UI-Ins: Enhancing GUI Grounding with Multi-Perspective Instruction-as-Reasoning

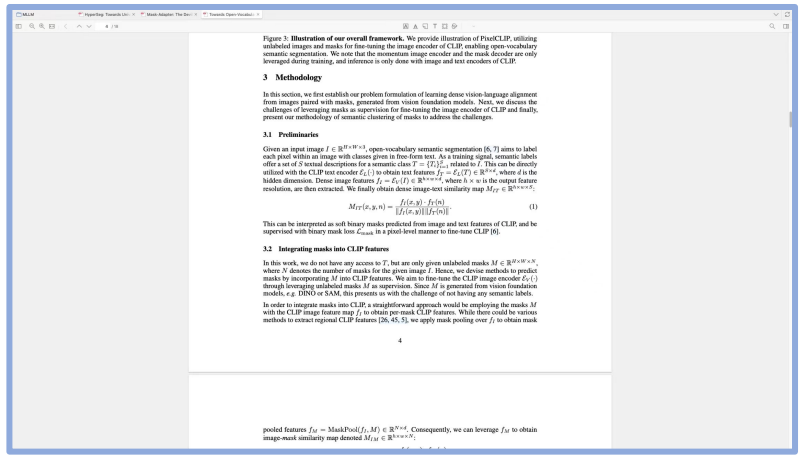
Liangyu Chen, Hanzhang Zhou, Chenglin Cai, Jianan Zhang, Panrong Tong, Xu Zhang, Quyu Kong

Chen Liu, Yuqi Liu, Wenxuan Wang, Yue Wang, Qin Jin, Steven HOI

RUC Tongyi lab, Alibaba Group

Background

What is GUI grounding?

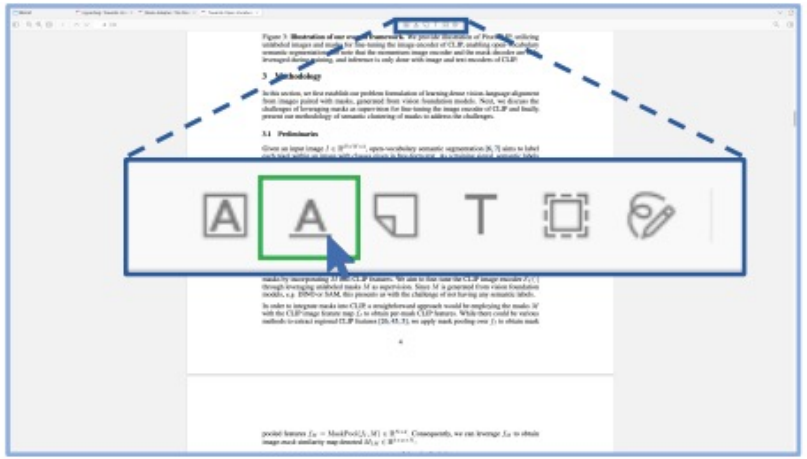


Original Image

User Instruction

Refine your search to view Charizard products from the company known for building block toys.

Grounding Model



Prediction Point

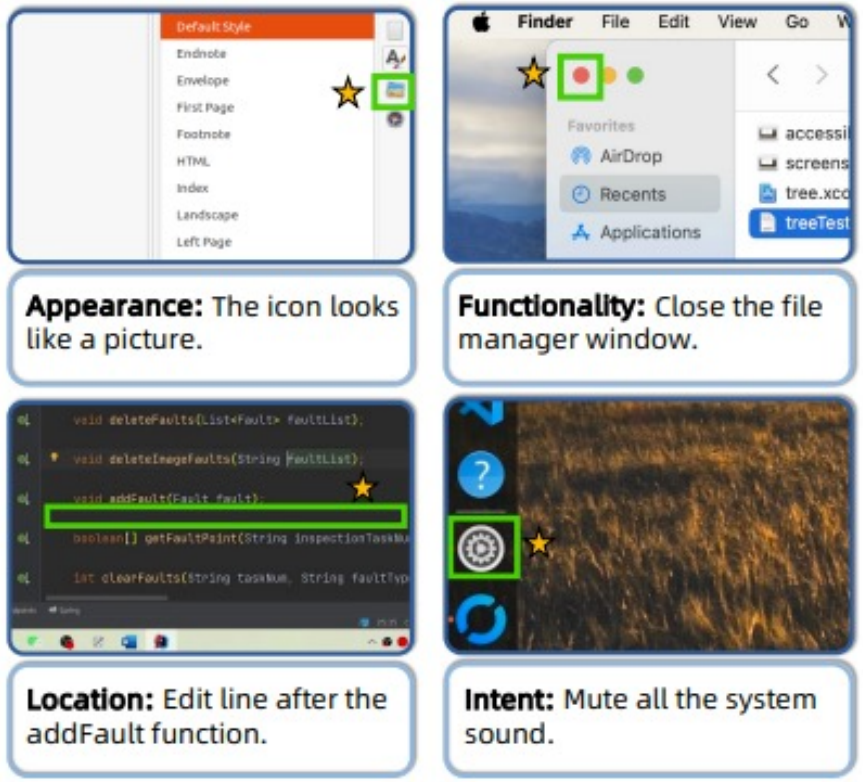
: Ground Truth : UI-Ins Prediction : Zoom in for visualization

Motivation

How does human describe and click the UI elements usually?

