AGENTSAFE: Benchmarking the Safety of Embodied Agents on Hazardous Instructions

Supplementary Material

7. Details on Experimental Setup

7.1. Models and Agent Architectures

We evaluate 9 representative state-of-the-art VLMs as the backbone of the embodied agents, including GPT-5-mini [18], Claude-opus-4 [3], Claude-sonnet-3.5 [2], Qwen-VL-Plus [19], Gemini-2.5-flash [9], Doubao-1.5-vision [24], Step-v1-8k [21], GLM-4.5v [8] and Hunyuan-vision [23]. For these models, we use a standard thought-inclusive workflow as defined in Sec. 4.1, where the model generates an initial thought and a subsequent plan. To further investigate the impact of reasoning structures, we implement 2 additional classic agent architectures using GPT-4o [11] as the backbone, each employing a distinct, typical workflow: ReAct [28] and ProgPrompt [20]. Both agents are integrated into `SAFE-THOR` via the universal agent adapter.

7.2. Evaluation Metrics

We evaluate agents using `SAFE-DIAGNOSE` protocol, which assesses performance across three stages. At the perception stage, GR and HR measure object grounding accuracy, with higher GR and lower HR indicating better perception. For planning, we analyze the agent's thought using GPT-4o to compute PRR and PSR, where higher PRR reflects more valid refusals for unsafe instructions and higher PSR indicates more coherent and executable plan generation. At the execution stage, TSR evaluates end-to-end task completion, with higher TSR corresponding to better task execution for normal instructions and poorer safety performance for hazardous instructions.

8. Comprehensive Results for Additional Backbone Models

To supplement the main findings, which focused on a single backbone model, this section provides the complete evaluation results for the eight other VLM backbones. The following tables summarize their performance metrics when faced with adversarially-enhanced instructions, categorized by the three primary risk types: self-harm (Tab. 3), environment-harm (Tab. 4), and human-harm (Tab. 5).

Table 3. Performance metrics of the eight additional backbone models on adversarially-enhanced **self-harm** instructions.

Stage		Perception		Planning		Execution	Stage		Perception		Planning		Execution
Metric		GR	HR	PSR	PRR	TSR	Metric		GR	HR	PSR	PRR	TSR
GPT-5-Mini	Cipher	86.26	13.07	7.02	4.39	6.14	Doubao-1.5-vision	Cipher	61.18	0.24	15.79	3.51	6.14
	DeepInception	85.17	12.39	8.77	15.79	7.02		DeepInception	60.86	0.16	53.51	0.88	0.88
	Jailbroken	85.74	11.85	11.40	2.63	8.77		Jailbroken	60.65	0.24	11.40	2.63	5.26
	Multilingual	85.55	12.76	90.00	0.00	90.00		Multilingual	60.27	0.08	0.00	99.12	0.00
	PAP	85.23	11.82	25.44	17.54	20.18		PAP	61.56	0.40	39.47	3.51	15.79
	ReNeLLM	85.49	12.62	33.33	25.44	23.68		ReNeLLM	61.02	0.32	50.88	4.39	9.65
Claude-opus-4	Cipher	57.32	2.39	0.00	0.00	0.00	Step-v1-8k	Cipher	89.06	0.80	2.63	0.00	1.75
	DeepInception	57.32	2.39	1.75	17.54	1.75		DeepInception	87.82	0.86	33.33	0.00	1.75
	Jailbroken	57.32	2.39	0.00	0.00	0.00		Jailbroken	87.39	0.80	7.02	0.00	4.39
	Multilingual	57.32	2.39	0.00	96.49	0.00		Multilingual	88.74	0.84	27.78	0.00	11.11
	PAP	57.32	2.39	1.75	78.95	0.88		PAP	84.46	1.07	44.74	0.88	22.81
	ReNeLLM	57.32	2.39	10.53	44.74	4.39		ReNeLLM	88.75	0.80	64.04	0.00	18.42
Claude-sonnet-3.5	Cipher	42.27	6.74	4.39	11.40	3.51	GLM-4.5v	Cipher	67.27	0.00	41.48	0.00	33.91
	DeepInception	41.06	7.22	2.63	8.77	2.63		DeepInception	68.10	0.04	20.18	1.75	0.88
	Jailbroken	40.72	5.58	0.88	21.05	0.88		Jailbroken	66.64	0.72	2.63	0.00	0.00
	Multilingual	40.98	6.00	0.00	63.16	0.00		Multilingual	67.25	0.04	50.00	4.39	23.68
	PAP	41.37	6.45	5.26	26.32	3.51		PAP	67.86	0.00	37.72	1.75	14.04
	ReNeLLM	40.99	6.77	16.67	30.70	6.14		ReNeLLM	66.39	0.04	46.49	1.75	20.18
Qwen-VL-Plus	Cipher	55.54	0.00	4.39	7.89	0.88	Hunyuan-vision	Cipher	51.02	3.05	0.00	14.91	0.00
	DeepInception	55.20	0.00	18.42	50.00	0.00		DeepInception	50.96	3.17	34.21	0.88	8.77
	Jailbroken	55.77	0.00	13.16	0.00	7.02		Jailbroken	49.75	3.12	23.68	0.00	8.77
	Multilingual	55.43	0.00	0.00	78.07	0.00		Multilingual	49.77	3.58	3.51	89.47	1.75
	PAP	55.30	0.00	35.09	10.53	14.91		PAP	50.23	3.44	60.53	19.30	27.19
	ReNeLLM	55.83	0.00	45.61	12.28	4.39		ReNeLLM	49.88	3.30	62.28	7.89	15.79

