MP-GUI: Modality Perception with MLLMs for GUI Understanding

Ziwei Wang[1*] Weizhi Chen[1*] Leyang Yang[1] Sheng Zhou[1] Shengchu Zhao[2]
Hanbei Zhan[1] Jiongchao Jin[2] Liangcheng Li[1] Zirui Shao[1] Jiajun Bu[1]
1 Zhejiang University 2 Ant Group

CVPR Nashville JUNE 11-15, 2025

https://github.com/BigTaige/MP-GUI

{wangziwei98, zhousheng_zju}@zju.edu.cn

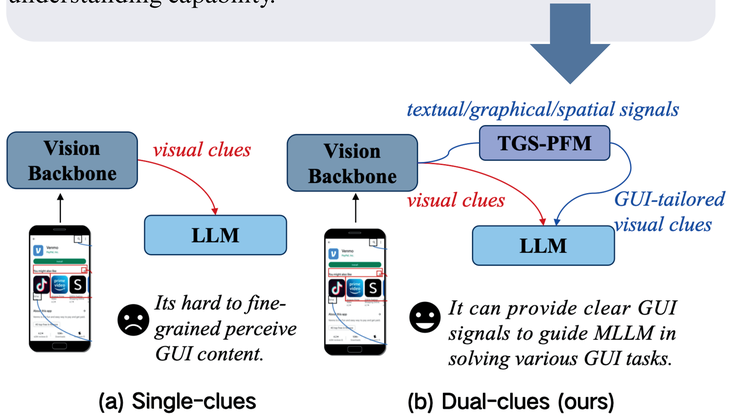
Background: GUI understanding

Challenge:

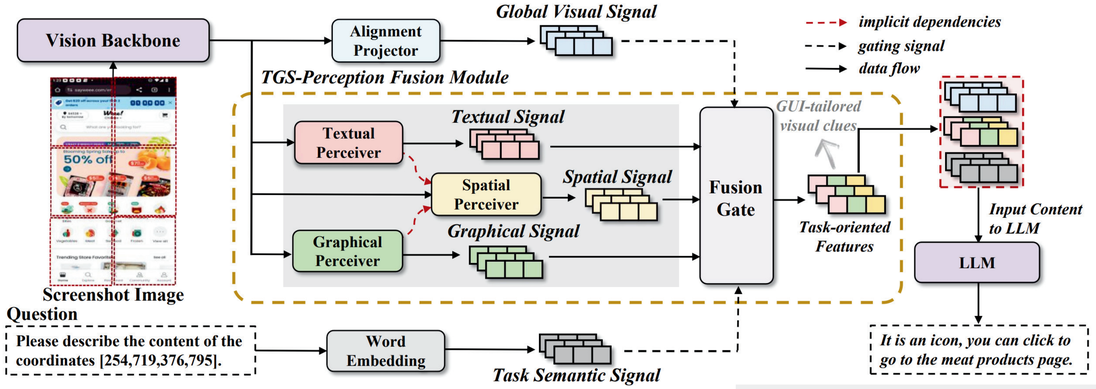
- Screen images with multimodal UI elements and complex and dense details increase the difficulty of grounding MLLM.
- The semantic associations between GUI elements need to be clearly identified, challenging MLLMs to grasp spatial GUI element relationships.

Introduction:

We propose **MP-GUI**, which **provides GUI-tailored visual clues for LLM via three perceivers and a semantically guided Fusion Gate**, endowing the MLLM with effective GUI perception and understanding capability.



Methodology: MP-GUI

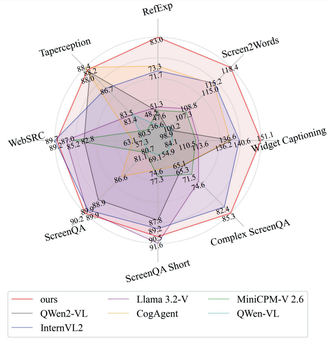


Overall of MP-GUI

- vision backbone providing visual clues of the screenshot.
- **TGS-Perception Fusion Module** including **three GUI-tailored perceivers** for extracting specific GUI modality signals and a **Fusion Gate** for dynamically fusing these signals based on task semantics to produce GUI-tailored visual clues.
- LLM generating results relying on screen visual clues, GUI-tailored visual clues, and task semantic signal.

Training Strategies:

- S1: Train the **Textual Perceiver** with **OCR data** .
- S2: Train the **Graphical Perceiver** with **graphical elements grounding data**.
- S3: Train the **Spatial Perceiver** with **constructed spatial relationship data**.
- S4: Train the **Fusion Gate** with **synthetic data**.



