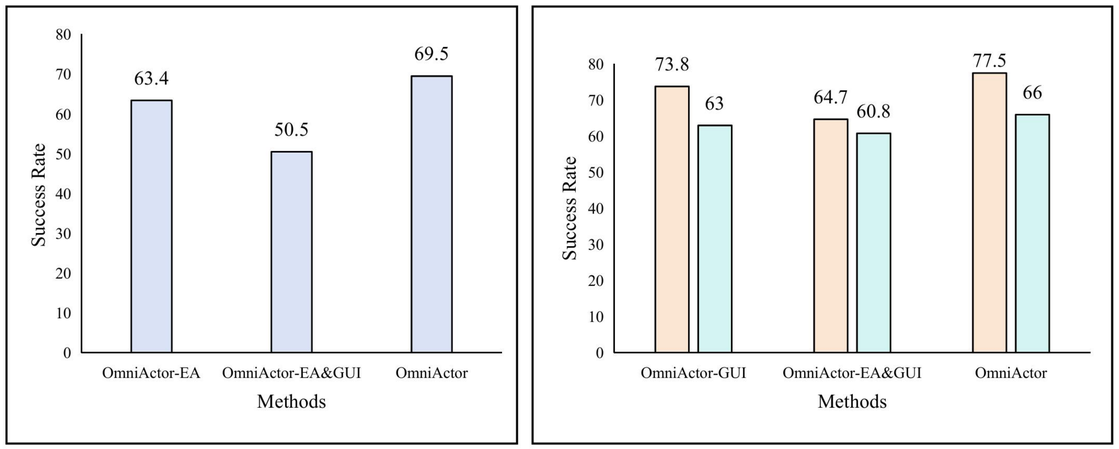
OmniActor: A Generalist GUI and Embodied Agent for 2D&3D Worlds

Longrong Yang[1,2], Zhixiong Zeng[2], Yufeng Zhong[2], Jing Huang[2], Liming Zheng[2], Lei Chen[2], Haibo Qiu[2], Zequn Qin[1], Lin Ma[2], Xi Li[1]

[1]Zhejiang University, [2]Meituan Technology

Motivation

- Humans can perform actions in different environments. ***Can an agent perform actions in both 2D and 3D worlds like humans, thereby achieving a generalist agent?***



- We initially mix GUI and embodied data to train, but find the performance degeneration brought by the data conflict, thus **aiming to utilize the synergy while eliminating theconflict**
  - OmniActor-GUI: the agent trained on GUI data
  - OmniActor-EA: the agent trained on embodied data
  - OmniActor EA&GUI: the agent trained on GUI and embodied data

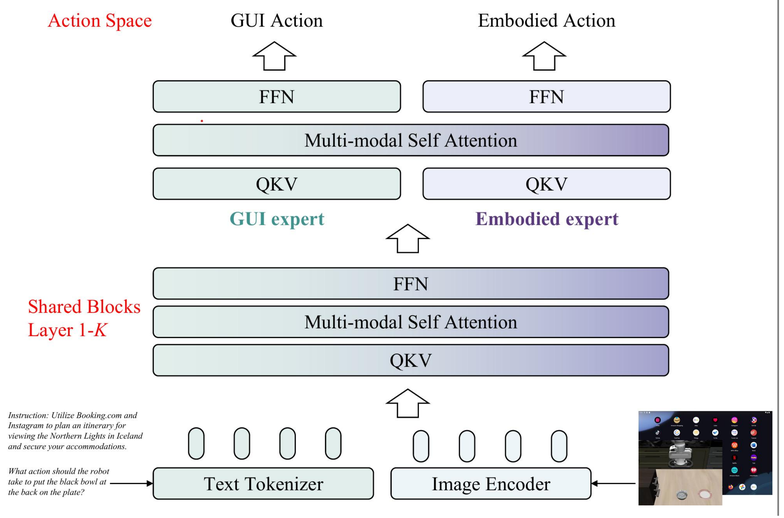
OmniActor: Cerebrum-cerebellum mechanism

