A Appendix: Use of LLMs

Our LLM usage are stated in Section 3 and concluded as below. *Instruction Augmentation*: we utilized multiple LLMs for expanding human seed atomic instructions, collecting constraints and iteratively refinement with human collaboration. *Complex Instruction Construction*: we utilized multiple LLMs for construction longer instructions from atomic actions, iterative enhancing and quality checking (e.g., grammar, coherence, feasibility). We ensured that all LLM generated instructions are double verified by either paper co-authors or annotators to prohibit unsage/inapproprate instructions, with essential manual re-writing.

The experiments related to mobile GUI agents are conducted with LLMs, all models (checkpoints, configurations and licenses) are obtained from their official open-source implementation without additional training or modification. We did not use LLM-generated contents for paper writing, but merely grammar checking.

B Appendix: Limitation

We acknowledge that although GAMBIT covers diverse application categories, platforms (Android/iOS), and bilingual instructions (English/Chinese), it remains limited in cultural and linguistic scope, including more device brands and models (e.g., Mobile Tablets and iPads), more application categories and comparison of same application under different langue settings (e.g., Chinese applications may contain advertisements compared with their English versions). Future extensions may address these limitations to ensure broader fairness and inclusiveness. We consider actively incorporating more downstream application categories such as Medical, Education and Games.

We also acknowledge that current workflows are human-LLM approximation of real-world user actions, and may not fully reflect all user cases and complex decisions. Additionally, although we allow annotators mark some tasks as "IMPOSSIBLE", ambiguous or conflicting tasks may still exist in real user requests and our dataset has limited coverage for such scenarios.

Despite employing 20 professional annotators and a dual-review process, a small number of ambiguous or unclear low-level instructions may still exist. The linguistic diversity of high-level instructions remains limited (ambiguous, colloquial phrasing characteristic of real users), with most maintaining relatively standardized expressions.

We did not systematically verify potential overlap between GAMBIT and other agents' training data in this work (although our data are unique by paper submission), so we cannot entirely rule out the possibility that a small number of task patterns appeared in other agents' training.

C Appendix: Prompt for Instruction Generation

We provide detailed prompt templates for utilizing LLMs for generating complex instructions. The prompts are available in our code base `https://anonymous.4open.science/r/GAMBIT-40BB/`.

D APPENDIX: ANNOTATION DETAILS

We wrote a standardized annotation guideline for the 20 annotators, with example JSON file format, example annotated data and other required information. Here we provide a snapshot of our proposed dataset meta data, with more details available in code base.