Experimental Results

Method	Size	Mobile		Desktop		Web		Avg.
		Text	Icon/Widget	Text	Icon/Widget	Text	Icon/Widget	
Llama 3.2-V [3]	11B	14.7%	5.7%	9.3%	4.3%	4.3%	4.4%	7.1%
GPT-4V[49]†	-	22.6%	24.5%	20.2%	11.8%	9.2%	8.8%	16.2%
Fuyu [9]†	8B	41.0%	1.3%	33.0%	3.6%	33.9%	4.4%	19.5%
InternVL2 [14]	8B	74.0 %	25.8%	54.6%	27.1%	38.3%	31.6%	41.9%
CogAgent [22]†	18B	67.0%	24.0%	**74.2%**	20.0%	**70.4%**	28.6%	47.4%
SeeClick [15]†	9.6B	78.0%	52.0%	72.2%	30.0%	55.7%	32.5%	53.4%
InternVL2-P	8B	83.2%	52.0%	63.4%	43.6%	47.0%	41.3%	55.1%
MP-GUI	8B	**86.8%**	**65.9%**	70.8%	**56.4%**	58.3%	**46.6%**	**64.1%**

Method	Cross-Task			Cross-Website			Cross-Domain		
	Ele.Acc	Op.F1	Step.SR	Ele.Acc	Op.F1	Step.SR	Ele.Acc	Op.F1	Step.SR
InternVL2 [14]	18.8	87.4	16.7	17.6	85.8	14.5	13.9	87.0	12.0
CogAgent [22]	22.4	53.0	17.6	18.4	42.4	13.4	20.6	42.0	15.5
SeeClick [15]	28.3	87.0	25.5	21.4	80.6	16.4	23.2	84.8	20.8
GPT-4 [2]	41.6	60.6	36.2	35.8	51.1	30.1	37.1	46.5	26.4
ShowUI [38]	39.7	88.0	36.9	**41.0**	83.6	**34.2**	**38.9**	85.3	**34.1**
InternVL2-P	27.4	87.8	24.0	27.4	86.1	23.1	24.3	87.1	21.1
MP-GUI	**42.1**	**89.0**	**38.1**	39.4	**87.1**	32.9	37.6	**87.4**	33.7

Method	General	Install	G.Apps	Single	WebShop	Overall
GPT-4V [57]	41.7	42.6	49.8	72.8	45.7	50.5
Qwen-VL [8]	49.5	59.9	46.9	64.7	50.7	54.3
OmniParser [45]	48.3	57.8	51.6	**77.4**	52.9	57.7
SeeClick [15]	54.0	66.4	54.9	63.5	57.6	59.3
InternVL2 [14]	58.1	65.3	56.8	68.7	61.1	62.0
ShowUI [38]	63.5	72.3	**66.0**	72.3	65.8	68.3
InternVL2-P	61.2	70.3	61.6	74.6	65.1	66.6
MP-GUI	**63.7**	**74.3**	65.3	75.4	**67.2**	**69.2**

Method	WC	S2W	RE	TP	WS	QA	QAS	CQA
w/o FG†	142.1 (-6.3%)	117.8 (-0.5%)	76.8 (-8.1%)	87.9 (-0.3%)	87.2 (-2.3%)	87.3 (-1.5%)	89.1 (-1.6%)	80.7 (-4.5%)
w/o FG‡	143.4 (-5.3%)	116.7 (-1.5%)	77.5 (-7.1%)	88.1 (-0.1%)	89.3 (+0.1%)	88.0 (-0.7%)	89.3 (-1.3%)	82.6 (-2.1%)
w/o TxP	142.6 (-5.9%)	115.2 (-2.8%)	78.8 (-5.3%)	88.3 (+0.1%)	89.1 (-0.1%)	87.5 (-1.3%)	89.4 (-1.2%)	80.8 (-4.3%)
w/o GaP	143.1 (-5.5%)	116.0 (-2.1%)	79.5 (-4.4%)	88.1 (-0.1%)	89.3 (+0.1%)	87.7 (-1.0%)	89.5 (-1.1%)	80.3 (-5.0%)
w/o SaP	141.9 (-6.4%)	116.2 (-1.9%)	78.4 (-5.9%)	88.2 (0.0%)	89.0 (-0.2%)	87.6 (-1.1%)	89.4 (-1.2%)	80.3 (-5.0%)
w/o MTS	148.3 (-1.8%)	117.0 (-1.2%)	82.4 (-0.7%)	87.2 (-1.1%)	86.9 (-2.6%)	87.4 (-1.4%)	88.3 (-2.4%)	83.5 (-1.0%)
MP-GUI	151.0	118.4	83.0	88.2	89.2	88.6	90.5	84.3