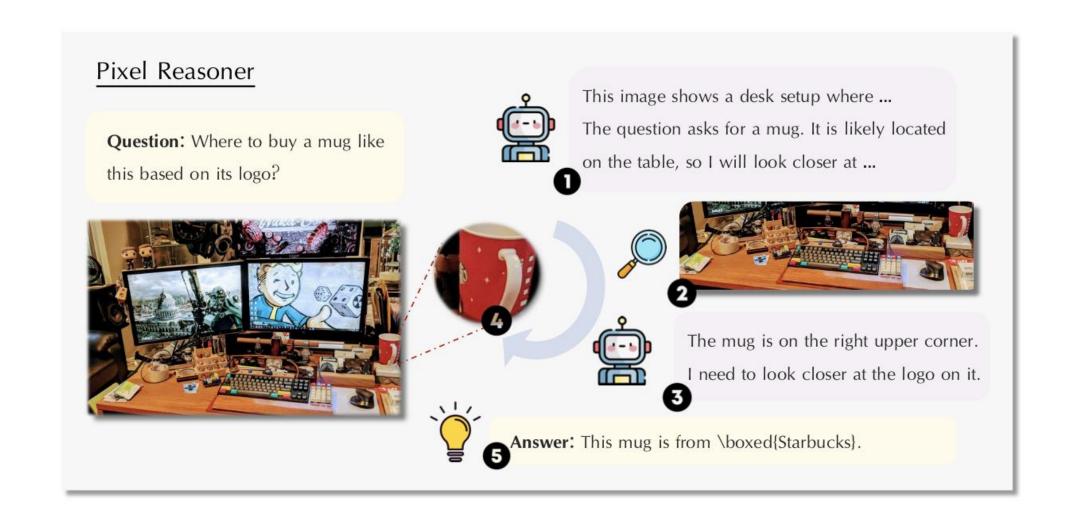# Pixel Reasoner: Overview

Pixel Reasoner: Overview

- How to enable VLMs to process multimodal inputs for better visual understanding?

- We focus on 2 visual operations: zoom-in and frame selection

- We propose a two-stage training method which involves instruction tuning stage and reinforcement learning stage.

Pixel Reasoner: Data Pipeline

Pixel Reasoner: Instruction Tuning

- Textual Reasoning

  - Mixed some data without visual operations.

  - Enables adaptive choice between pixel–space and textual reasoning.

- Loss Masking

  - Apply loss masks on execution outputs and erroneous operations.

  - Prevents model from learning incorrect actions.

Pixel Reasoner: Curiosity–driven RL

$$r'(\mathbf{x}, \mathbf{y}) = r(\mathbf{x}, \mathbf{y}) + \alpha \cdot r_{\text{curiosity}}(\mathbf{x}, \mathbf{y}) + \beta \cdot r_{\text{penalty}}(\mathbf{y}),$$

$$\text{where} \quad r_{\text{curiosity}}(\mathbf{x}, \mathbf{y}) = \max(\text{H} - \mathbf{RaPR}(\mathbf{x}), 0) \cdot \mathbf{1}_{\text{PR}}(\mathbf{y})$$

$$r_{\text{penalty}}(\mathbf{y}) = \min(\text{N} - \mathbf{n}_{\text{vo}}(\mathbf{y}), 0)$$

Pixel Reasoner: Performance

Table 1: Our main results on the four evaluated benchmarks.

Model	Size	V* Bench	TallyQA-complex	MVBench-test	InfoVQA-test
Metric		Acc	Acc	Acc	ANLS
Models w/o Tools					
GPT-4o	-	62.8	73.0	**64.6**	80.7
Gemini-2.0-Flash	-	73.2	73.8	-	**86.5**
Gemini-2.5-Pro	-	79.2	**74.0**	-	84.0
Qwen2.5-VL	7B	70.4	68.6	63.8	80.7
Video-R1	7B	51.2	42.6	63.9	67.9
LongLLava	13B	68.5	64.6	54.6	65.4
Gemma3	27B	62.3	54.3	56.8	59.4
Models with Tools					
Visual Sketchpad (GPT-4o)	-	80.4	-	-	-
IVM-Enhance (GPT-4V)	-	**81.2**	-	-	-
PaLI-3-VPD	5B	70.9	-	-	-
SEAL	7B	74.8	-	-	-
PaLI-X-VPD	55B	76.6	-	-	-
Ours (Initialized from Qwen2.5-VL-7B)					
Pixel-Reasoner	7B	**84.3**	**73.8**	**67.8**	**84.0**
Ablation Baselines (Ablated from Pixel-Reasoner)					
Warm-Start Model (w/o RL)	7B	79.0	67.9	59.0	74.3
RL w/o Curiosity	7B	81.1	71.8	66.4	80.7
RL w/o Warm-Start	7B	81.7	72.2	65.6	81.2
RL w/o Correction-Data	7B	80.1	69.8	63.6	78.2

Pixel Reasoner: Some Analysis