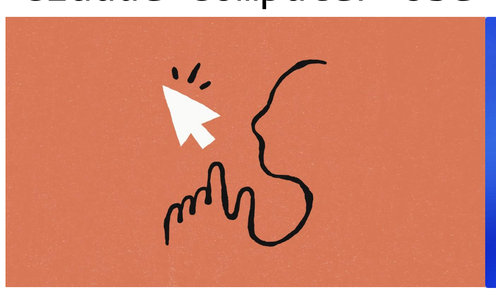
UI-RFT: Reinforcement Fine-Tuning for GUI Grounding

Yanheng He
Shanghai Jiao Tong University

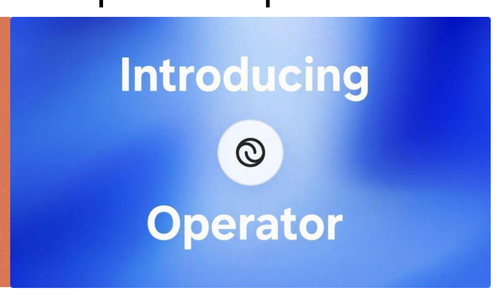
Background

GUI Agents

Claude Computer Use

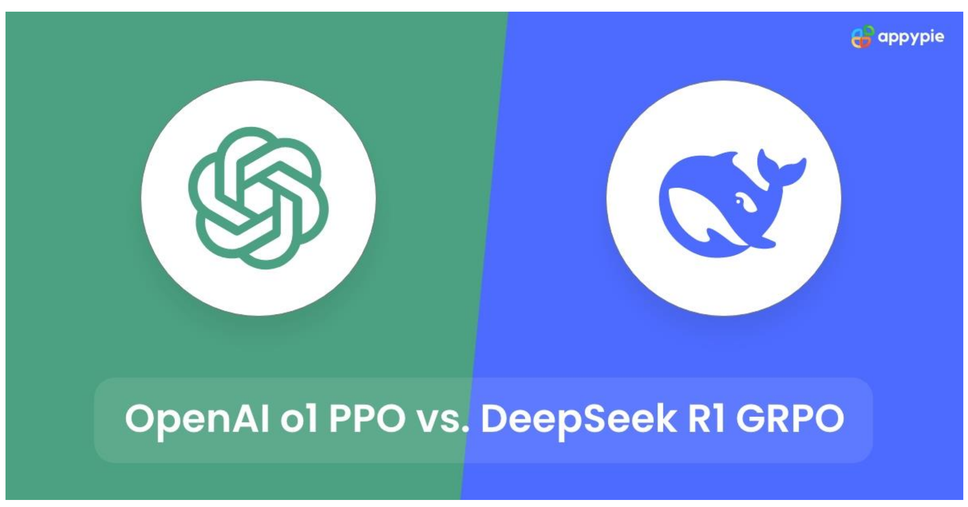


OpenAI Operator



Use computers like human do!

Rule-based RL



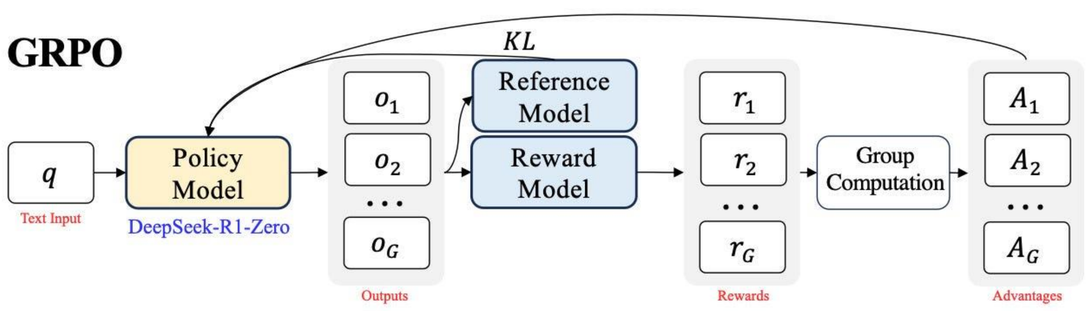
Task

GUI grounding is a fundamental visual capability for agents to effectively use computers.