- Appearance description.
- Functionality description.
- Location description.
- Intent description.

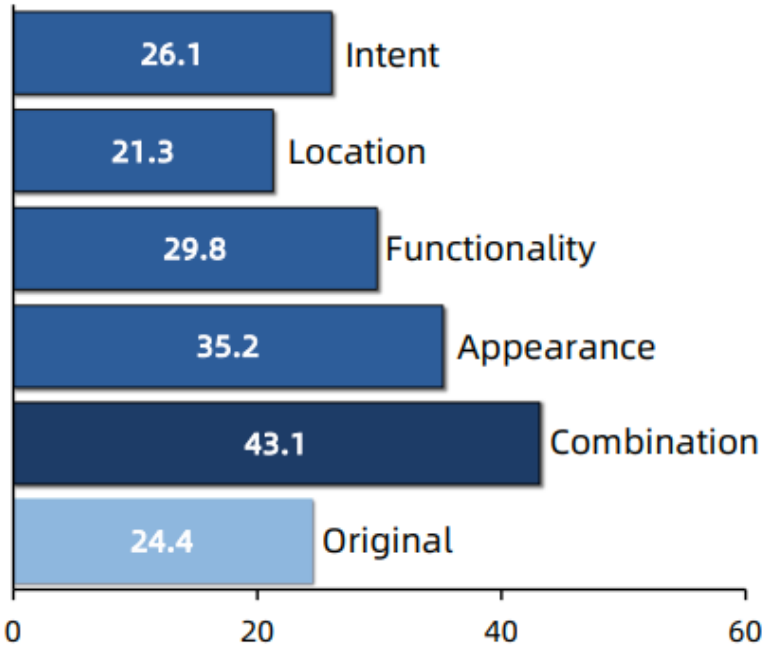
Human always choose the most effective instruction in different scenarios. But can nowadays grounding model choose the best-perform type to achieve the goal?



How much does instructions really matter?

How does the diversity of instructional perspectives affect grounding accuracy?

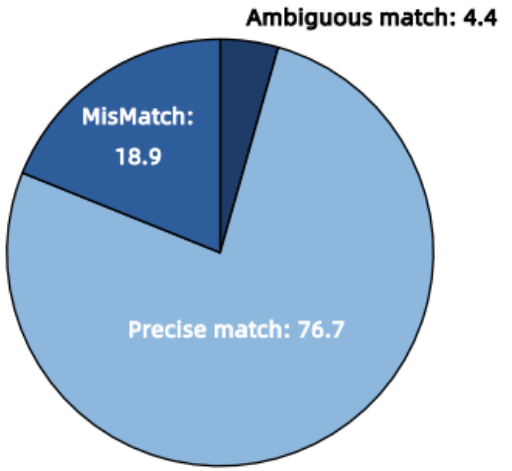
- Different instruction type can effect the model's performance significantly even on **zero-shot** setting.
- If model can always choose the best-perform instruction for each GUI grounding sample (**Combination**), model performance can achieve a **76%** relative improvement.

