SpiritSight Agent: Advanced GUI Agent with One Look

Zhiyuan Huang[1], Ziming Cheng[1 2], Junting Pan[3], Zhaohui Hou[1], Mingjie Zhan[1]

[1]SenseTime Research, [2]Beijing University of Posts and Telecommunications, [3]Chinese University of Hong Kong

CVPR Nashville JUNE 11-15, 2025

Abstract

Graphical User Interface (GUI) agents show amazing abilities in assisting human-computer interaction, automating human user's navigation on digital devices. An ideal GUI agent is expected to achieve high accuracy, low latency, and compatibility for different GUI platforms. Recent vision-based approaches have shown promise by leveraging advanced Vision Language Models (VLMs). While they generally meet the requirements of compatibility and low latency, these vision-based GUI agents tend to have low accuracy due to their limitations in element grounding. To address this issue, we propose **SpiritSight**, a vision-based, end-to-end GUI agent that excels in GUI navigation tasks across various GUI platforms. First, we create a multi-level, large-scale, high-quality GUI dataset called **GUI-Lasagne** using scalable methods, empowering SpiritSight with robust GUI understanding and grounding capabilities. Second, we introduce the **Universal Block Parsing (UBP)** method to resolve the ambiguity problem in dynamic high-resolution of visual inputs, further enhancing SpiritSight's ability to ground GUI objects. Through these efforts, SpiritSight agent outperforms other advanced methods on diverse GUI benchmarks, demonstrating its superior capability and compatibility in GUI navigation tasks.

