

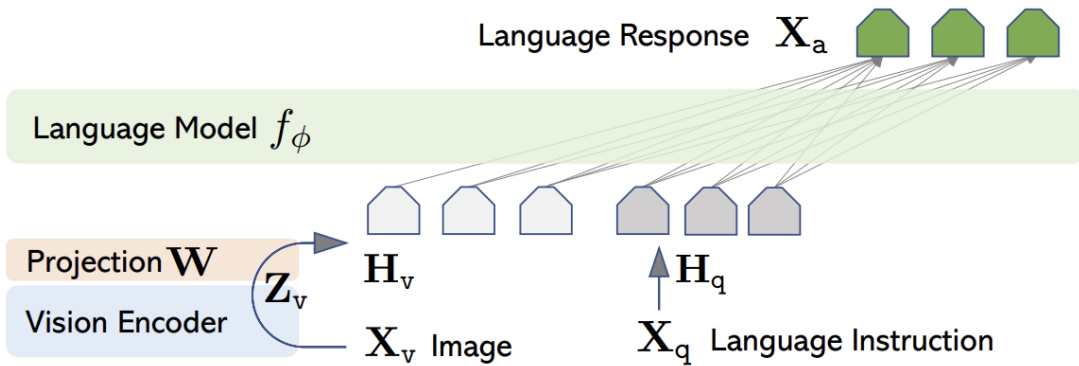
ERGO: Efficient High-Resolution Visual Understanding for Vision-Language Models

Jewon Lee*, Wooksu Shin*, Seungmin Yang, Ki-Ung Song, DongUk Lim, Jaeyeon Kim, Tae-Ho Kim†, Bo-Kyeong Kim†

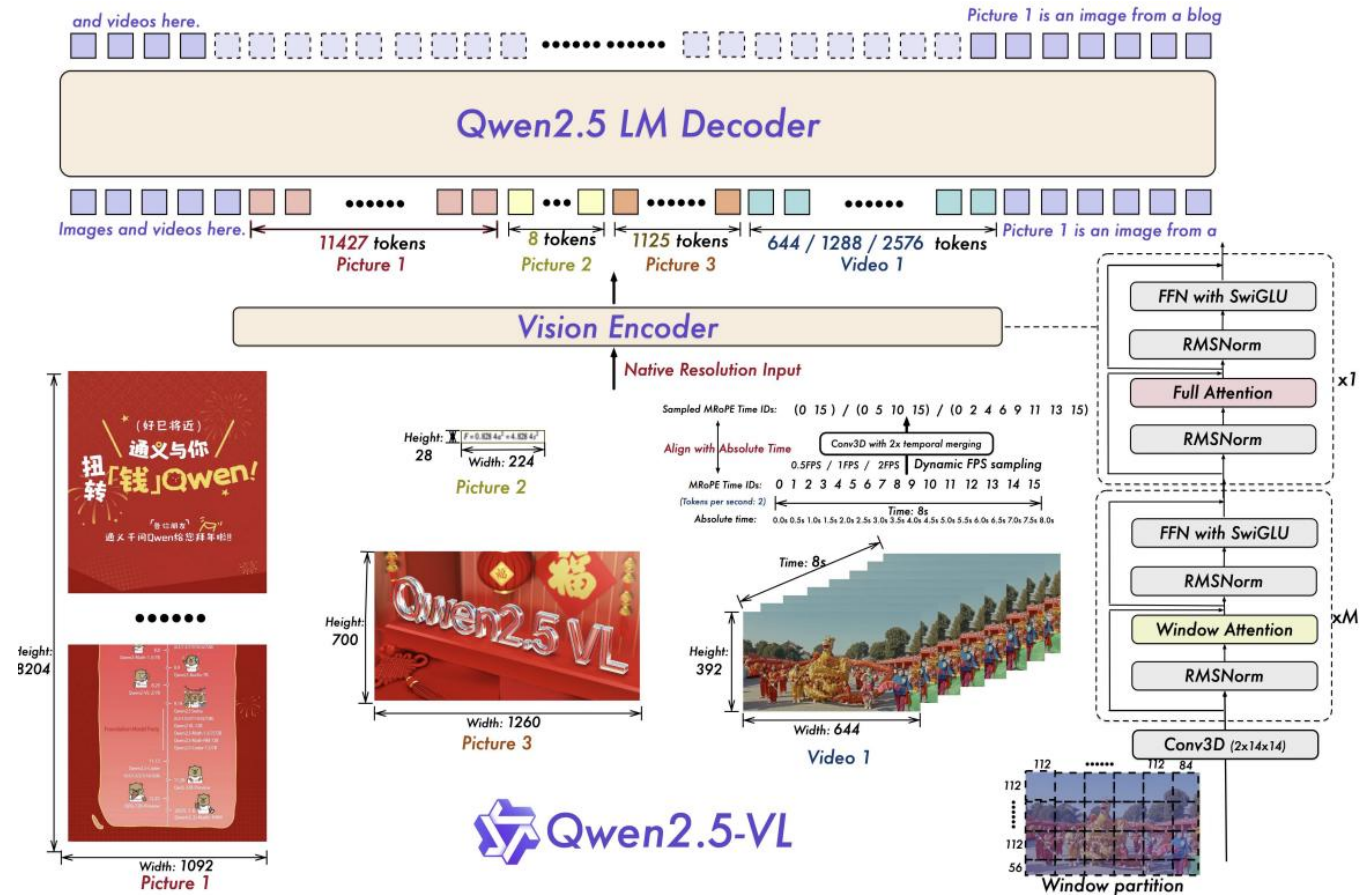
International Conference on Learning Representations (ICLR), 2026