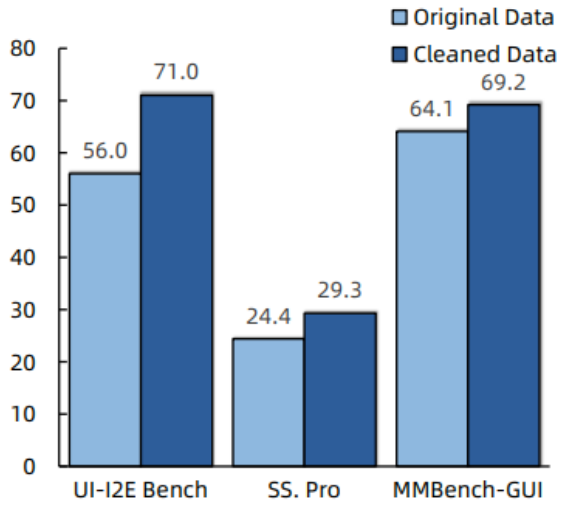


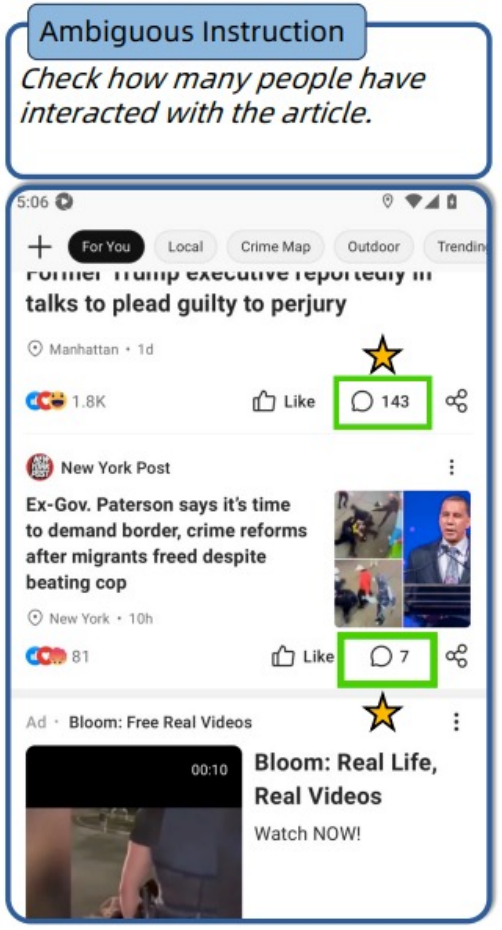
How much does instructions really matter?

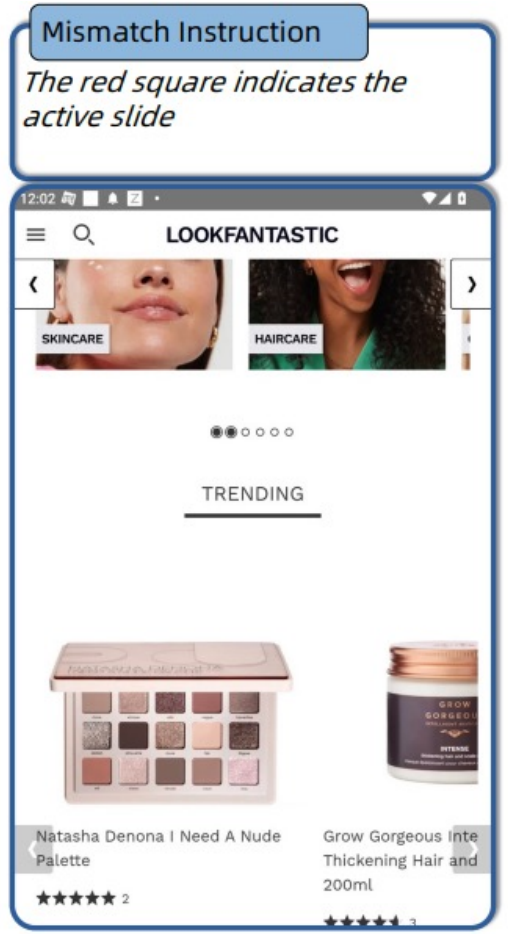
What is the state of instruction quality in GUI grounding datasets, and what is its impact?

- About **23%** samples in open-source datasets exhibit substantive flaws.
- Cleaned data can help model improve performance by simply SFT training (ori: 21w, cleaned: 18w).





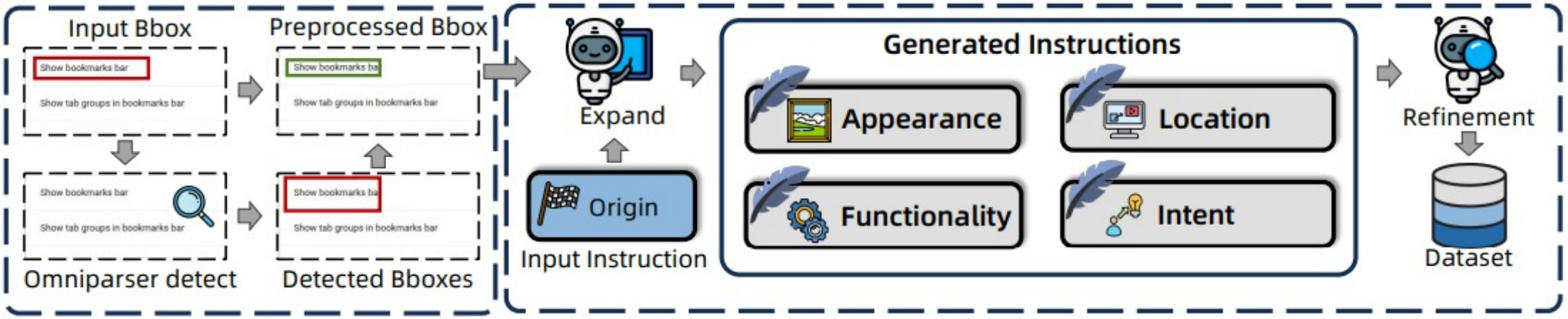




Method – Data Pipeline

How to improve the quality of GUI grounding data?