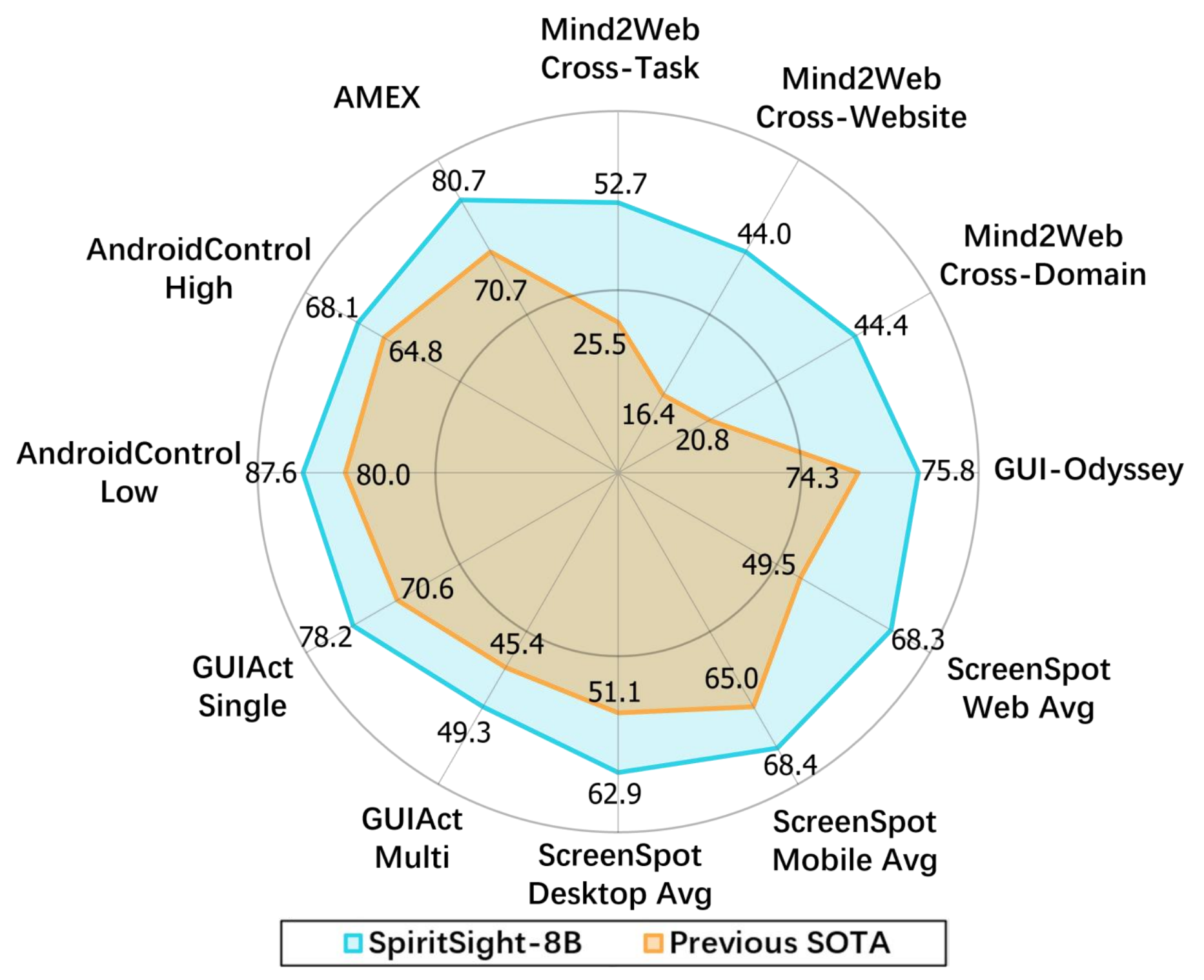


Fig 1. Overall Results