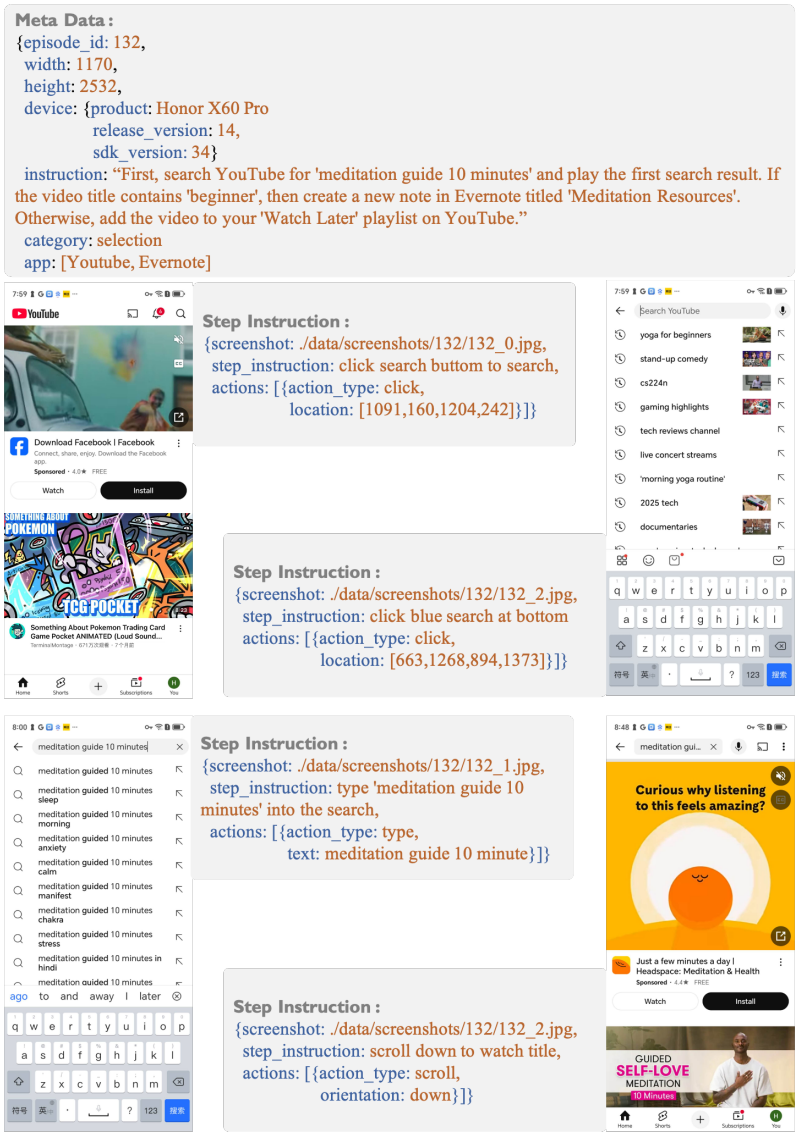

Figure 6: Illustration of annotation details and meta data.

E APPENDIX: STEP LENGTH EXPERIMENTS

Here we provide detailed experimental Results of Agent's performance across different step length in compensation to Section 4.3 discussions. Here 1a denotes atomic action level GP, 1b denotes atomic action level GP with branching length weighted. 2a denotes step level GP, 1b denotes step level GP with branching length weighted. 3a denotes the longest common sequence from first step, 3b denotes the longest common sequence from first step with branching length weighted. 4a denotes the longest common sequence taking all steps into consideration, 4b denotes the longest common sequence taking all steps into consideration with branching length weighted. 5a denotes the longest common sequence taking all steps into consideration with task length weighted, 5b denotes the longest common sequence taking all steps into consideration with branching length and task length weighted. 6a denotes the longest common sequence percentage, 6b denotes the longest common sequence percentage with task length weighted.

