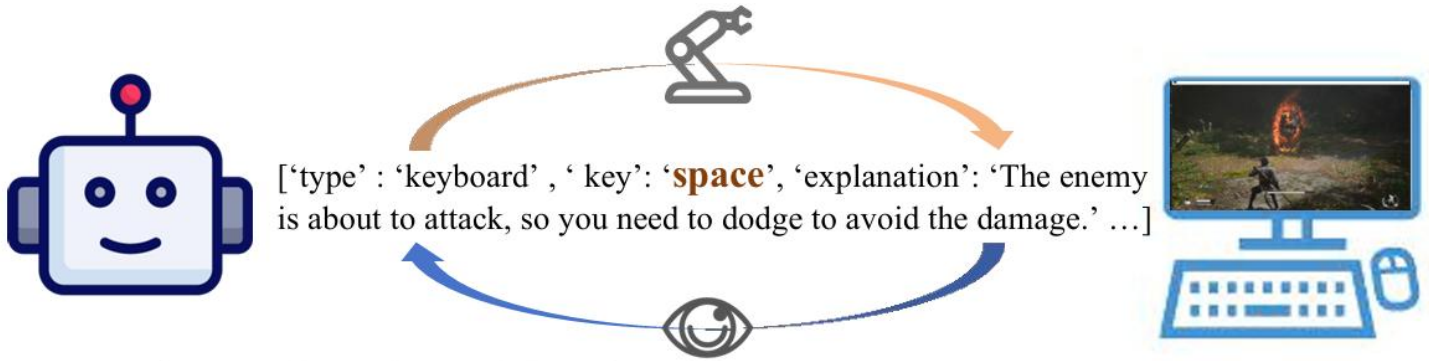
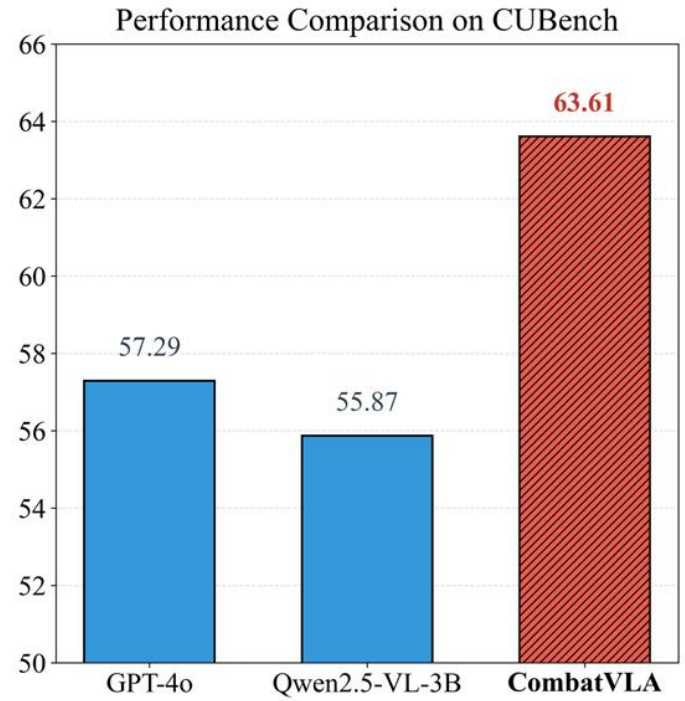
CombatVLA: An Efficient Vision-Language-Action Model for Combat Tasks in 3D Action Role-Playing Games

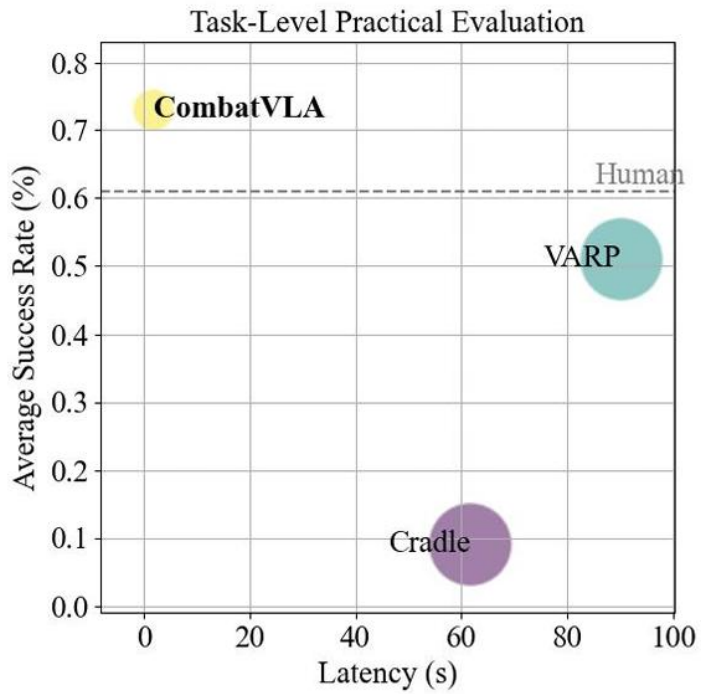
ICCV 2025

Peng Chen, Pi Bu, Yingyao Wang, Xinyi Wang, Ziming Wang, Jie Guo
Yingxiu Zhao, Qi Zhu, Jun Song, Siran Yang, Jiamang Wang, Bo Zheng