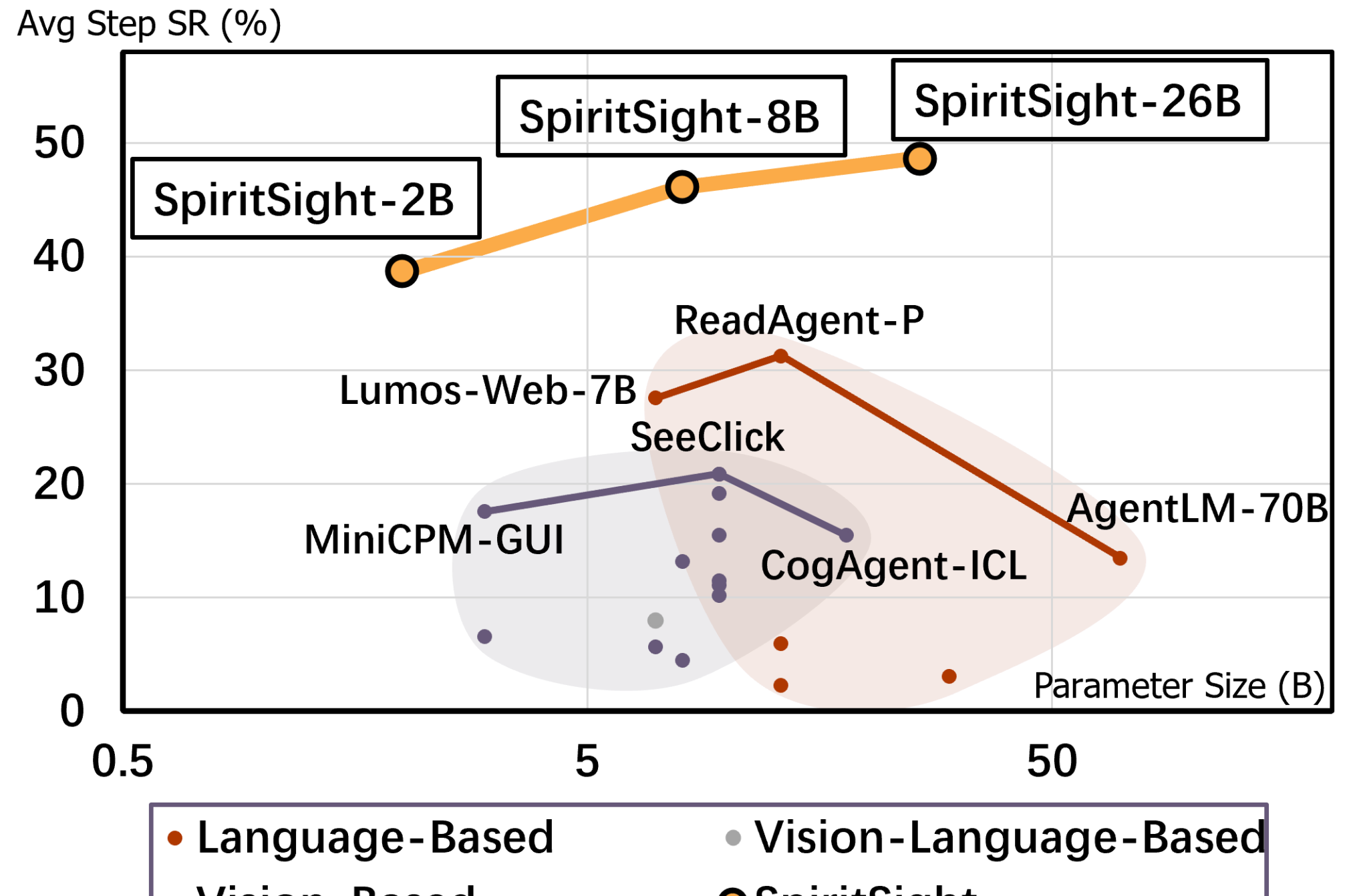
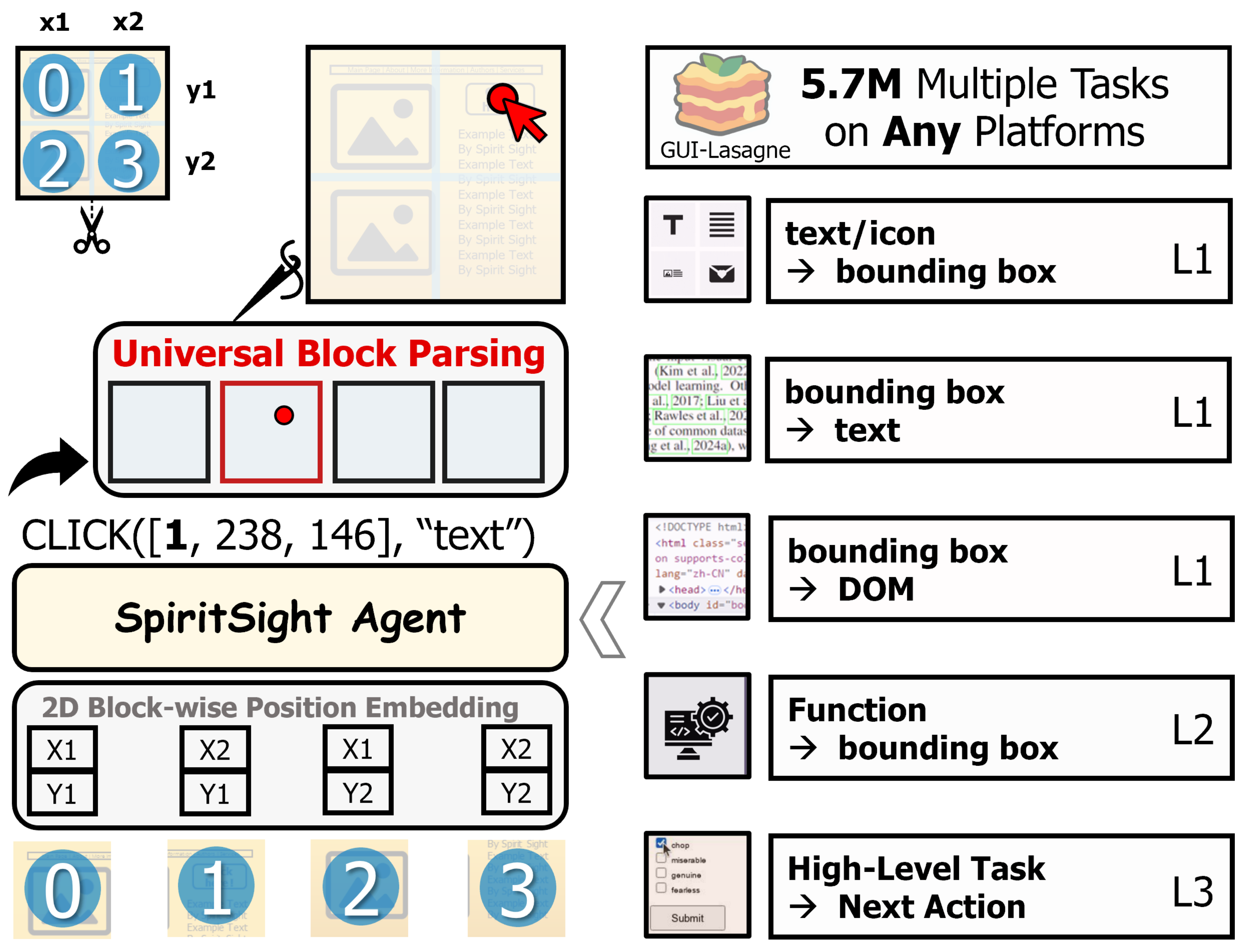