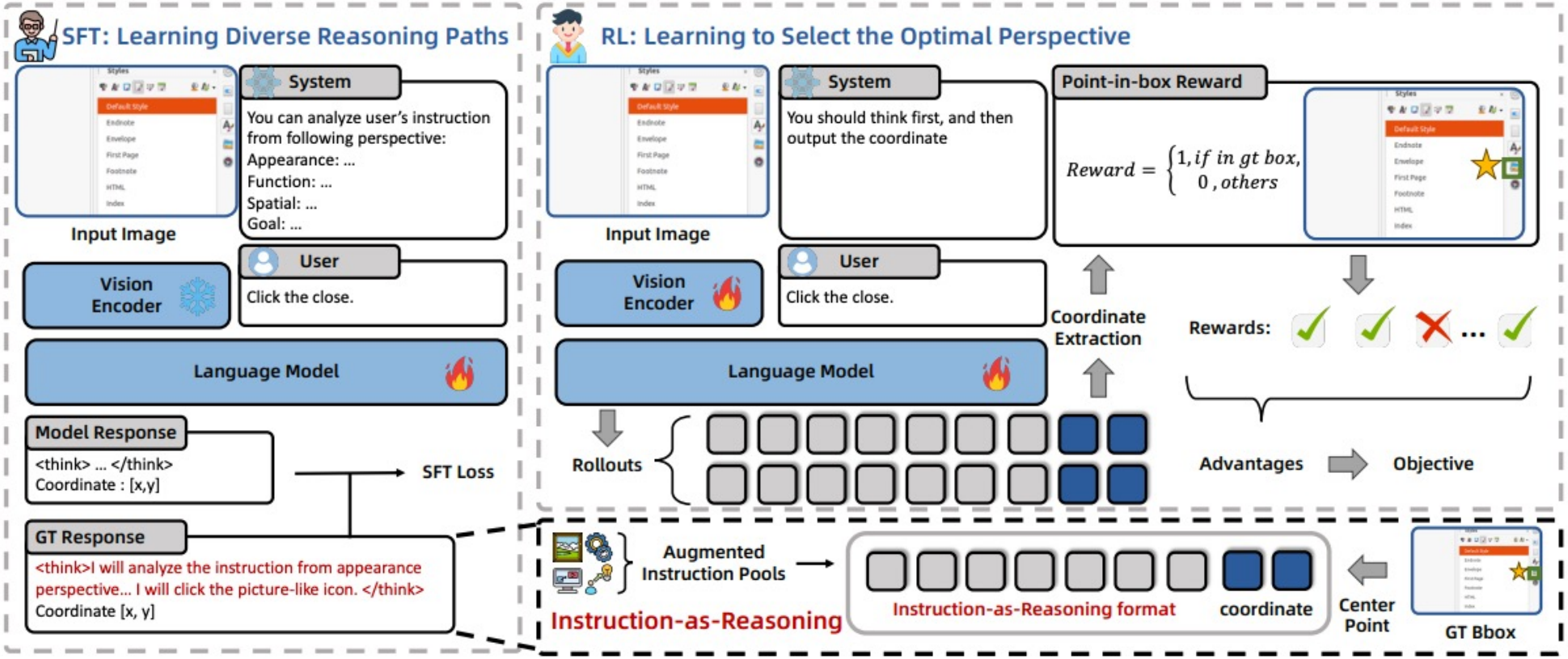
- Bbox preprocess: Use Omniparser V2 (Microsoft detection model) detect all UI elements in image, modify the GT bbox with the intersection.
- Multi-perspective instruction generation: four perspectives.
- Instruction-GT Bbox correspondence refinement: filter the inaccurate instructions.



Method – Instruction-as-Reasoning

How can model reasoning in diverse perspective?