Alibaba Group


['type' : 'keyboard' , ' key': 'space', 'explanation': 'The enemy is about to attack, so you need to dodge to avoid the damage.' …]


Performance Comparison on CUBench
66
64
62
60
58
56
54
52
50
63.61
57.29
55.87
GPT-4o
Qwen2.5-VL-3B
CombatVLA


Task-Level Practical Evaluation
Average Success Rate (%)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
CombatVLA
Human
VARP
Cradle
0
20
40
60
80
100
Latency (s)

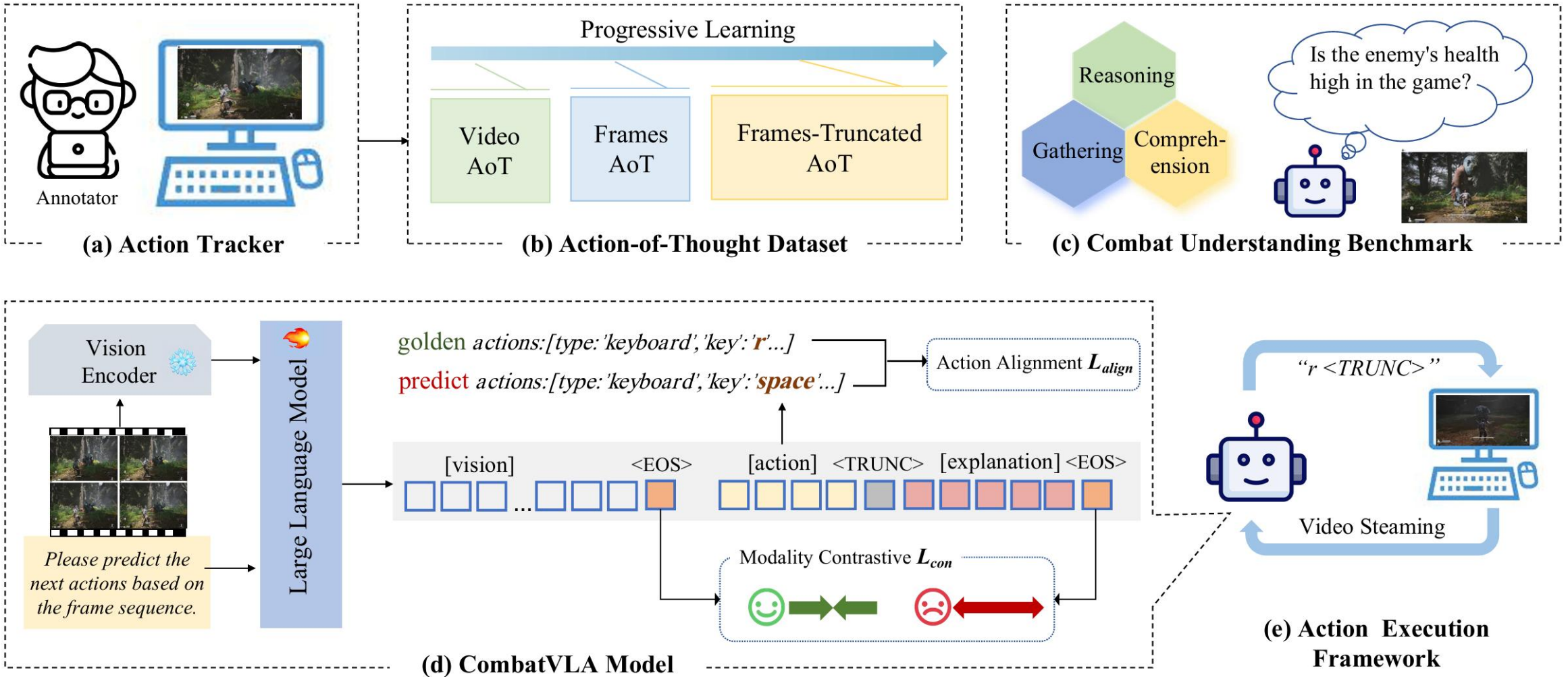

Annotator
(a) Action Tracker
Progressive Learning
Video AoT
Frames AoT
Frames-Truncated AoT
(b) Action-of-Thought Dataset
Reasoning
Gathering
Compreh-ension
Is the enemy's health high in the game?
(c) Combat Understanding Benchmark
Vision Encoder
Please predict the next actions based on the frame sequence.
Large Language Model
golden actions:[type:'keyboard','key':'r'...]
predict actions:[type:'keyboard','key':'space'...]
Action Alignment L_{align}
[vision] <EOS> [action] <TRUNC> [explanation] <EOS>
Modality Contrastive L_{con}
(d) CombatVLA Model
"r <TRUNC>"
Video Steaming
(e) Action Execution Framework

Table 2. Performance comparison of closed source and open source LVLMs on the combat understanding benchmark and general benchmark. The highest scores among models in each metric are highlighted in **FirstBest**.