Fig 2. Comparing on Mind2Web



Universal Block Parsing

CLICK([**1**, 238, 146], "text")

SpiritSight Agent

2D Block-wise Position Embedding

X1 Y1 | X2 Y1 | X1 Y2 | X2 Y2

5.7M Multiple Tasks on **Any** Platforms

text/icon → bounding box L1

bounding box → text L1

bounding box → DOM L1

Function → bounding box L2

High-Level Task → Next Action L3

Fig 3. Overview of SpiritSight Agent's Solution

Three level of GUI-Lasagne Dataset

Level One: Visual-Text Alignment
1.9M Web / **1.1M** Mobile bbox

Level two: Visual-Function Alignment
1.5M function2bbox

Level three: Visual GUI Navigation
0.64M CoT-style + Opensource data

Universal Block Parsing enhance Grounding

Grounding ambiguity because of dynamic high-resolution strategy's **flattening operation**
→Solution:
2D Block-wise Position Embedding

Block-specific coordinate representation
[x, y] → [block_id, x, y]

Experiments

Advanced Vision-based GUI Agent

	Model	Input	Select	Cross-Task			Cross-Website			Cross-Domain		
	Size	Modality	From Top	Ele.Acc	Op.F1	Step SR	Ele.Acc	Op.F1	Step SR	Ele.Acc	Op.F1	Step SR
AutoWebGLM [26]	6B	Text	✓	-	-	66.4%	-	-	56.4%	-	-	55.8%
LLaMA2-7B [26]	7B	Text	✓	-	-	52.7%	-	-	47.1%	-	-	50.3%
CogAgent [19]	18B	Image	✓	-	-	62.3%	-	-	54.0%	-	-	59.4%
HTML-T5-XL [17]	3B	Text	✓	76.4%	78.8%	71.5%	68.4%	71.0%	62.2%	73.0%	75.6%	67.1%
SeeAct [74]	-	Text+Image	×	46.4%	73.4%	40.2%	38.0%	67.8%	32.4%	42.4%	69.3%	36.8%
ReadAgent-P [28]	340B	Text	×	33.7%	72.5%	29.2%	37.4%	75.1%	31.1%	37.2%	76.3%	33.4%
MiniCPM-GUI [7]	3B	Image	×	23.8%	86.8%	20.8%	20.3%	81.7%	17.3%	17.9%	74.5%	14.6%
Fuyu-GUI [4]	8B	Image	×	19.1%	86.1%	15.6%	13.9%	80.7%	12.2%	14.2%	83.1%	11.7%
SeeClick [11]	9.6B	Image	×	28.3%	87.0%	25.5%	21.4%	80.6%	16.4%	23.2%	84.8%	20.8%
OmniParser [53]	-	Image	×	42.4%	87.6%	39.4%	41.0%	84.8%	36.5%	45.5%	85.7%	42.0%
SpiritSight-2B	2B	Image	×	51.7%	87.2%	44.9%	44.0%	83.6%	37.8%	42.4%	83.5%	36.9%
SpiritSight-8B	8B	Image	×	59.2%	88.9%	52.7%	52.2%	84.7%	44.0%	50.1%	86.0%	44.4%
SpiritSight-26B	26B	Image	×	**60.5%**	**89.7%**	**54.7%**	**57.0%**	**85.7%**	**48.1%**	**54.1%**	**87.2%**	**49.2%**