Large Vision-Language Model (LVLM)

- High-resolution LVLMs are costly due to excessive vision tokens



LLaVA, NeurIPS 2023



Qwen2.5-VL

Qwen2.5-VL, ArXiv: 2502.13923

High-Resolution Visual Reasoning




What is the color of the straw?

2250×1500
Example from V* Bench

“Thinking with Images” via RL

Q: What is the direction the arrow below text "PARKING" point?

Original Image:



Question: What is the direction the arrow below text "PRAKING" point?

Turn1: The user is asking for the direction of an arrow located below the text "PRAKING". ... I can see a sign on the right side of the road, below a traffic light. ... It is likely that this sign has the text "PARKING" on it. ... The sign is located approximately in the region defined by the coordinates [0.7, 0.4, 1.0, 0.7]. This area seems to contain the sign and the arrow in question.

Turn2: Observation 2

Turn4: Observation 4

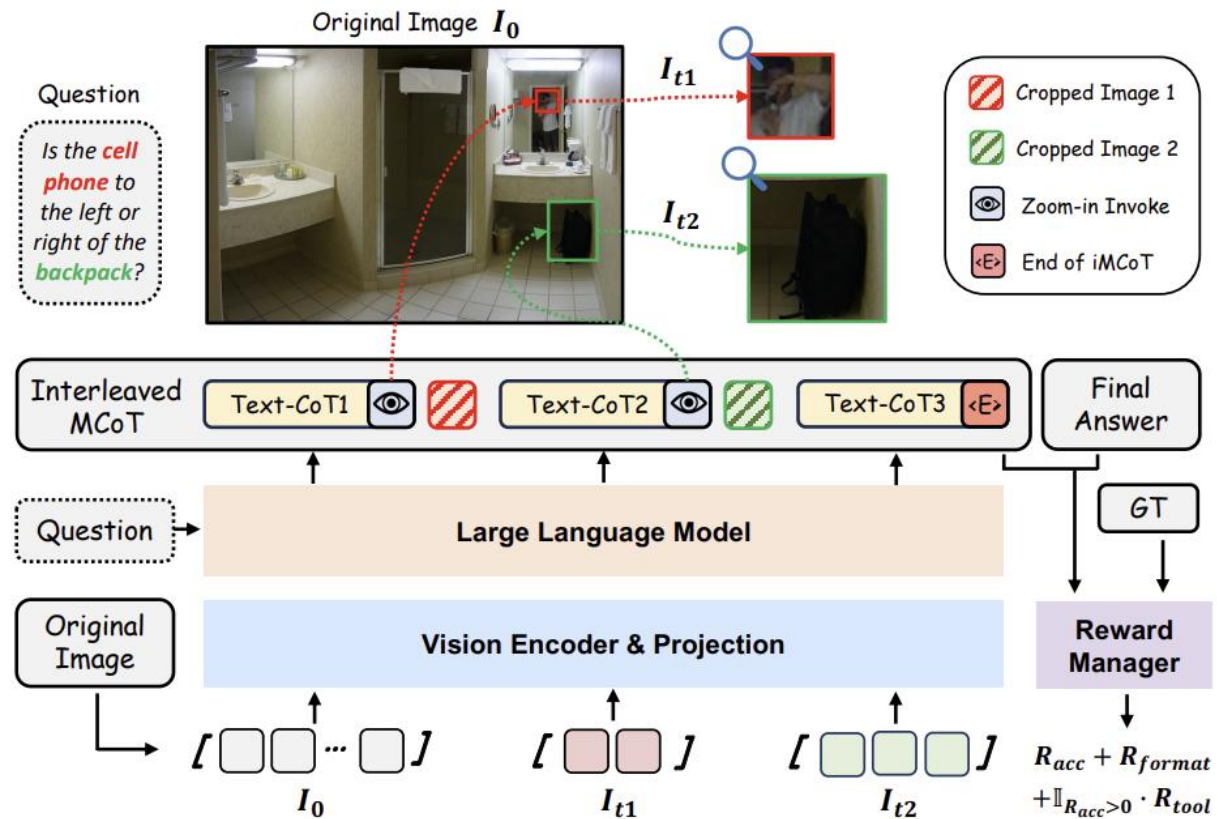
Turn1: Observation 1

Turn5: Observation 5