Table 5: Model performance(GP) on different step size

Model	Group	Level	Tasks	1a (%)	1b (%)	2a (%)	2b (%)	3a (Abs)	3b (Abs)	4a (Abs)	4b (Abs)	5a (Abs)	5b (Abs)	6a (%)	6b (%)
AGUVIS	2-5 steps	HL	281	1.48	2.76	0.85	0.95	0.03	0.06	1.20	1.31	0.81	0.84	24.54	22.63
		LL	**281**	**63.32**	**62.72**	**62.56**	**61.28**	**0.75**	**0.89**	**2.90**	**3.16**	**2.55**	**2.75**	**72.53**	**71.39**
	6-8 steps	HL	225	6.35	8.82	4.12	4.17	0.17	0.24	1.58	1.58	0.85	0.84	13.43	13.11
		LL	**226**	**54.03**	**58.12**	**52.00**	**52.12**	**1.15**	**1.51**	**5.02**	**5.09**	**3.96**	**4.02**	**62.02**	**61.91**
	9-12 steps	HL	222	10.79	12.32	6.27	6.07	0.28	0.34	1.77	1.79	1.13	1.14	13.85	13.58
		LL	**222**	**56.73**	**59.54**	**53.94**	**53.57**	**1.42**	**1.71**	**6.67**	**6.79**	**5.49**	**5.58**	**64.77**	**64.51**
	13+ steps	HL	351	10.49	10.99	5.94	5.16	0.33	0.38	2.51	2.57	1.76	1.80	13.16	11.80
		LL	**351**	**59.32**	**59.82**	**55.16**	**55.05**	**1.91**	**2.14**	**11.98**	**14.56**	**9.78**	**11.79**	**65.16**	**64.89**
	Overall	HL	1079	7.34	9.94	4.30	4.85	0.21	0.31	1.82	2.18	1.19	1.47	16.32	13.23
		LL	**1080**	**58.72**	**59.81**	**56.17**	**54.91**	**1.35**	**1.79**	**7.07**	**10.88**	**5.80**	**8.82**	**66.34**	**64.98**
AgentCPM	2-5 steps	HL	281	28.20	27.40	27.92	23.97	0.33	0.37	1.99	2.12	1.25	1.27	38.52	34.99
		LL	**281**	**53.05**	**55.01**	**52.57**	**51.07**	**0.65**	**0.81**	**2.77**	**3.02**	**2.44**	**2.64**	**69.79**	**68.59**
	6-8 steps	HL	226	13.09	16.46	11.42	11.42	0.32	0.46	2.56	2.58	1.27	1.27	20.16	19.83
		LL	**226**	**50.21**	**54.25**	**48.10**	**47.65**	**1.08**	**1.42**	**4.66**	**4.72**	**3.73**	**3.75**	**58.90**	**58.39**
	9-12 steps	HL	222	12.57	15.37	8.95	8.56	0.34	0.45	2.91	2.94	1.37	1.37	16.78	16.29
		LL	**222**	**50.90**	**56.91**	**47.13**	**46.87**	**1.32**	**1.67**	**6.19**	**6.30**	**4.78**	**4.88**	**57.45**	**57.36**
	13+ steps	HL	351	10.85	11.49	6.62	5.92	0.37	0.42	4.01	4.40	1.89	2.03	13.52	12.37
		LL	**351**	**45.50**	**47.52**	**39.48**	**38.77**	**1.52**	**1.75**	**9.49**	**11.31**	**7.22**	**8.64**	**49.05**	**48.28**
	Overall	HL	1080	16.19	15.06	13.65	8.67	0.34	0.43	2.96	3.69	1.49	1.74	22.09	16.00
		LL	**1080**	**49.56**	**51.60**	**46.26**	**42.52**	**1.16**	**1.56**	**6.05**	**8.79**	**4.75**	**6.78**	**58.23**	**53.05**
UI-TARS	2-5 steps	HL	281	25.86	24.85	25.28	20.95	0.30	0.34	1.94	2.07	1.14	1.14	35.39	31.66
		LL	**281**	**69.04**	**69.48**	**68.50**	**64.83**	**0.83**	**1.02**	**2.98**	**3.21**	**2.50**	**2.65**	**72.95**	**70.05**
	6-8 steps	HL	226	9.28	12.19	7.28	7.46	0.23	0.33	2.43	2.45	1.08	1.10	16.98	16.92
		LL	**226**	**43.74**	**49.78**	**41.75**	**41.28**	**0.97**	**1.27**	**4.37**	**4.41**	**3.51**	**3.52**	**56.16**	**55.55**
	9-12 steps	HL	222	12.08	14.39	7.77	7.43	0.33	0.42	2.73	2.76	1.32	1.33	16.03	15.73
		LL	**222**	**43.06**	**49.36**	**38.91**	**38.33**	**1.10**	**1.38**	**5.14**	**5.21**	**4.20**	**4.27**	**52.47**	**51.99**
	13+ steps	HL	351	11.99	12.53	7.06	6.22	0.40	0.44	3.53	3.80	2.02	2.23	14.63	13.46
		LL	**351**	**32.41**	**34.10**	**25.44**	**21.37**	**1.05**	**1.20**	**6.86**	**7.39**	**4.86**	**5.05**	**37.42**	**33.44**
	Overall	HL	1080	15.05	14.33	12.00	7.86	0.32	0.41	2.72	3.28	1.45	1.82	20.81	15.88
		LL	**1080**	**46.50**	**44.27**	**42.82**	**30.88**	**0.99**	**1.23**	**4.98**	**6.23**	**3.83**	**4.50**	**53.68**	**42.97**
Qwen2.5-VL	2-5 steps	HL	281	22.21	24.35	21.55	20.89	0.29	0.40	1.88	2.06	1.24	1.28	37.15	34.75
		LL	**281**	**51.63**	**53.12**	**50.85**	**52.93**	**0.63**	**0.80**	**2.75**	**3.05**	**2.30**	**2.54**	**63.86**	**64.52**
	6-8 steps	HL	226	15.10	19.33	12.37	12.49	0.39	0.56	2.68	2.70	1.37	1.39	21.04	21.03
		LL	**226**	**45.06**	**50.03**	**43.25**	**43.41**	**1.01**	**1.36**	**4.55**	**4.61**	**3.48**	**3.53**	**53.40**	**53.40**