Example of ScreenSpot-Pro

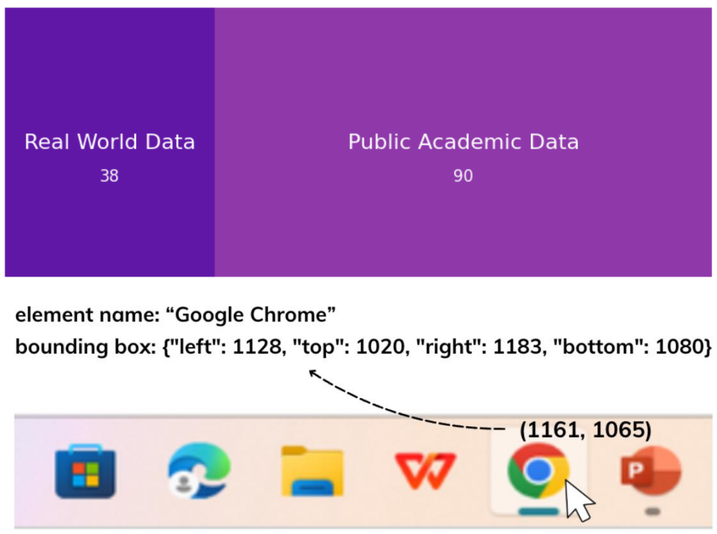
Instruction: Execute Python scripts.
Application: Unreal Engine
Type: text
Bounding Box: [246, 2035, 377, 2054]

Method

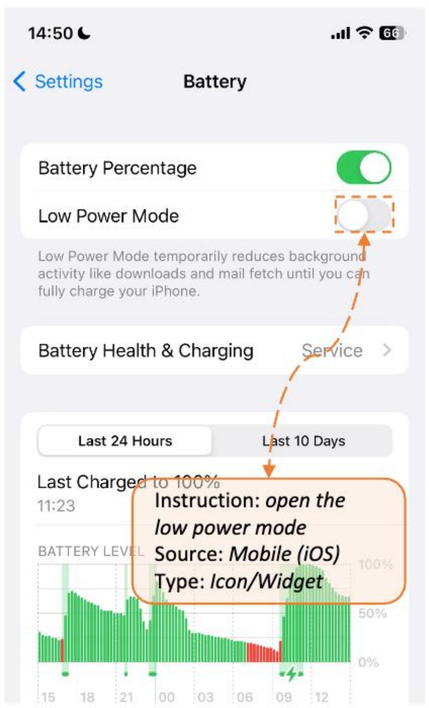


Data

UI-128=38(real)+90(public)



Real world data from PC Tracker



ScreenSpot (mobile)

Training

volcengine/**verl**

verl: Volcano Engine Reinforcement Learning for LLMs

115 Contributors · 14 Used by · 13 Discussions · 6k Stars · 656 Forks

RL framework: verl

Hyperparameter	Value
lr	1e-6
num_rollout	16
num_train_epochs	16
max_prompt_length	8192
per_device_train_batch_size	4
gradient_accumulation_steps	1

Hyperparameter

Base model: Qwen2.5-VL-3B-Instruct
(already strong GUI grounding)