Turn4: The user is asking for the direction of the arrow below the text "PRAKING". In Observation 4, I have a zoomed-in view of the signs on the right side of the intersection. In the bottom right corner of this image, I can see another street sign. It's a bit small, but it appears to have the word "PARKING". ... I need to zoom in even closer on this specific sign to see the direction of the arrow clearly. The sign is in the bottom right corner of Observation 4, within the region defined approximately by the coordinates [0.7, 0.7, 1.0, 1.0].

Turn5: The user is asking for the direction of the arrow below the text "PRAKING". In Observation 5, I have successfully located the relevant sign. The sign is clearly visible and reads "PARKING" with an arrow pointing to the right. ... Therefore, I have found the information needed to answer the question. The direction of the arrow is to the right.

A: Right



✓ Processing high-resolution input images at the initial stage is costly

- (left) Mini-o3, ICLR 2026
- (right) DeepEyes, ICLR 2026

Effectiveness of High-Resolution Task-Relevant Cues

- Input: Question + Low-resolution original image + High-resolution critical GT area

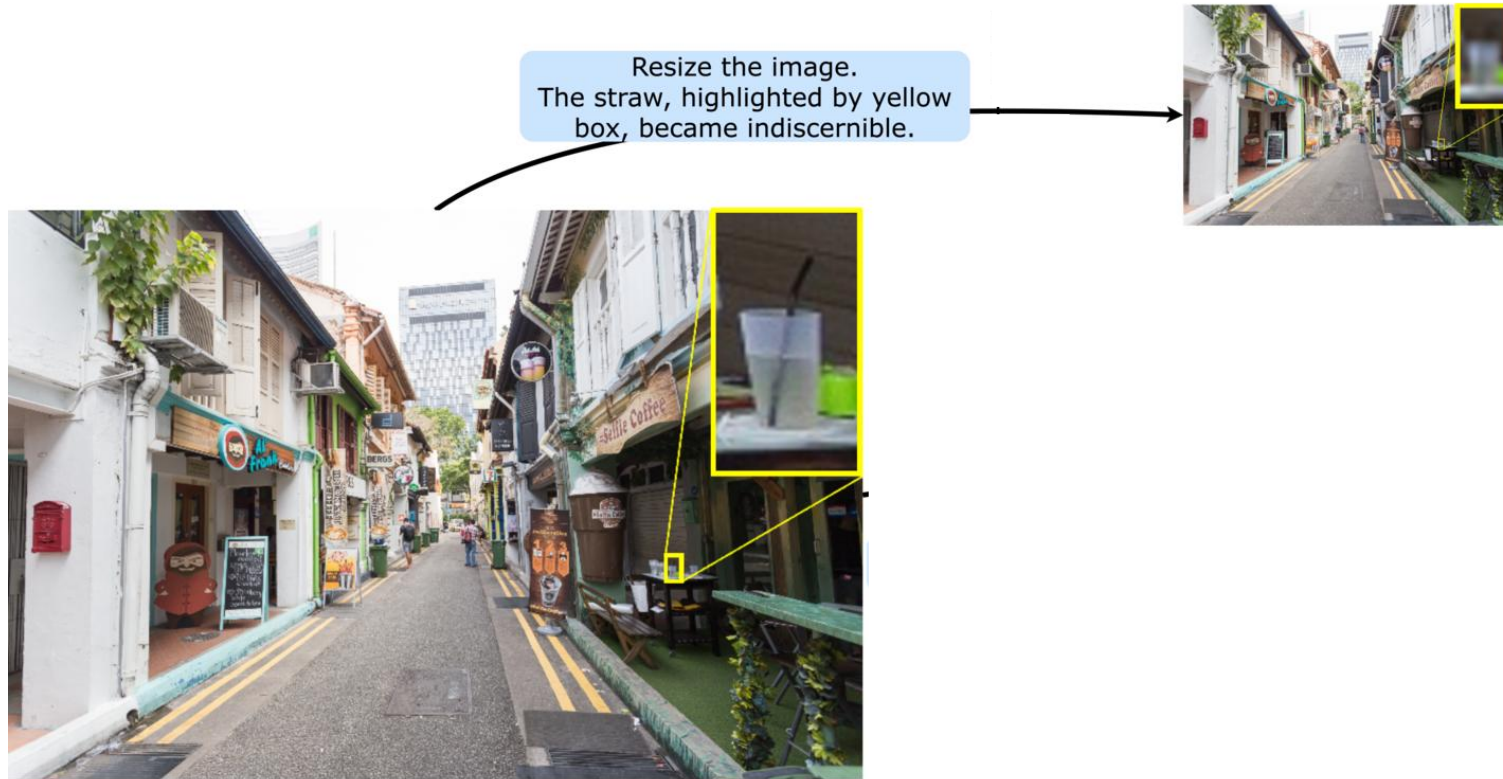
What is the color of the straw?