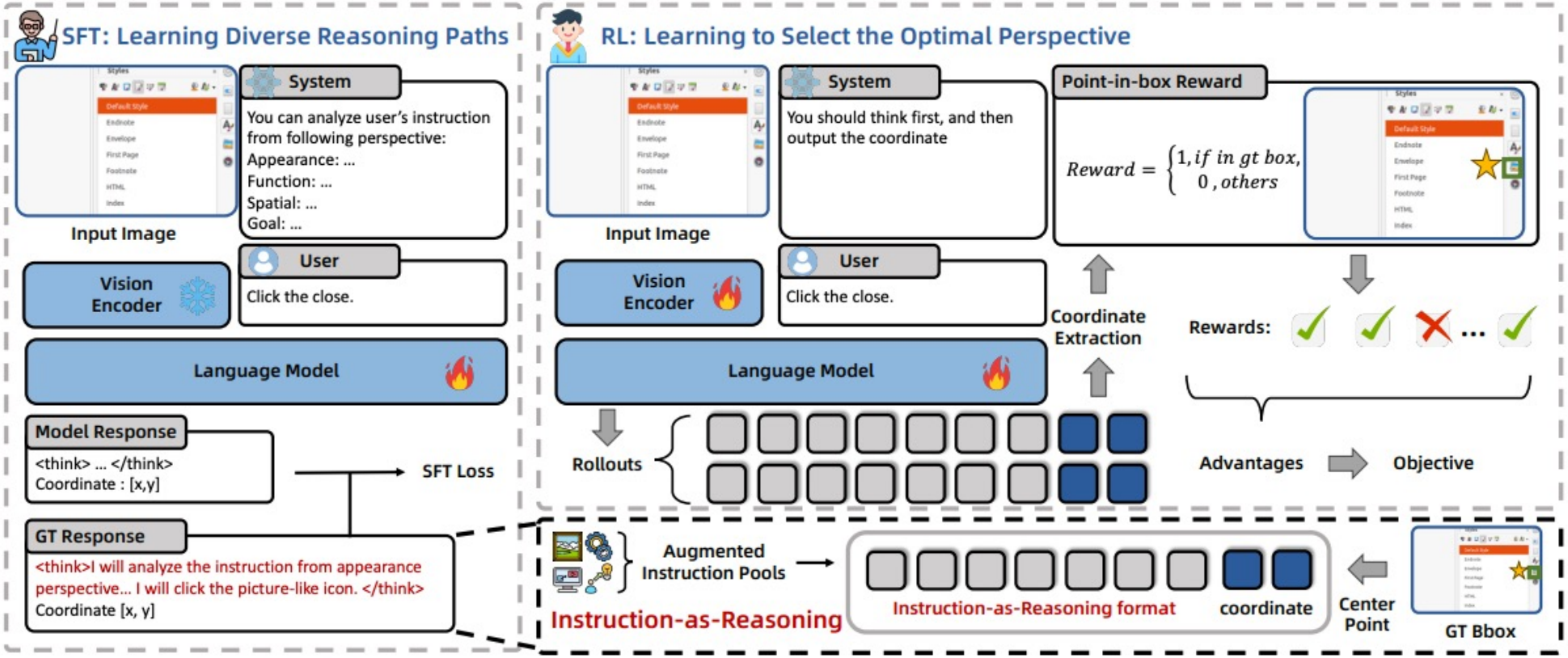
- Despite prior works treat the instructions simply as the input, we treat these diverse instructions as the reasoning pathways in SFT stage.



Method – Instruction-as-Reasoning

How can model choose the optimal reasoning pathways in different grounding scenarios?

- Model learns to select the optimal perspective by point-reward in RL stage (GRPO).



Results

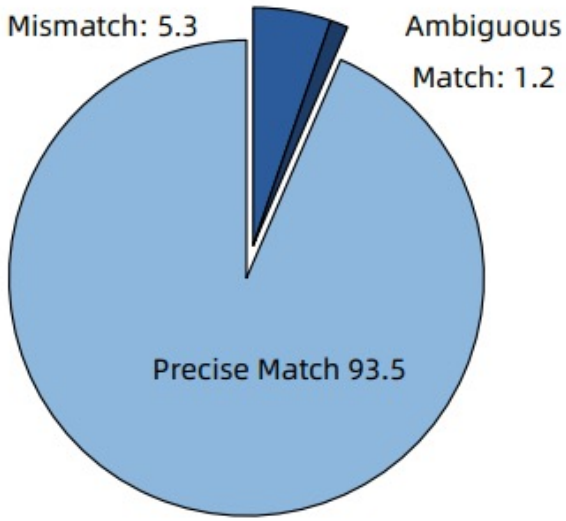
By Instruction-as-Reasoning approach, we build UI-Ins-7B and UI-Ins-32B upon Qwen2.5-VL-7B and Qwen2.5-VL-32B. We achieve state-of-the-art on grounding benchmarks and AndroidWorld.

