. Compared with other MLLMs, our MP-GUI has advantages in difference calculation (as shown in Examples 1 and 4) and quantity counting (as shown in Examples 2 and 3) in Fig. 13. More qualitative results of our MP-GUI are shown in Fig. 14.



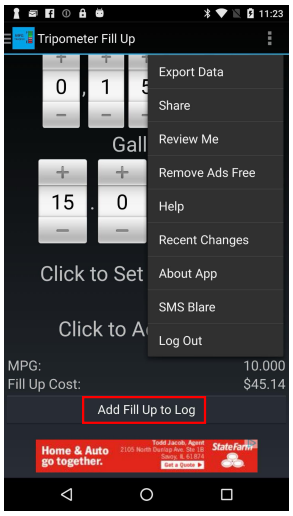
Q: What is the position of Will Orban?

Ground-truth:
DEF
Llama 3.2-V :
4th
MiniCPM-V 2.6 :
4
CogAgent:
DEF
Qwen2-VL:
24
InternVL2:
MID
MP-GUI:
DEF



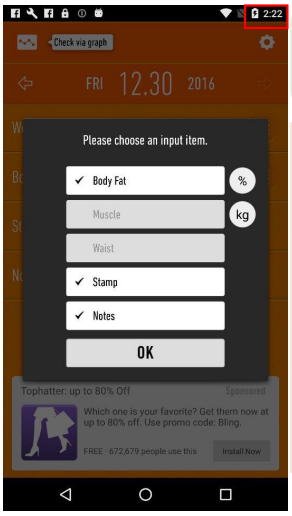
Q: On which date 694KB memory has shown?

Ground-truth:
no answer
Llama 3.2-V :
Mar 29
MiniCPM-V 2.6 :
Mar 29
CogAgent:
Mar 29
Qwen2-VL:
Mar 29
InternVL2:
Mar 29
MP-GUI:
no answer



Q: Where do we have to add fill up?

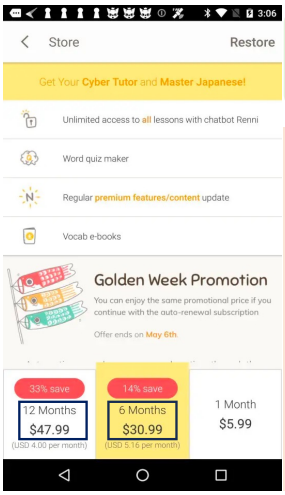
Ground-truth:
to "Log"
Llama 3.2-V :
no answer
MiniCPM-V 2.6 :
no answer
CogAgent:
click to add fill up
Qwen2-VL:
no answer
InternVL2:
no answer
MP-GUI:
Log



Q: What is the time?

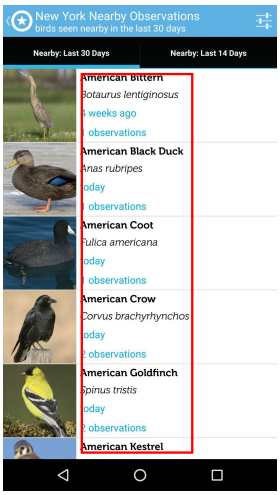
Ground-truth:
2:22
Llama 3.2-V :
12:30
MiniCPM-V 2.6 :
2:22 p.m.
CogAgent:
2:15
Qwen2-VL:
12:30
InternVL2:
12:30
MP-GUI:
2:22

Figure 12. Comparisons on ScreenQA Short [5].



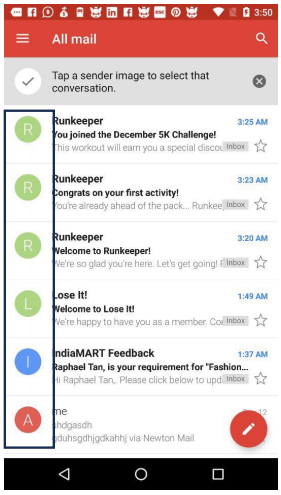
Q: How much more does the 12 month subscription cost than the 6 month subscription?

Ground-truth:
47.99 – $30.99 = $17.00
Llama 3.2-V :
$5.00
MiniCPM-V 2.6 :
$17.0
CogAgent:
$47.00
Qwen2-VL:
$37.00
InternVL2:
$17.99
MP-GUI:
$17.00



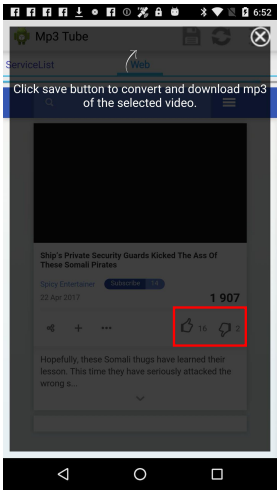
Q: How many birds have been seen nearby in the last 30 days?

Ground-truth:
6
Llama 3.2-V :
5
MiniCPM-V 2.6 :
7
CogAgent:
4
Qwen2-VL:
4
InternVL2:
5
MP-GUI:
6



Q: How many items are in the inbox?