- Results: High-resolution task-relevant regions are sufficient for VQA, even without training

Approx. resol.	LR orig img	HR critical GT area	V*
	Pixel const.	Task-relevant region	
3584×3584 →	16384×28×28	✗	77.0
1000×1000 →	1280×28×28	✗	64.9
700×700 →	640×28×28	✗	56.5
	640×28×28	✓	77.0

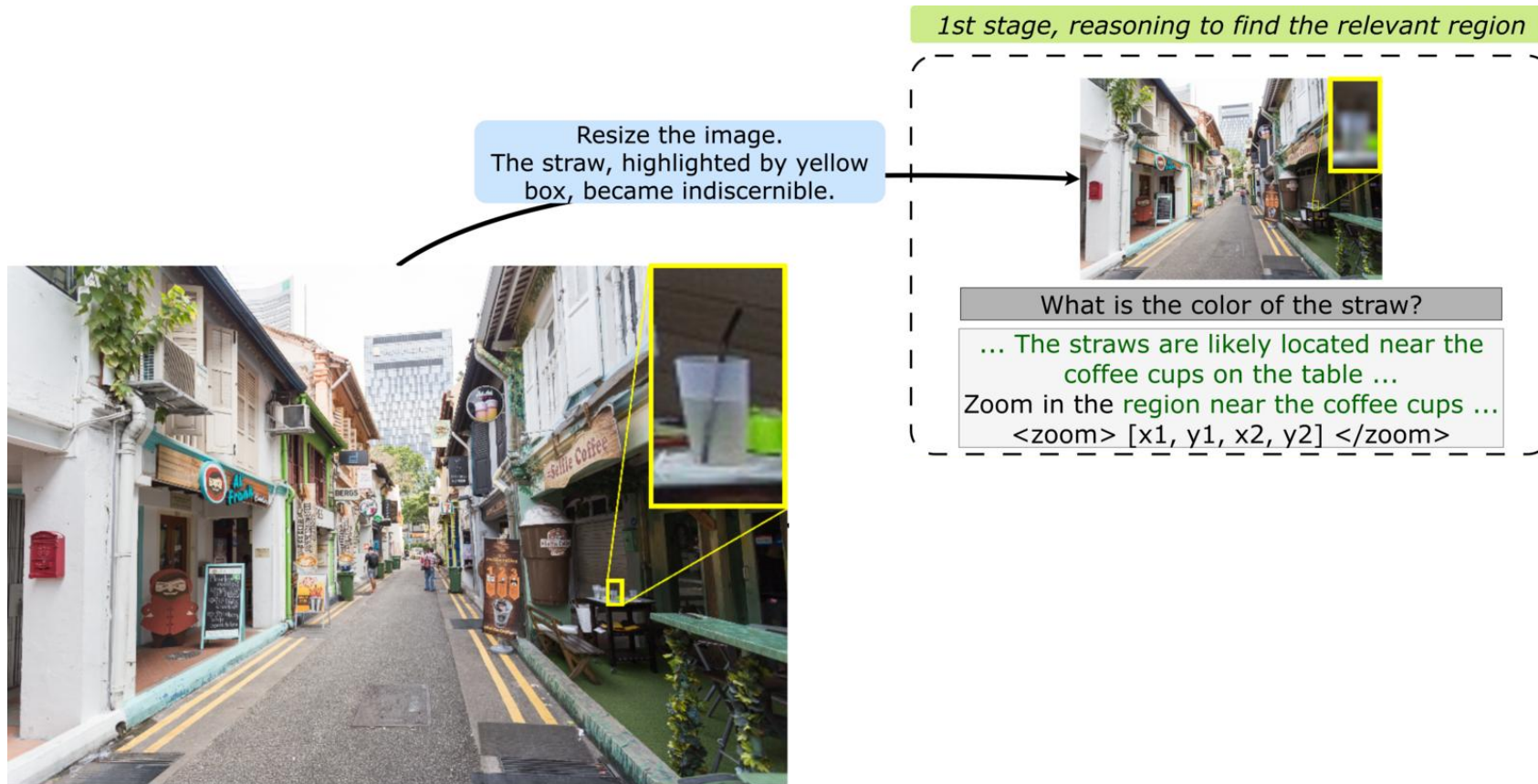
Coarse-to-Fine Reasoning Pipeline



For efficient (①)
visual reasoning,

- ① Downsample high-resolution input image
 - small objects blurred

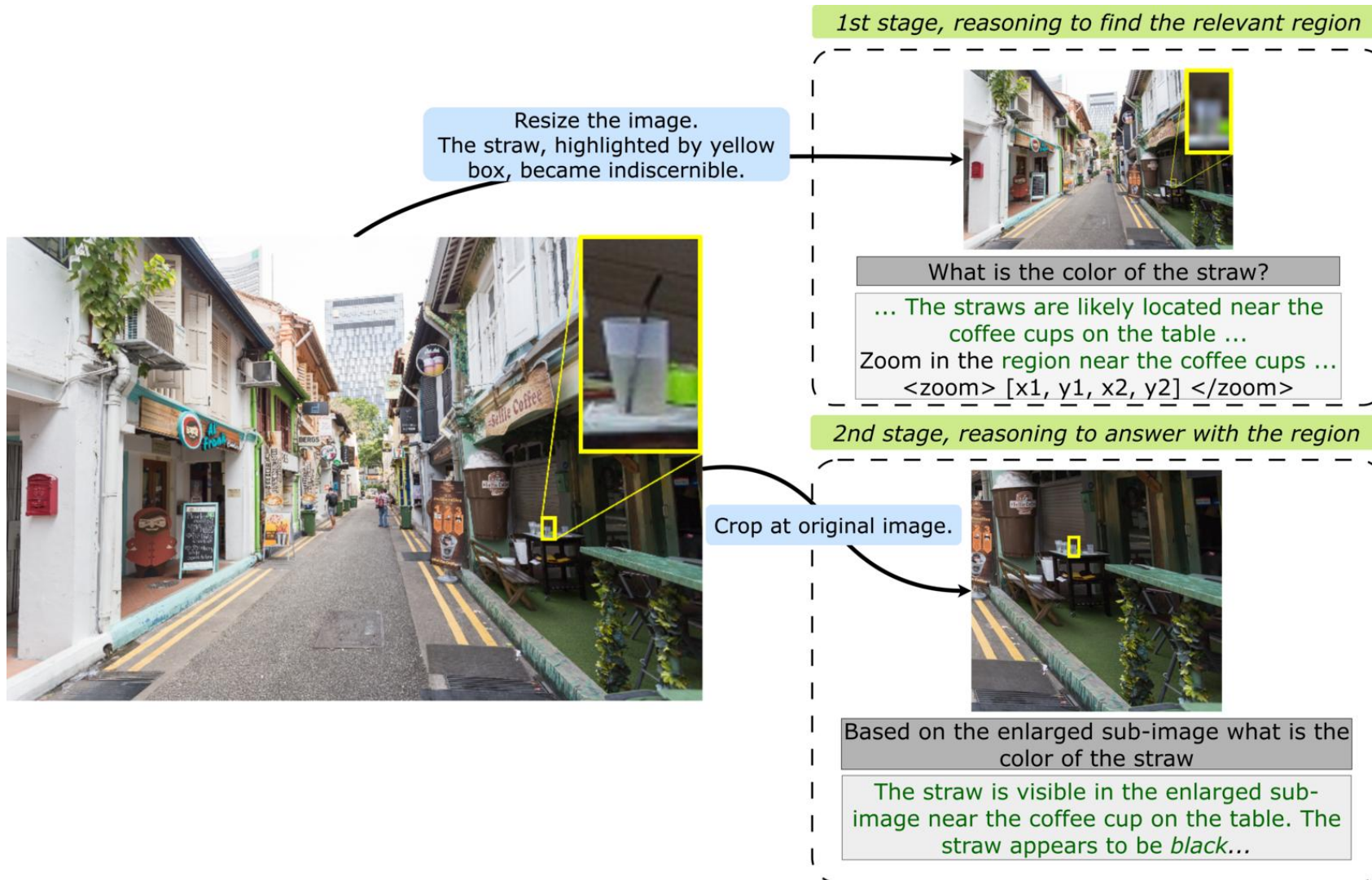
Coarse-to-Fine Reasoning Pipeline



For efficient (①) & accurate (②) visual reasoning,

- ① Downsample high-resolution input image
 - small objects blurred
- ② Use coarse context cues (e.g., cup, table)
 - to find relevant regions

Coarse-to-Fine Reasoning Pipeline



For efficient (①) & accurate (②-③) visual reasoning,

- ① Downsample high-resolution input image
 - small objects blurred
- ② Use coarse context cues (e.g., cup, table)
 - to find relevant regions
- ③ Re-encode selected regions at high resolution
 - to recover fine details

Prior Work: Sensitive to Resolution

(a) Vision Token Efficiency ✗
Accurate Reasoning ✓

The straw is tiny but still discernible.