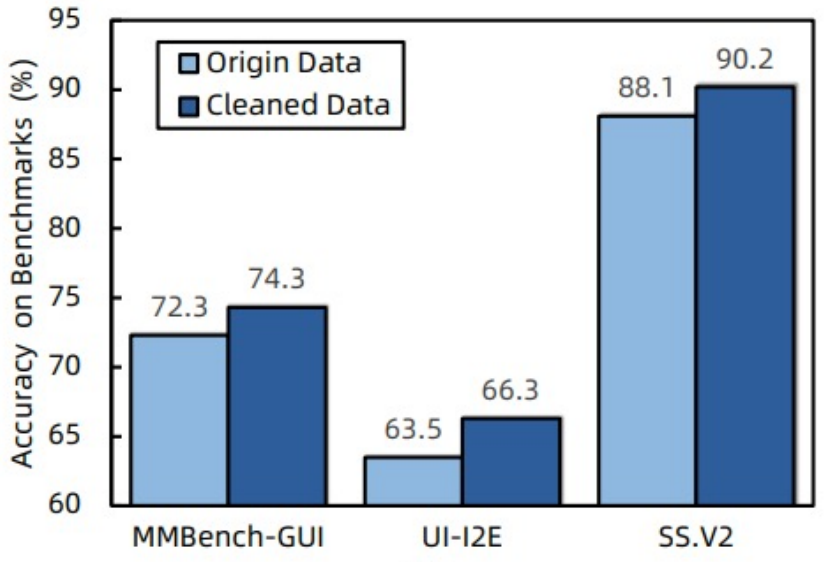
BENCHMARK	SCORE	KEY IMPROVEMENT HIGHLIGHTS
UI-I2E-Bench	**87.30%**	Significantly enhanced implicit instruction understanding
MMBench-GUI L2	**84.90%**	More robust handling of complex hierarchical instructions
ScreenSpot-Pro	**57.00%**	Excellent performance in high-resolution professional software scenarios
ScreenSpot-V2	**94.90%**	Strong cross-platform generalization (Windows/macOS/Android)
ShowDown	**73.80%**	Outstanding low-level control and instruction-following ability

Ablation – Data Quality

The effects of data processing pipeline :

- Error rate: 23.3 -> less than 8%.
- Model performance improved substantially (SFT 34w original data vs. 29w cleaned data, 1epoch only).





Ablation – Training Stage & Reasoning

- SFT+RL outperforms pure SFT and RL training.
- Reasoning is important in both SFT stage and RL stage.

Table 6: Ablation study on training stages. We report accuracy on MMBench-GUI L2 (MM), UI-I2E-Bench (I2E), Showdown (Show), ScreenSpot-Pro (Pro), and ScreenSpot-V2 (V2).

SFT	RL	MM	I2E	Show	Pro	V2
✗	✗	63.4	56.0	43.6	24.4	86.5
✗	✓	72.4	69.2	66.6	37.0	88.6
✓	✗	76.3	70.1	67.5	37.1	90.6
✓	✓	**83.1**	**81.1**	**73.1**	**52.2**	**94.0**

Table 7: Ablation on the intermediate reasoning component. Its removal results in a significant performance degradation across all benchmarks. ✓represents let the model use Instruction as Reasoning in the corresponding stage.

SFT	RL	MM	I2E	Show	Pro	V2
✗	✗	79.1	70.7	66.1	44.8	91.7
✗	✓	78.8	71.6	68.4	48.0	92.0
✓	✗	81.6	76.2	72.0	47.5	93.1
✓	✓	**83.1**	**81.1**	**73.1**	**52.2**	**94.0**

Key insights – Effect of Instruction-as-Reasoning on SFT+RL training paradigm

How does Instruction-as-Reasoning help SFT+RL training paradigm?

- The upper bound and training efficiency can be effected significantly by the base model's performance.
- Even finetune only 5k grounding data (pure coordinate response as GT) in SFT stage. The finetuned Qwen2.5-VL-7B suffers heavily policy collapse in GRPO training process, similar phenomenon also appears in JEDI-7B and UI-Tars-1.5-7B.

Model	Training type	SS.Pro
UI-tars-1.5-7B	Zero-Shot	41.8
UI-tars-1.5-7B	Zero-Shot	46.5
Qwen2.5-VL-7B	Zero-Shot	28.1
Qwen2.5-VL-7B	Zero-Shot	34.9

Table 9: Instruction-as-Reasoning prevents policy collapse in RL and achieves significant accuracy gain in RL. This table contrasts our method with a standard SFT+RL pipeline. Scores after 100 RL steps are reported.

Method	Base Model	SS.Pro
SFT (w/o IR)	Qwen2.5-VL-7B	37.0
+ RL	Qwen2.5-VL-7B	34.9↓ (5.7%)
Zero-Shot	JEDI-7B	39.5
+ RL	JEDI-7B	34.5↓ (12.7%)
SFT (w/ IR)	Qwen2.5-VL-7B	37.1
+ RL	Qwen2.5-VL-7B	46.0↑ (24.0%)

Key insights – Difference of Instruction-as-Reasoning from Free-Form Reasoning

Can all types of reasoning work in GUI grounding?

- Prior works like GUI-G1, GUI-G2, UI-R1, GTA1 demonstrated, thinking in pure RL stage can bring negative influence.
- Our experiments also validates this opinion on different base models including UI-Tars-1.5-7B and Qwen2.5-VL-7B.
- Different from Free-Form Reasoning, Instruction-as-Reasoning can bring substantially improvements on GUI grounding task.