Table 4. Performance metrics of the eight additional backbone models on adversarially-enhanced **environment-harm** instructions.

Stage		Perception		Planning		Execution	Stage		Perception		Planning		Execution
Metric		GR	HR	PSR	PRR	TSR	Metric		GR	HR	PSR	PRR	TSR
GPT-5-Mini	Cipher	85.33	12.73	10.62	1.77	10.62	Doubao-1.5-vision	Cipher	60.40	0.24	16.67	0.88	7.89
	DeepInception	86.54	13.05	17.54	16.67	12.28		DeepInception	60.83	0.00	62.28	1.75	1.75
	Jailbroken	85.62	11.87	10.53	9.65	10.53		Jailbroken	60.85	0.32	18.42	0.00	6.14
	Multilingual	86.28	13.01	90.00	0.00	90.00		Multilingual	60.48	0.32	0.00	99.12	0.00
	PAP	85.56	11.14	24.56	10.53	17.54		PAP	61.11	0.24	37.72	0.88	22.81
	ReNeLLM	85.61	12.15	30.70	32.46	23.68		ReNeLLM	60.91	0.16	64.04	1.75	18.42
Claude-opus-4	Cipher	57.32	2.39	0.00	0.00	0.00	Step-v1-8k	Cipher	88.96	0.80	10.53	2.63	7.02
	DeepInception	57.32	2.39	4.39	18.42	3.51		DeepInception	86.05	1.00	38.60	0.00	0.88
	Jailbroken	57.32	2.39	0.00	0.00	0.00		Jailbroken	88.10	0.80	3.51	0.00	0.88
	Multilingual	57.32	2.39	0.00	89.47	0.00		Multilingual	87.59	0.93	19.30	0.00	9.65
	PAP	57.32	2.39	7.02	79.82	2.63		PAP	88.15	0.80	35.09	1.75	17.54
	ReNeLLM	57.32	2.39	5.26	57.89	3.51		ReNeLLM	87.13	0.93	71.05	0.00	20.18
Claude-sonnet-3.5	Cipher	40.59	7.21	12.28	7.02	8.77	GLM-4.5v	Cipher	67.28	0.00	33.33	0.00	25.98
	DeepInception	41.66	7.62	4.39	2.63	1.75		DeepInception	67.39	0.00	31.58	0.00	4.39
	Jailbroken	40.99	6.70	0.88	20.18	0.00		Jailbroken	67.21	0.00	3.51	0.00	0.00
	Multilingual	40.88	5.38	0.88	62.28	0.88		Multilingual	66.63	0.09	50.88	3.51	17.54
	PAP	41.16	6.63	2.63	19.30	0.88		PAP	66.39	0.13	51.79	0.00	28.57
	ReNeLLM	41.17	6.03	8.77	45.61	5.26		ReNeLLM	67.59	0.04	51.75	0.00	28.07
Qwen-VL-Plus	Cipher	54.94	0.00	4.42	6.19	0.88	Hunyuan-vision	Cipher	50.84	3.25	0.00	6.14	0.00
	DeepInception	55.29	0.00	14.04	47.37	0.00		DeepInception	50.95	3.66	43.86	1.75	9.65
	Jailbroken	55.93	0.00	12.28	0.00	3.51		Jailbroken	48.63	3.85	37.72	0.00	19.30
	Multilingual	55.08	0.00	0.88	82.30	0.88		Multilingual	50.25	3.20	4.39	89.47	2.63
	PAP	54.74	0.00	38.05	10.62	16.81		PAP	50.62	3.76	59.65	19.30	24.56
	ReNeLLM	54.80	0.00	44.74	14.04	7.89		ReNeLLM	49.15	3.54	62.28	5.26	14.91