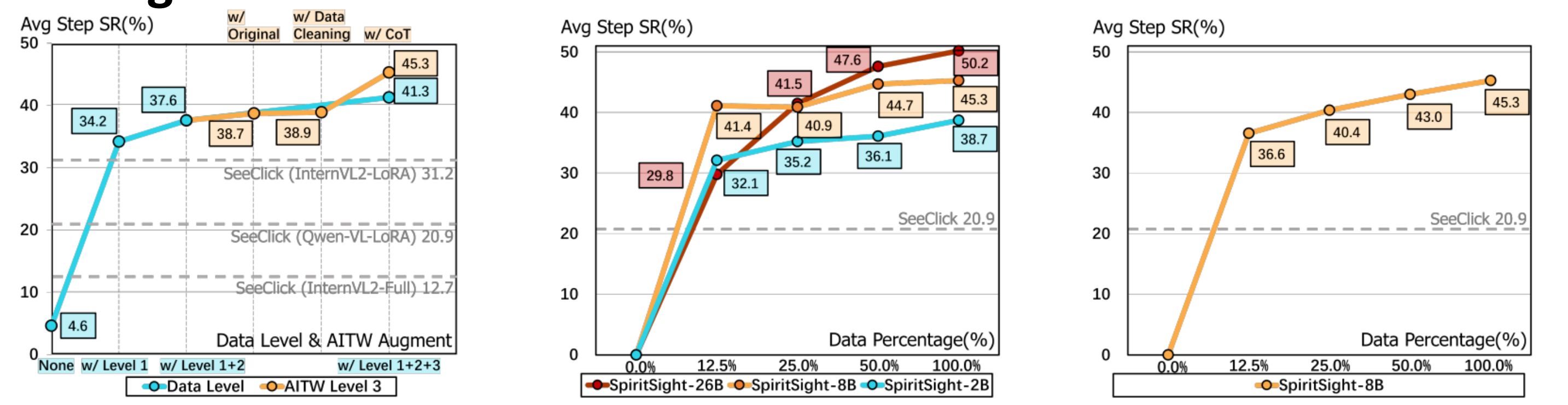
Strong Cross-Platform Compatibility

Agent	Odyssey	AMEX	AndroidCtrl		GUIAct	
	High	High	High	Low	Multi	Single
GPT-4o[42]	20.4%	-	21.2%	28.4%	-	41.8%
Previous SOTA	74.3%	70.7%	64.8%	80.0%	45.4%	74.9%
SpiritSight-2B	72.3%	74.5%	64.9%	86.3%	45.5%	76.0%
SpiritSight-8B	**75.8%**	**80.7%**	**68.1%**	**87.6%**	**49.3%**	**78.2%**

Strong Grounding Compatibility

Agent	Model	ScreenSpot		
	Size	Web	Mobile	Desktop
GPT4V[1]	-	5.0%	7.5%	4.6%
Qwen-VL[3]	9.6B	3.0%	7.2%	5.4%
Fuyu[4]	8B	19.2%	21.2%	18.3%
CogAgent[19]	18B	49.5%	45.5%	47.1%
SeeClick[11]	9.6B	44.1%	65.0%	51.1%
SpiritSight-2B	2B	63.6%	62.5%	61.8%
SpiritSight-8B	8B	**68.3%**	**68.4%**	**62.9%**