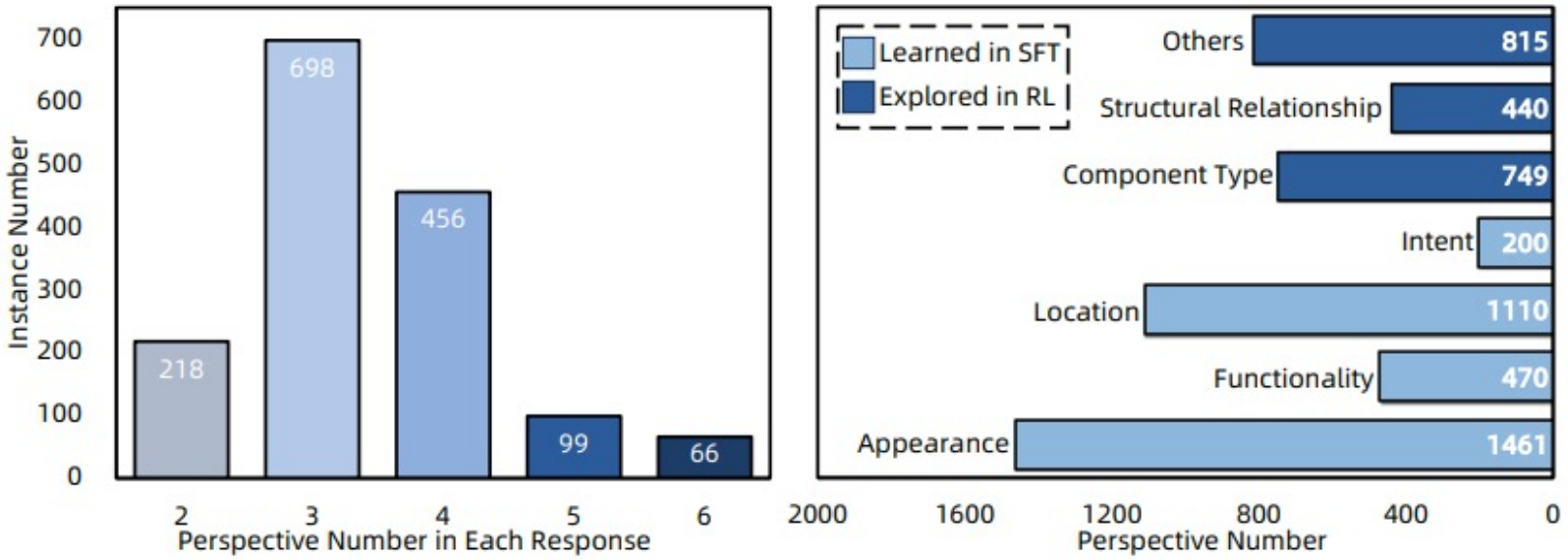
Table 8: Comparison between free-form reasoning (FFR) and Instruction as Reasoning (IR) in RL. Our Instruction-as-Reasoning is the key to unlocking effective reasoning for GUI grounding.

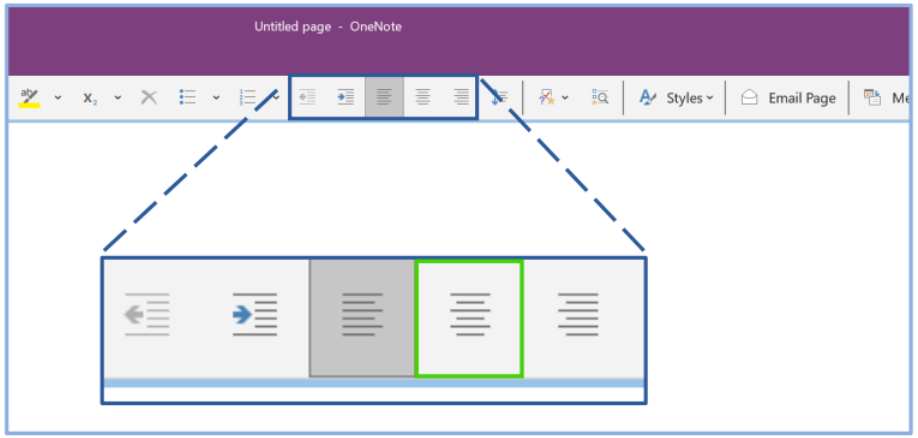
Method	Base Model	SS.Pro
Free-Form Reasoning (FFR) in RL		
RL (w/o FFR)	UI-Tars-1.5-7B	50.1
RL (w/ FFR)	UI-Tars-1.5-7B	46.9↓ (6.4)%
RL (w/o FFR)	Qwen2.5-VL-7B	36.4
RL (w/ FFR)	Qwen2.5-VL-7B	36.4↓ (0)%
Instruction-as-Reasoning (IR) in RL		
RL (w/o IR)	UI-Tars-1.5-7B	48.7
RL (w/ IR)	UI-Tars-1.5-7B	51.2↑ (5.1%)
RL (w/o IR)	Qwen2.5-VL-7B	47.5
RL (w/ IR)	Qwen2.5-VL-7B	52.2↑ (9.9%)

Visualization

UI-Ins can

- Select the optimal reasoning pathway.
- Combine different reasoning pathways.
- Explore emergent reasoning perspectives after RL.