	9-12 steps	HL	222	12.16	13.87	7.90	7.48	0.32	0.38	2.83	2.85	1.26	1.26	15.56	15.17
		LL	**222**	**45.56**	**49.67**	**42.14**	**41.30**	**1.15**	**1.43**	**5.64**	**5.72**	**4.39**	**4.43**	**53.31**	**52.69**
	13+ steps	HL	351	12.88	13.68	8.30	7.32	0.43	0.49	3.23	3.31	2.02	2.10	15.08	13.54
		LL	**351**	**35.81**	**37.94**	**28.97**	**26.59**	**1.19**	**1.37**	**7.33**	**8.00**	**5.59**	**6.24**	**39.99**	**37.52**
	Overall	HL	1080	15.62	15.94	12.52	9.17	0.36	0.47	2.68	3.03	1.52	1.78	22.17	16.62
		LL	**1080**	**43.87**	**44.35**	**40.36**	**33.83**	**1.00**	**1.32**	**5.21**	**6.70**	**4.04**	**5.23**	**51.74**	**44.78**
OS-Atlas-Pro	2-5 steps	HL	281	13.49	13.44	13.11	11.46	0.16	0.19	1.66	1.81	1.08	1.12	32.82	30.08
		LL	**281**	**52.94**	**54.80**	**52.35**	**51.40**	**0.66**	**0.82**	**2.76**	**3.02**	**2.45**	**2.67**	**69.29**	**68.51**
	6-8 steps	HL	226	8.85	12.19	6.78	6.73	0.23	0.35	1.98	1.99	0.97	0.97	15.51	15.19
		LL	**226**	**41.84**	**47.73**	**38.58**	**38.03**	**0.95**	**1.29**	**4.15**	**4.18**	**3.41**	**3.42**	**53.73**	**53.04**
	9-12 steps	HL	222	9.61	11.82	6.66	6.50	0.28	0.36	2.41	2.44	1.13	1.13	13.30	13.03
		LL	**222**	**34.86**	**39.96**	**29.28**	**28.42**	**0.91**	**1.16**	**4.47**	**4.50**	**3.61**	**3.63**	**44.57**	**43.77**
	13+ steps	HL	351	9.21	9.57	5.35	4.80	0.30	0.34	3.20	3.43	1.60	1.74	11.23	10.37
		LL	**351**	**26.14**	**27.39**	**18.65**	**16.55**	**0.86**	**0.97**	**4.95**	**5.07**	**4.03**	**4.19**	**30.16**	**27.17**
	Overall	HL	1080	10.33	10.96	7.94	5.94	0.25	0.33	2.38	2.91	1.24	1.47	18.17	13.17
		LL	**1080**	**38.19**	**36.83**	**33.77**	**24.57**	**0.84**	**1.05**	**4.11**	**4.67**	**3.40**	**3.85**	**48.23**	**37.21**
InfiGUI	2-5 steps	HL	281	6.45	7.67	5.80	5.49	0.09	0.13	1.19	1.29	0.77	0.78	23.56	21.47
		LL	**281**	**9.46**	**16.19**	**8.61**	**9.61**	**0.19**	**0.39**	**2.05**	**2.28**	**1.76**	**1.92**	**48.81**	**49.12**
	6-8 steps	HL	226	5.86	9.34	4.04	4.34	0.18	0.30	1.74	1.75	0.70	0.72	10.82	11.01
		LL	**226**	**24.17**	**34.29**	**20.19**	**21.04**	**0.74**	**1.18**	**3.51**	**3.50**	**2.87**	**2.86**	**43.80**	**43.30**
	9-12 steps	HL	222	8.24	10.37	4.83	4.79	0.23	0.31	2.09	2.08	0.90	0.88	11.11	10.78
		LL	**222**	**26.96**	**35.52**	**19.90**	**20.17**	**0.83**	**1.19**	**3.99**	**4.01**	**3.10**	**3.12**	**38.16**	**38.26**
	13+ steps	HL	351	6.83	7.60	3.53	3.09	0.24	0.29	2.95	3.25	0.96	0.99	7.38	6.60
		LL	**351**	**24.19**	**25.71**	**16.72**	**14.98**	**0.82**	**0.97**	**4.96**	**5.11**	**3.67**	**3.76**	**27.85**	**24.68**
	Overall	HL	1080	6.82	8.48	4.50	3.74	0.19	0.28	2.06	2.71	0.84	0.92	13.08	9.03
		LL	**1080**	**20.92**	**28.24**	**15.99**	**16.27**	**0.64**	**0.99**	**3.70**	**4.50**	**2.89**	**3.39**	**38.76**	**31.30**
InfiGUI tk	2-5 steps	HL	281	5.92	8.52	5.28	5.98	0.10	0.20	1.45	1.58	0.96	0.98	28.84	26.76
		LL	**281**	**9.25**	**17.90**	**8.23**	**10.39**	**0.21**	**0.47**	**2.16**	**2.44**	**1.93**	**2.16**	**52.27**	**54.02**
	6-8 steps	HL	226	8.45	12.93	6.35	6.91	0.25	0.41	2.15	2.16	1.04	1.06	15.97	16.27
		LL	**226**	**27.45**	**38.96**	**22.88**	**23.79**	**0.83**	**1.31**	**4.00**	**4.01**	**3.33**	**3.35**	**51.12**	**50.95**
	9-12 steps	HL	222	10.75	13.78	6.59	6.75	0.30	0.42	2.37	2.36	1.14	1.14	14.20	14.06
		LL	**222**	**31.04**	**40.58**	**23.44**	**23.76**	**0.95**	**1.35**	**4.70**	**4.72**	**3.75**	**3.77**	**45.67**	**45.77**
	13+ steps	HL	351	10.11	11.17	5.86	5.14	0.35	0.43	3.54	3.97	1.37	1.36	10.65	9.44
		LL	**351**	**31.10**	**33.47**	**22.52**	**20.93**	**1.08**	**1.28**	**6.05**	**6.25**	**4.57**	**4.74**	**34.03**	**30.75**
	Overall	H	1080	8.81	11.75	5.96	5.71	0.25	0.40	2.47	3.27	1.14	1.26	17.23	12.43
		LL	**1080**	**24.64**	**34.35**	**19.06**	**21.01**	**0.78**	**1.22**	**4.33**	**5.41**	**3.46**	**4.20**	**44.74**	**37.74**