1. Data Processing: GUI and embodied data are unified for large-scale training

- Unification of Data Format: we convert all GUI and embodied data into the ShareGPT format
  - Each sample includes a system prompt, an image, a task instruction, and an action
- Tokenization of Embodied Actions: we use a special tokenizer to convert embodied actions into tokens
  - 6-DoF displacement ranges from [-1, 1], whiech is discretized into K intervals. Each interval is a token ID
  - GUI and embodied actions share the same vocabulary

2. Layer-heterogeneity MoE

- Observation: We design a novel metric *parameter update similarity*, and find *parameter update similarity* in the shallow layers is much higher than that in the deep layers
- Based on the above oservation, we share parameters in the shallow layers for synergy, while separate parameters in the deep layers for eliminating conflict



Experiments

Comparison between different agents

	Agents	Source	Embodied Tasks LIBERO-90	GUI Tasks AndroidControl-Low	AndroidControl-High	GUI Odyssey
GUI Agents	Claude	In-house	-	19.4	12.5	3.1
	GPT-4o	In-house	-	19.4	20.8	3.3
	Qwen2-VL-7B	In-house	-	82.6	69.7	60.2
	UI-TARS-7B	In-house	-	90.8	72.5	87.0
	SeeClick	ACL'24	-	75.0	59.1	53.9
	Aria-UI	ACL'25	-	67.3	10.2	36.5
	OS-Atlas-7B	ICLR'25	-	85.2	71.2	62.0
	Aguvis-7B	ICML'25	-	80.5	61.5	63.8
	ScaleTrack-7B	arXiv'25	-	86.6	77.9	65.3
Embodied Agents	MUTEX	CoRL'23	53.0	-	-	-
	Distill-D	CoRL'23	49.9	-	-	-
	ACT	RSS'23	46.6	-	-	-
	MaIL	CoRL'24	60.3	-	-	-
	PRISE	ICML'24	54.4	-	-	-
	ATM	RSS'24	48.4	-	-	-
	MDT	RSS'24	67.2	-	-	-
Generalist Agents	Magma	CVPR'25	34.7	52.1	32.7	51.0
	GEA	CVPR'25	48.0	-	57.3	-
	NaviMaster	arXiv'25	-	68.9	54.0	-
	OmniActor-GUI	-	-	89.4	73.8	63.0
	OmniActor-EA	-	63.4	-	-	-
	OmniActor	-	69.5	86.4	77.5	66.0

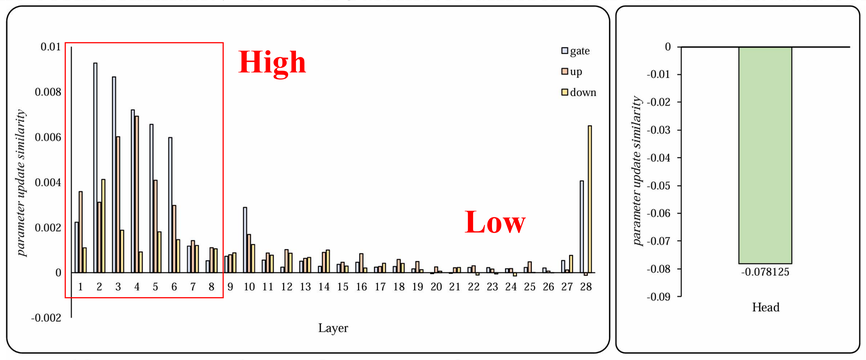
Study about different OmniActor variants

- OmniActor *hard*: Parameters of attention heads, FFNs, and the classification head are fully separated
- OmniActor *router*: A router is incorporated prior to each MoE layer, which takes token features as input and outputs theprobabilities of assigning the token to different experts

Models	Embodied Tasks LIBERO-90	GUI Tasks AndroidControl-Low	AndroidControl-High	GUI Odyssey	Avg
OmniActor-GUI	-	88.4	74.5	63.0	-
OmniActor-EA	63.4	-	-	-	-
OmniActor-EA&GUI	50.5	86.3	71.0	60.8	67.2
OmniActor *hard*	59.5	85.6	75.2	63.9	71.1
OmniActor	69.5	87.5	77.1	66.0	75.0
OmniActor *router*	64.0	86.0	72.6	66.1	72.2

Parameter update similarity

- The cosine similarity between parameter difference from different data before and after training



(a) *Parameter update similarity* statistic for FFNs (b) Statistic for prediction head

Feel free to communicate if you have any questions.