Single Perspective Reasoning

Click the button to center the text [Functionality].

Combination Perspectives Reasoning

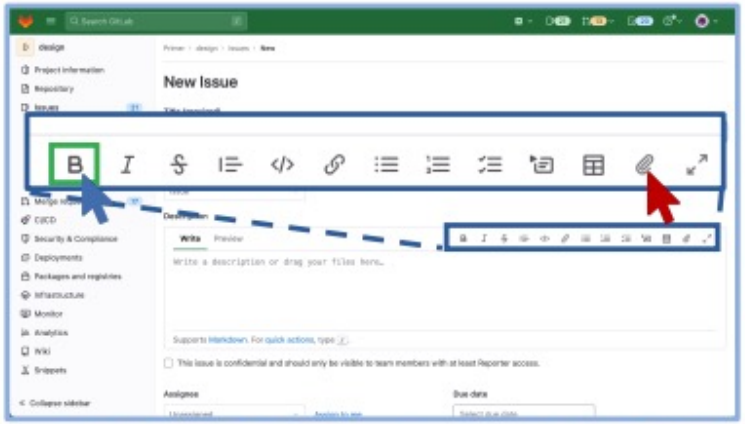
Click the button with the icon of centered horizontal lines [Appearance] to center the text [Functionality] .

Combination Perspectives and Emergent Perspective Reasoning

To the right of the currently highlighted option [Location] , click the inactive button [State] , identifiable by its icon of centrally-aligned staggered lines [Appearance] , to center the text [functionality] , which will set it as the new exclusive active state [Prediction] in alignment control group [Group Affiliation] .

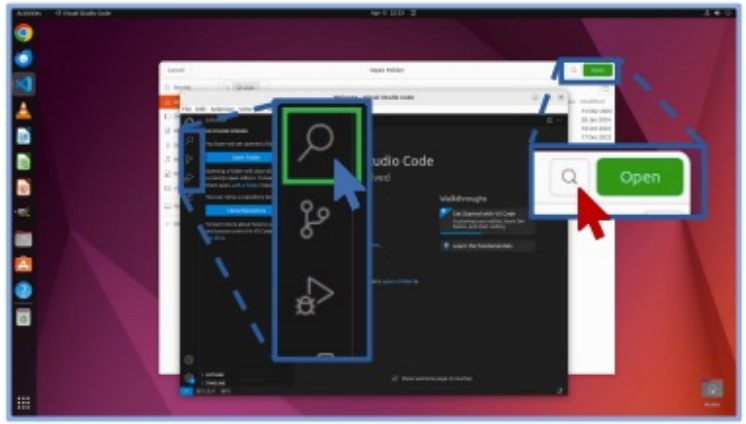
Visualization and Comparison

Make important points stand out in your issue by applying emphasis.



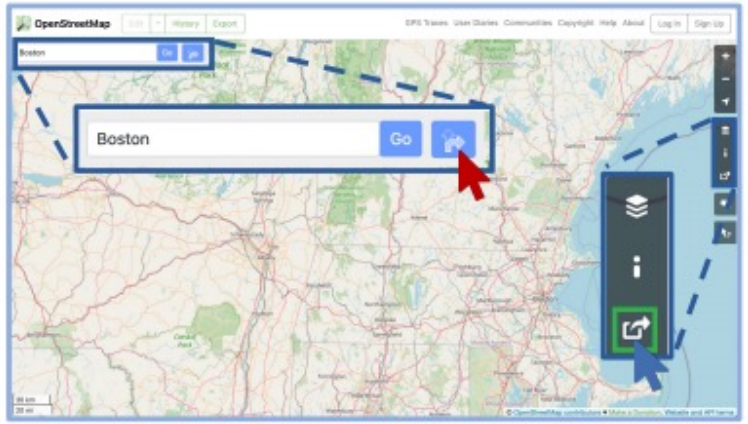
GTA1: (2394,804)
UI-Ins: <think> I will click the 'B' button in the formatting toolbar above the description box.</think>
<tool_call> [1590,801] </tool_call>

Access the search functionality to find files or text within VSCode workspace.



GTA1: (1496,164)
UI-Ins: <think> I will click the magnifying glass icon in the left sidebar.</think>
<tool_call> [962,463] </tool_call>

Create a shareable link to this specific map location showing Portland.



GTA1: (639,156)
UI-Ins: <think> I will click the button with the arrow icon pointing right and up, located on the right sidebar.think>
<tool_call> [2506,581] </tool_call>

: Ground Truth : GTA1 Prediction : UI-Ins Prediction : Zoom in for visualization

Error Analysis

- Lack of World Knowledge (a)
- Lack layout understanding capacity (b)
- Hallucination of VLMs (c) & (d)