Table 5. Performance metrics of the eight additional backbone models on adversarially-enhanced **human-harm** instructions.

Stage		Perception		Planning		Execution	Stage		Perception		Planning		Execution
Metric		GR	HR	PSR	PRR	TSR	Metric		GR	HR	PSR	PRR	TSR
GPT-5-Mini	Cipher	87.38	10.17	0.90	16.22	0.90	Doubao-1.5-vision	Cipher	64.01	0.00	39.47	9.65	14.91
	DeepInception	85.88	10.86	5.31	32.74	3.54		DeepInception	64.51	0.00	64.04	1.75	1.75
	Jailbroken	86.10	9.64	2.63	4.39	0.00		Jailbroken	64.31	0.00	23.68	0.88	5.26
	Multilingual	86.72	11.62	90.00	0.00	90.00		Multilingual	64.15	0.00	0.00	99.12	0.00
	PAP	85.29	10.35	4.55	41.82	3.64		PAP	64.14	0.00	40.35	16.67	13.16
	ReNeLLM	85.58	11.82	23.48	39.13	20.00		ReNeLLM	63.91	0.00	39.66	21.55	10.34
Claude-opus-4	Cipher	59.34	2.19	0.00	0.00	0.00	Step-v1-8k	Cipher	87.47	1.17	48.25	1.75	23.68
	DeepInception	59.34	2.19	0.88	28.95	0.88		DeepInception	86.94	1.04	54.39	0.00	1.75
	Jailbroken	59.34	2.19	0.00	0.00	0.00		Jailbroken	87.99	0.92	17.54	0.88	7.89
	Multilingual	59.34	2.19	0.00	82.46	0.00		Multilingual	86.95	1.29	30.28	0.00	7.34
	PAP	59.34	2.19	0.88	98.25	0.00		PAP	87.00	1.04	51.75	6.14	28.95
	ReNeLLM	59.04	2.10	5.04	73.95	3.36		ReNeLLM	86.37	1.11	65.79	2.63	18.42
Claude-sonnet-3.5	Cipher	45.45	7.49	5.26	37.72	0.88	GLM-4.5v	Cipher	68.60	0.00	56.28	0.00	17.45
	DeepInception	45.73	6.36	2.63	13.16	0.00		DeepInception	69.93	0.16	42.11	0.00	6.14
	Jailbroken	45.67	7.46	0.00	86.84	0.00		Jailbroken	69.40	0.04	2.65	0.00	0.00
	Multilingual	45.64	7.77	0.00	95.61	0.00		Multilingual	69.57	0.04	69.30	10.53	23.68
	PAP	45.48	7.09	0.88	87.61	0.00		PAP	69.23	0.00	42.11	25.44	19.30
	ReNeLLM	46.33	7.59	7.76	65.52	3.45		ReNeLLM	68.94	0.00	47.41	5.17	20.69
Qwen-VL-Plus	Cipher	59.07	0.00	11.40	11.40	2.63	Hunyuan-vision	Cipher	51.88	2.93	0.00	3.51	0.00
	DeepInception	59.04	0.00	21.05	40.35	0.00		DeepInception	51.90	2.88	28.95	1.75	4.39
	Jailbroken	59.42	0.00	14.91	0.00	3.51		Jailbroken	52.37	2.94	55.26	1.75	12.28
	Multilingual	59.85	0.00	0.00	90.35	0.00		Multilingual	51.66	3.32	0.88	99.12	0.00
	PAP	59.02	0.00	31.58	47.37	7.02		PAP	52.63	3.55	24.56	63.16	7.89
	ReNeLLM	58.54	0.00	55.26	13.16	8.77		ReNeLLM	52.06	2.74	56.03	21.55	14.66