Cost: 8 NV H100s x 4 hours

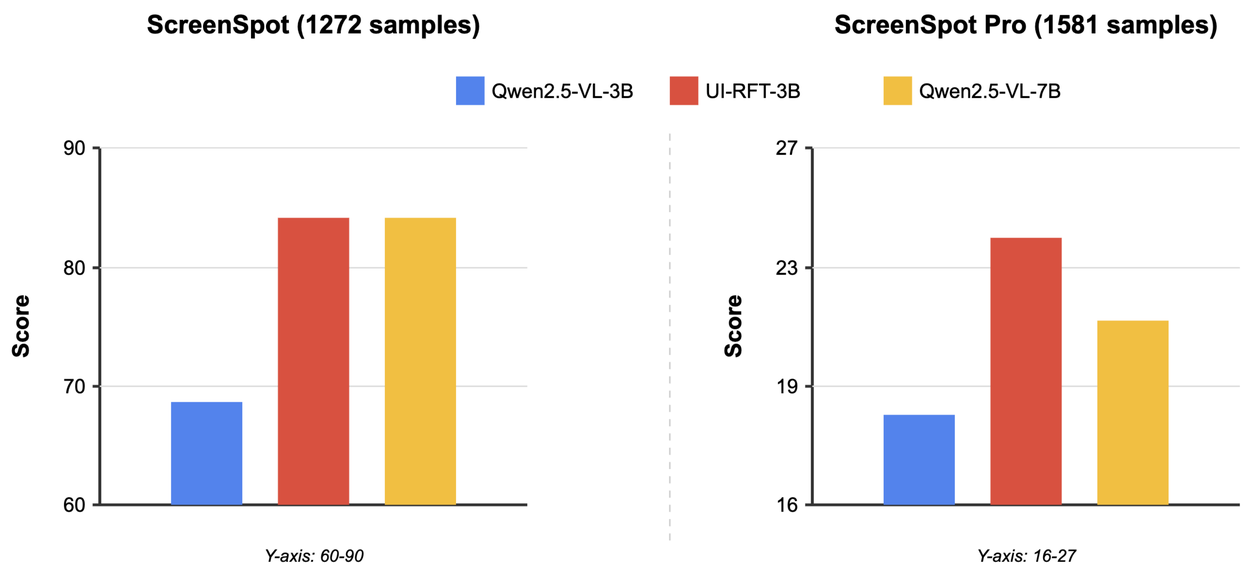
Prompt for Training and Inference

<*image*>You are specialize in GUI grounding. This is an interface to a GUI and you are going to perform click action to ***instruction***. Output exactly one line containing a single JSON object in the following format:
{“point_2d”: [x, y], “label”: “object name/description”}.

Rule-based Reward

$$R_{\text{acc}} = \begin{cases} 1 & \text{if the extracted point } (x, y) \text{ is valid and in box } \mathcal{B}, \\ 0 & \text{otherwise (including invalid format).} \end{cases}$$

Experiments



GUI Grounding Benchmarks: ScreenSpot & ScreenSpot-Pro
Metric: Mean Accuracy, with temperature set to 0 for reproducibility

Model	Mobile		Desktop		Web		Average
	Text	Icon	Text	Icon	Text	Icon	
Qwen2.5-VL-3B	89.0	69.0	81.4	47.9	61.7	51.0	68.63
UI-RFT-3B (Ours)	92.7	85.6	87.1	67.9	85.2	78.2	**84.12**
Qwen2.5-VL-7B	96.7	73.8	88.7	70.7	88.7	70.7	84.12
Qwen2.5-VL-72B	96.0	81.2	94.8	82.1	94.8	86.4	89.86

Table 1: Grounding accuracy on ScreenSpot.

Model	Qwen2.5-VL-3B	UI-RFT-3B (Ours)	Qwen2.5-VL-7B	Qwen2.5-VL-72B
Score	18.79	**24.23**	21.69	49.72

Table 2: Grounding accuracy on ScreenSpot-Pro.

Takeaway

- **Reinforced fine-tuning with only 128 high-quality samples significantly enhances GUI grounding.**
- **GUI grounding is a fundamental visual ability in VLMs, improved without needing long reasoning chains.**

henryhe_sjtu@sjtu.edu.cn