What is the color of the straw?

Looking at the image, we focus on the coffee cups and straws.

Successfully located the straw



... the image shows that the straws are black. ✓

Powerful, but too much computation cost from massive image tokens.

(b) Vision Token Efficiency ✓
Accurate Reasoning ✗

The straw is indiscernible by downsampling



What is the color of the straw?

The signs are colorful, and there are straws visible in the image. The straws appear to be white.

The model fails to find the straw and instead reasons about a false object.



Crop at high-resolution

The zoom-in tool confirms that the straws are indeed white... The color of the straw is **white**. ✗

Failed to solve the task with downsampled image.

Perception-Driven Reasoning (e.g., DeepEyes)


Ours: Accurate Even at Low Resolution

(a) Vision Token Efficiency ✗
Accurate Reasoning ✓

The straw is tiny but still discernible.



What is the color of the straw?
Looking at the image, we focus on the coffee cups and straws.
Successfully located the straw




... the image shows that the straws are black. ✓


Powerful, but too much computation cost from massive image tokens.

(b) Vision Token Efficiency ✓
Accurate Reasoning ✗

The straw is indiscernible by downsampling



What is the color of the straw?
The signs are colorful, and there are straws visible in the image. The straws appear to be white.
The model fails to find the straw and instead reasons about a false object.




Crop at high-resolution

The zoom-in tool confirms that the straws are indeed white... The color of the straw is **white**. ✗


Failed to solve the task with downsampled image.

(c) Vision Token Efficiency ✓
Accurate Reasoning ✓

The straw is indiscernible by downsampling



What is the color of the straw?
... The straws are likely located near the coffee cups on the table ...
Zoom in the region near the coffee cups ...
The model uses **coarse cues** (coffee cup, table) to locate the straw.



Crop at high-resolution

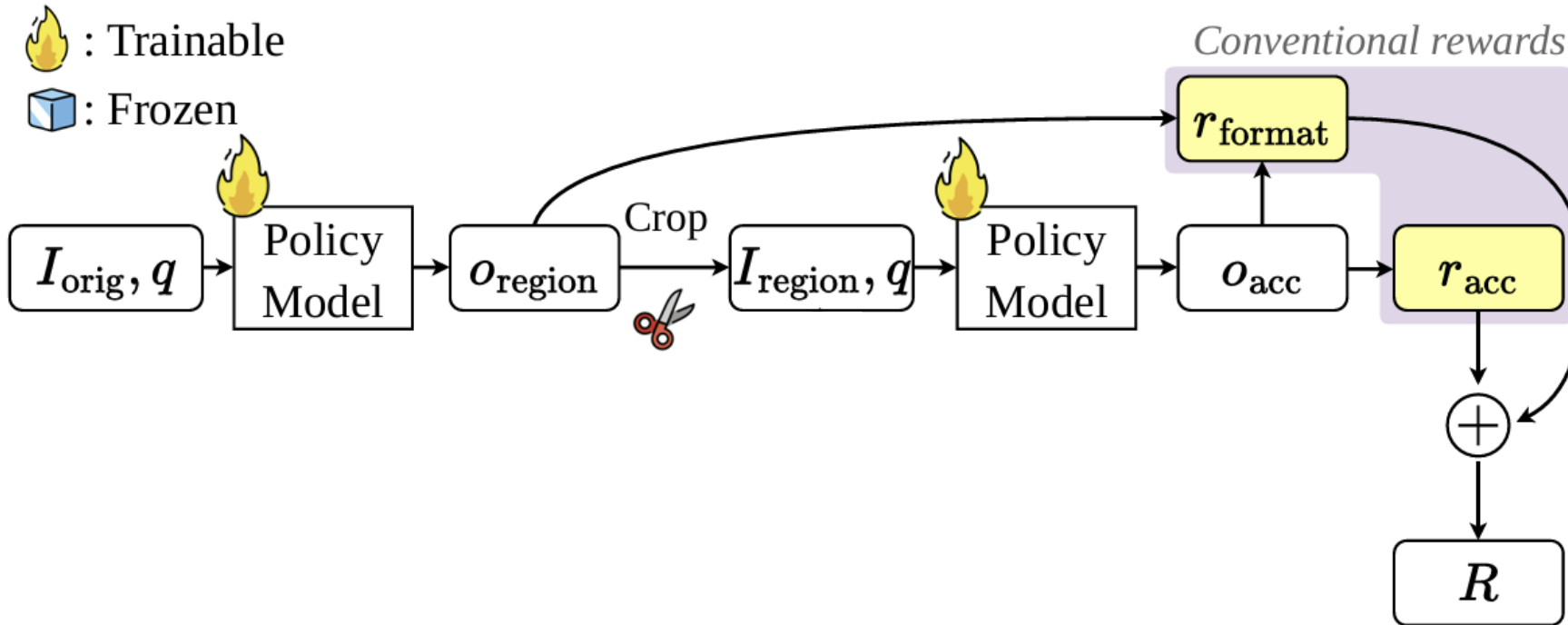
The straw is visible in the enlarged sub-image near the coffee cup on the table. The straw appears to be **black**... ✓

Correctly identified the straw's location and color via effective reasoning.