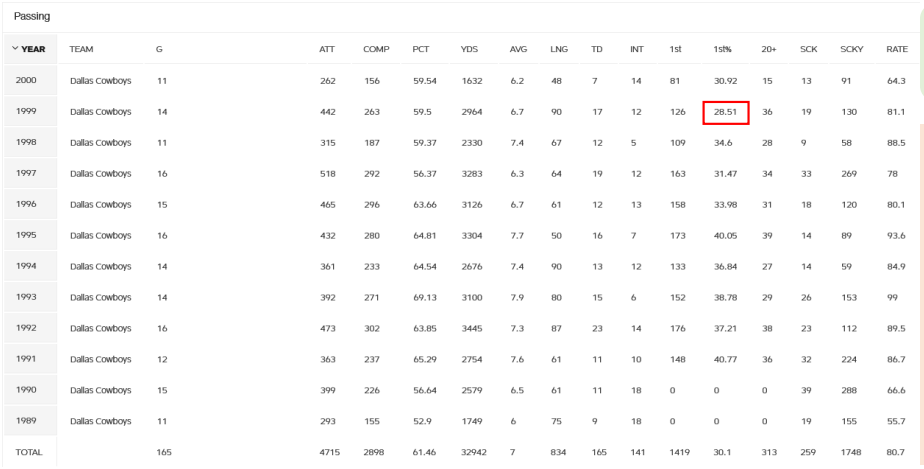
Ground-truth:
6
Llama 3.2-V :
5
MiniCPM-V 2.6 :
5
CogAgent:
5
Qwen2-VL:
3
InternVL2:
12
MP-GUI:
6



Q: How many more thumbs up than thumbs down are there on the video?

Ground-truth:
16 – 2 = 14
Llama 3.2-V :
3
MiniCPM-V 2.6 :
13
CogAgent:
16
Qwen2-VL:
16
InternVL2:
14
MP-GUI:
14

Figure 13. Comparisons on Complex ScreenQA [5].

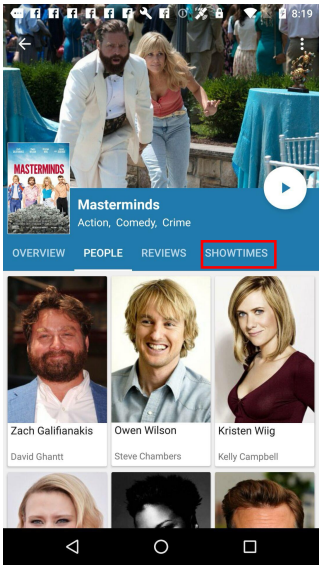


Passing

YEAR	TEAM	G	ATT	COMP	PCT	YDS	AVG	LNG	TD	INT	1st	1st%	20+	SCK	SCKY	RATE
2000	Dallas Cowboys	11	262	156	59.54	1632	6.2	48	7	14	81	30.92	15	13	91	64.3
1999	Dallas Cowboys	14	442	263	59.5	2964	6.7	90	17	12	126	28.51	36	19	130	81.1
1998	Dallas Cowboys	11	315	187	59.37	2330	7.4	67	12	5	109	34.6	28	9	58	88.5
1997	Dallas Cowboys	16	518	292	56.37	3283	6.3	64	19	12	163	31.47	34	33	269	78
1996	Dallas Cowboys	15	465	296	63.66	3126	6.7	61	12	13	158	33.98	31	18	120	80.1
1995	Dallas Cowboys	16	432	280	64.81	3304	7.7	50	16	7	173	40.05	39	14	89	93.6
1994	Dallas Cowboys	14	361	233	64.54	2676	7.4	90	13	12	133	36.84	27	14	59	84.9
1993	Dallas Cowboys	14	392	271	69.13	3100	7.9	80	15	6	152	38.78	29	26	153	99
1992	Dallas Cowboys	16	473	302	63.85	3445	7.3	87	23	14	176	37.21	38	23	112	89.5
1991	Dallas Cowboys	12	363	237	65.29	2754	7.6	61	11	10	148	40.77	36	32	224	86.7
1990	Dallas Cowboys	15	399	226	56.64	2579	6.5	61	11	18	0	0	0	39	288	66.6
1989	Dallas Cowboys	11	293	155	52.9	1749	6	75	9	18	0	0	0	19	155	55.7
TOTAL		165	4715	2898	61.46	32942	7	834	165	141	1419	30.1	313	259	1748	80.7

Q: In 1999, what was this player's 1st%?

Ground-truth:
28.51
Llama 3.2-V :
36
MiniCPM-V 2.6 :
126
CogAgent:
14
Qwen2-VL:
28.51
InternVL2:
126
MP-GUI:
28.51



Q: Is [box] tappable? Answer yes or no.

Ground-truth:
yes
Llama 3.2-V :
no
MiniCPM-V 2.6 :
no
CogAgent:
yes
Qwen2-VL:
yes
InternVL2:
no
MP-GUI:
yes

Figure 14. More qualitative results.