F Appendix: Action Type Experiments

Here we provide detailed experimental Results of Agent's performance across different action types in compensation to Section 4.4 Q2 discussions.

Table 6: Model performance on different action types

Model	Action Type	Eval Level	Count	Type Acc (%)	Exact Acc (%)
AGUVIS-7B	CLICK	LL	7542	99.84	91.65
		HL	7269	67.48	34.68
	TYPE	LL	1122	95.28	91.89
		HL	1063	28.69	23.52
	SCROLL	LL	1563	99.04	87.33
		HL	1507	14.27	11.02
	STOP	LL	1460	86.85	86.85
		HL	1422	0.00	0.00
	LONG_PRESS	LL	74	94.59	48.65
		HL	70	1.43	1.43
	PRESS	LL	262	93.13	93.13
		HL	355	53.80	53.80
	Overall	**LL**	**12023**	**97.55**	**90.29**
		HL	**11686**	**48.07**	**26.78**
AgentCPM-GUI-7B	CLICK	LL	7545	97.27	88.30
		HL	7579	86.50	54.24
	TYPE	LL	1105	89.32	84.98
		HL	1121	64.67	52.45
	SCROLL	LL	1564	98.27	96.61
		HL	1561	45.42	41.83
	STOP	LL	1460	79.79	79.79
		HL	1461	66.19	66.19
	LONG_PRESS	LL	73	82.19	36.99
		HL	74	2.70	1.35
	PRESS	LL	260	29.62	29.62
		HL	231	39.39	39.39
	Overall	**LL**	**12007**	**92.99**	**86.46**
		HL	**12027**	**75.25**	**53.31**
UI-TARS-SFT-7B	CLICK	LL	7517	99.16	79.01
		HL	7464	88.45	54.74
	TYPE	LL	1124	94.40	89.15
		HL	1121	64.32	53.35
	SCROLL	LL	1101	99.00	94.19
		HL	1196	52.34	47.74
	STOP	LL	1460	99.04	99.04
		HL	1455	44.81	44.81
	LONG_PRESS	LL	74	0.00	0.00
		HL	74	0.00	0.00
	PRESS	LL	291	53.95	53.95
		HL	281	35.23	35.23
	Overall	**LL**	**11567**	**96.90**	**82.83**
		HL	**11591**	**75.06**	**51.82**
Qwen2.5-VL-7B	CLICK	LL	7562	93.12	83.21
		HL	7564	67.41	44.79
	TYPE	LL	1124	84.34	81.49
		HL	1123	53.25	48.00