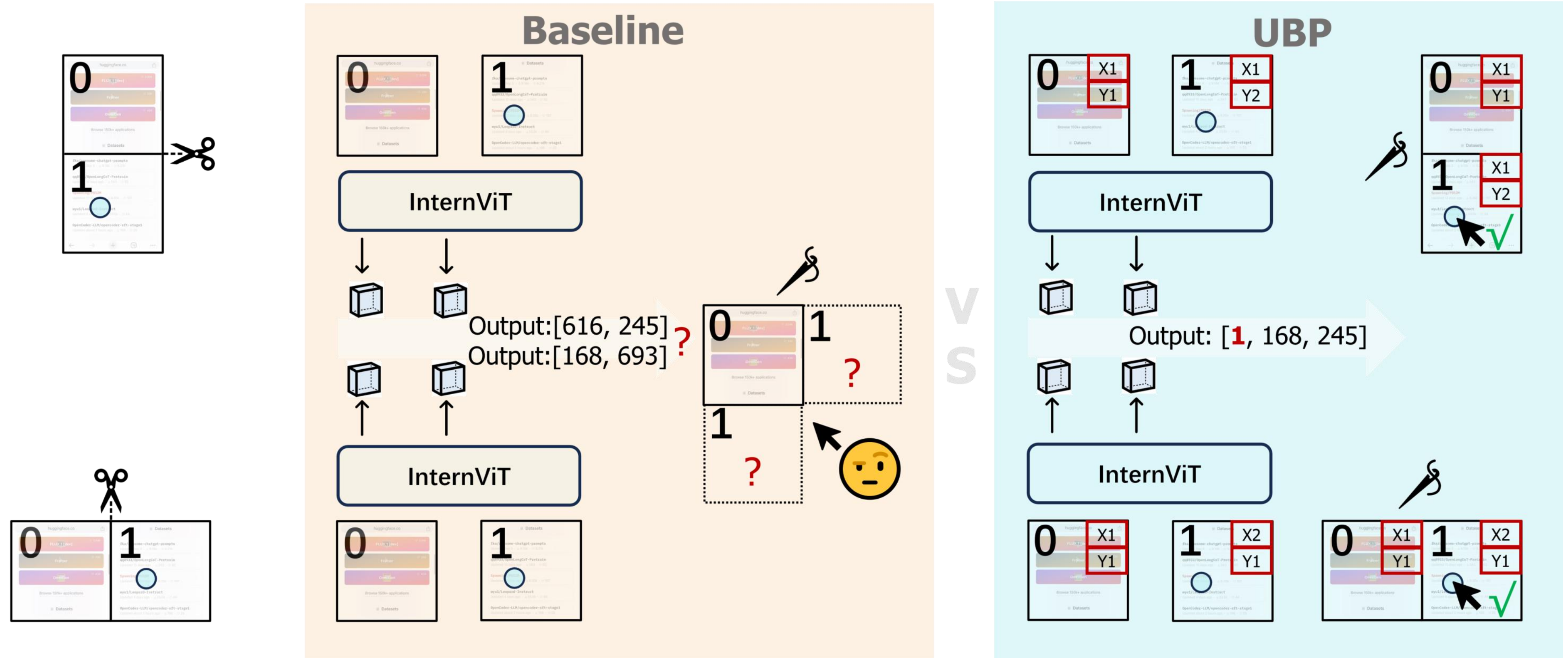
Scaling Effects on Dataset and Model Size



(a) Effect of three data levels and data augmentation for level-3 data

(b) Ablation Study: Effect of training data percentages on model performance.

(c) Ablation Study: Effect of data percentages on downstream task.



Output:[616, 245]
Output:[168, 693]

Output: [**1**, 168, 245]

Fig 4. Demonstration of Universal Block Parsing