Model	Combat Understanding				General Benchmark		
	Gathering	Comprehension	Reasoning	Avg.	MME	VideoMME	OCRBench
Closed-Source Large Vision Language Models							
GPT-4o-0513	58.06	**66.67**	47.14	57.29	2328	71.9	736
GPT-4o-mini-0718	59.44	66.18	42.57	56.06	2003	64.8	785
GPT-4-vision-preview	52.78	53.92	43.71	50.14	1926	59.9	645
Gemini-2.0-flash	58.61	64.22	50.86	**57.90**	–	–	–
Gemini-1.5-pro	**64.44**	62.75	41.71	56.30	2110	75.0	754
Claude35-sonnet	53.89	57.35	**55.43**	55.56	1920	60.0	788
Open-Source Large Vision Language Models							
LLaVA-1.5-7B	50.56	60.29	42.86	51.24	1510	–	–
InternVL2.5-4B	53.89	48.04	43.71	48.55	2337	62.3	828
Qwen2-VL-7B	55.28	59.80	43.14	52.74	2326	63.3	866
Qwen2-VL-2B	53.33	46.57	42.86	47.59	1872	55.6	809
Qwen2.5-VL-7B	45.56	52.94	50.57	49.69	2347	65.1	864
Qwen2.5-VL-3B	53.61	56.86	57.14	55.87	2157	61.5	797
CombatVLA-3B (Ours)	**60.83**	**60.29**	**69.71**	**63.61**	2141	58.7	741

Table 1. Task definitions in *Black Myth: Wukong* (BMW) and *Sekiro: Shadows Die Twice* (SSDT).

Game	Task ID	Description	Diffuculty	Zero-Shot
BMW	1	Defeat WolfScout	Easy	✓
	2	Defeat WolfStalwart	Easy	✓
	3	Defeat WolfSwornsword	Easy	✓
	4	Defeat WolfSoldier	Easy	✓
	5	Defeat Croaky	Easy	✓
	6	Defeat Crow Diviner	Middle	✓
	7	Defeat Bandit Chief	Middle	✓
	8	Defeat Bullguard	Hard	✓
	9	Defeat Wandering Wight	Very Hard	✗
	10	Defeat Guangzhi	Very Hard	✗
SSDT	11	Defeat Katana	Easy	✓
	12	Defeat Hassou Stance	Middle	✓
	13	Defeat Shigenori Yamauchi	Hard	✓

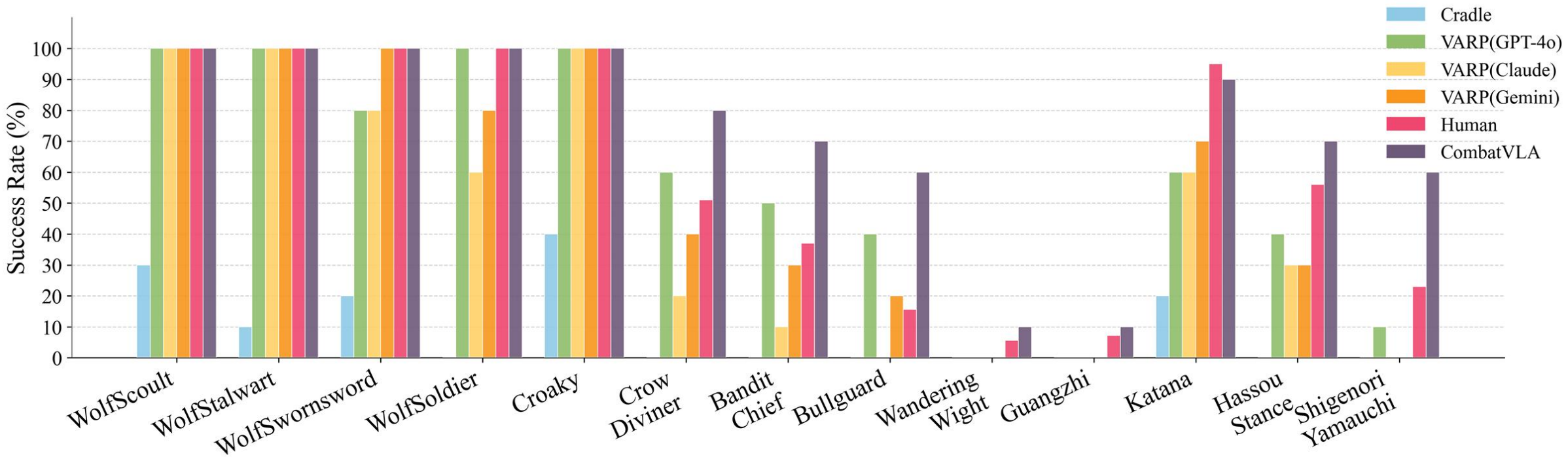


Figure 5. Comparison of task-level practical tests. Our CombatVLA not only outperforms all VLM-based agents (i.e., Cradle and VARP) but also has a higher task success rate than human players.

Project: https://combatvla.github.io/