Perception-Driven Reasoning (e.g., DeepEyes)

Reasoning-driven Perception (Ours)

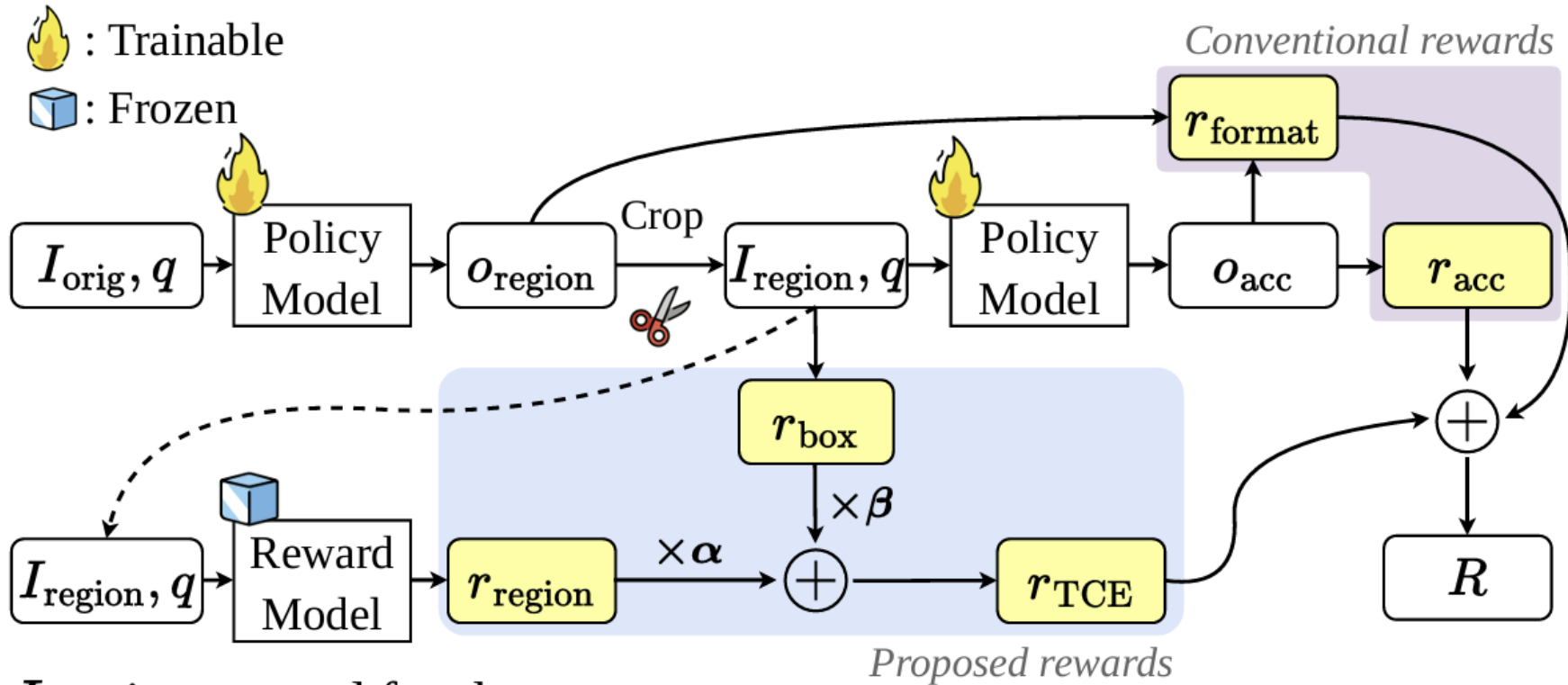
RL-based Training Pipeline



Conventional reward

- Format consistency
 - `<think>...</think>`,
 - `<zoom>...</zoom>`,
 - `<answer>...</answer>`
- Accuracy
 - Model answer vs. GT

RL-based Training Pipeline



I_{orig} is **not** used for the r_{region} .

Conventional reward

- Format consistency
 - `<think>...</think>`,
 - `<zoom>...</zoom>`,
 - `<answer>...</answer>`
- Accuracy
 - Model answer vs. GT

Task-driven Contextual Exploration (TCE) reward

- Region verification
 - Cropped region alone suffices to answer the question.
- Box adjustment
 - Crop remains compact, avoiding full-image cropping.