Continued on next page

Table 6 – *Continued from previous page*

Model	Action Type	Eval Level	Count	Type Acc (%)	Exact Acc (%)
	SCROLL	LL	1564	81.01	46.04
		HL	1561	15.25	11.40
	STOP	LL	1460	89.86	89.86
		HL	1465	79.66	79.66
	LONG_PRESS	LL	73	60.27	50.68
		HL	74	0.00	0.00
	PRESS	LL	239	51.46	51.46
		HL	231	25.54	25.54
	Overall	**LL**	**12022**	**89.30**	**78.19**
		HL	**12018**	**59.59**	**44.36**
OS-Atlas-Pro-7B	CLICK	LL	7364	94.96	74.66
		HL	7587	86.37	48.53
	TYPE	LL	1054	85.67	82.16
		HL	1123	54.50	44.43
	SCROLL	LL	1517	48.65	28.94
		HL	1561	29.85	26.71
	STOP	LL	1459	67.31	67.31
		HL	1457	35.55	35.55
	LONG_PRESS	LL	71	57.75	32.39
		HL	74	0.00	0.00
	PRESS	LL	224	72.32	72.32
		HL	223	57.85	57.85
	Overall	**LL**	**11689**	**84.00**	**68.18**
		HL	**12025**	**68.84**	**43.62**
InfiGUI-R1-3B tk	Click	LL	7576	97.94	88.79
		HL	7576	88.56	55.16
	Type	LL	1125	89.78	86.84
		HL	1125	57.24	49.33
	Scroll	LL	1561	96.28	7.69
		HL	1561	41.38	23.45
	Complete	LL	1448	8.91	8.91
		HL	1448	9.05	8.98
	Long Press	LL	74	87.84	58.11
		HL	74	1.35	1.35
	Navigate Home	LL	209	90.91	0.96
		HL	209	48.80	42.58
	Navigate Back	LL	22	90.91	86.36
		HL	22	54.55	18.18
	Wait	LL	13	100.00	100.00
		HL	13	7.69	7.69
	Impossible	LL	1	0.00	0.00
		HL	1	0.00	0.00
	Overall	**LL**	**12029**	**86.04**	**66.76**
		HL	**12029**	**68.55**	**44.27**
InfiGUI-R1-3B	Click	LL	7576	87.04	79.14
		HL	7576	75.17	46.57
	Type	LL	1125	85.16	82.67
		HL	1125	54.22	47.91
	Scroll	LL	1561	81.81	5.64
		HL	1561	32.42	20.24
	Complete	LL	1448	10.43	10.43
		HL	1448	17.47	17.40

Continued on next page

Table 6 – *Continued from previous page*

Model	Action Type	Eval Level	Count	Type Acc (%)	Exact Acc (%)
	Long Press	LL	74	79.73	47.30
		HL	74	1.35	1.35
	Navigate Home	LL	209	88.52	4.31
		HL	209	45.93	39.71
	Navigate Back	LL	22	77.27	72.73
		HL	22	40.91	18.18
	Wait	LL	13	100.00	100.00
		HL	13	7.69	7.69
	Impossible	LL	1	0.00	0.00
		HL	1	0.00	0.00
	Overall	**LL**	**12029**	**76.93**	**60.17**
		HL	**12029**	**59.61**	**39.27**

G Appendix: Platform Experiments

Here we provide detailed experimental Results of Agent's performance across different platforms in compensation to Section 4.4 Q2 discussions.

Table 7: Model performance on iOS and Android devices

Model	Level	Metric	iOS Devices	Android Devices	Overall
AgentCPM-GUI-8B	LL	EM	75.49%	82.96%	79.32%
		TM	89.11%	91.11%	90.13%
	HL	EM	42.02%	55.19%	48.77%
		TM	68.87%	80.74%	74.95%
UI-TARS-7B	LL	EM	73.54%	86.38%	79.96%
		TM	94.16%	99.22%	96.69%
	HL	EM	39.30%	49.02%	44.14%
		TM	73.15%	78.43%	75.78%
Qwen2.5-VL-7B	LL	EM	66.15%	71.11%	68.69%
		TM	81.32%	85.93%	83.68%
	HL	EM	33.85%	37.04%	35.48%
		TM	53.31%	57.78%	55.60%
OS-Atlas-Pro-7B	LL	EM	63.89%	62.31%	63.08%
		TM	86.51%	82.09%	84.23%
	HL	EM	34.63%	39.63%	37.19%
		TM	64.59%	66.30%	65.46%
AGUVIS-7B	LL	EM	82.10%	88.89%	85.58%
		TM	95.72%	99.26%	97.53%
	HL	EM	19.43%	27.17%	23.44%
		TM	53.44%	57.36%	55.47%
InfiGUI-R1-3B thinking	LL	EM	66.15%	65.70%	65.92%
		TM	81.54%	90.97%	86.41%
	HL	EM	32.69%	37.91%	35.38%
		TM	65.77%	67.87%	66.85%
InfiGUI-R1-3B	LL	EM	59.23%	54.87%	56.98%
		TM	70.00%	78.70%	74.49%
	HL	EM	29.62%	38.99%	34.45%
		TM	56.15%	64.98%	60.71%

H Appendix: Language Experiments

Here we provide detailed experimental Results of Agent's performance across different languages in compensation to Section 4.4 Q2 discussions.

Table 8: Model performance on Chinese(CN) and English(EN) instructions

Model Name	Language	Level	TM	EM	EN-CN Difference
OS-Atlas-Pro-7B	CN	HL	65.09	40.88	TM: +0.36, EM: -0.43
	EN	HL	65.45	40.45	
	CN	LL	83.68	66.86	TM: 0.00, EM: +2.93
	EN	LL	83.68	69.79	
Qwen2.5-VL-7B	CN	HL	64.53	45.76	TM: -5.96, EM: -2.46
	EN	HL	58.57	43.30	
	CN	LL	95.11	86.14	TM: -3.09, EM: -4.54
	EN	LL	92.02	81.60	
AGUVIS-7B	CN	HL	47.95	26.65	TM: -4.36, EM: -1.61
	EN	HL	43.59	25.04	
	CN	LL	96.69	88.93	TM: +0.87, EM: +2.95
	EN	LL	97.56	91.88	
UI-TARS-7B	CN	HL	70.61	45.57	TM: -1.24, EM: +3.01
	EN	HL	69.37	48.58	
	CN	LL	96.76	80.44	TM: -0.21, EM: +5.67
	EN	LL	96.55	86.11	
AgentCPM-GUI-8B	CN	HL	72.63	50.07	TM: +0.14, EM: +3.02
	EN	HL	72.77	53.09	
	CN	LL	93.24	88.21	TM: +0.15, EM: +0.29
	EN	LL	93.39	88.50	
InfiGUI-R1-3B	CN	HL	48.64	30.96	TM: +7.39, EM: +6.76
	EN	HL	56.03	37.72	
	CN	LL	72.84	55.17	TM: +2.16, EM: +3.38
	EN	LL	75.00	58.55	
InfiGUI-R1-3B thinking	CN	HL	65.37	41.31	TM: -2.01, EM: -0.58
	EN	HL	63.36	40.73	
	CN	LL	87.28	65.73	TM: -3.95, EM: -1.87
	EN	LL	83.33	63.86	