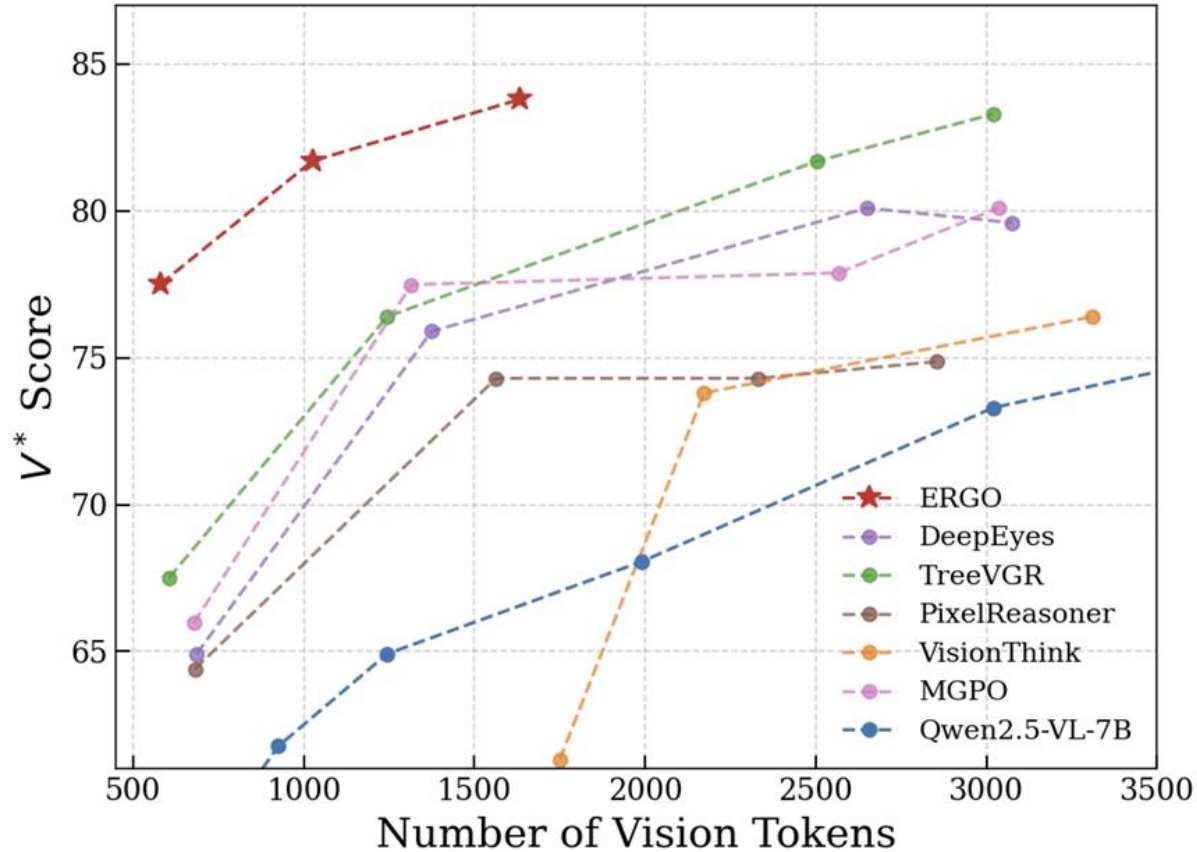
Performance Comparison under Efficiency Constraints

- ERGO: **best** performance on high-resolution benchmarks

Approx Resol.	Pixel Const.	Model	V* Bench	HR Bench ^{4K}	HR Bench ^{8K}	MME-RW ^{Lite}	TreeBench	VisualProbe	Average	
3584×3584	16384×28×28	Qwen2.5-VL-7B-Inst.	77.0	71.1	67.1	46.7	40.0	29.2	52.4	
		Qwen2.5-VL-7B-Inst.	64.9	65.6	56.5	42.6	40.0	15.2	44.8	
1000×1000	1280×28×28	<i>Non-efficiency-oriented Post Training Methods</i>								
		PixelReasoner (Su et al., 2025a)	74.5	66.9	61.3	49.8	40.4	31.8	52.1	
		DeepEyes (Zheng et al., 2025b)	78.5	66.0	60.0	48.9	40.2	39.0	52.5	
		TreeVGR [‡] (Wang et al., 2025a)	76.4	66.4	60.4	47.5	48.2	26.6	53.1	
		MiniO3 (Lai et al., 2025)	81.2	70.0	65.9	36.7	36.3	39.0	52.2	
		<i>Efficiency-oriented Post Training Methods</i>								
		MGPO [†] (Huang et al., 2025b)	77.5	69.8	61.1	44.4	39.3	33.4	52.6	
		VisionThink [‡] (Yang et al., 2025)	73.8	66.1	65.8	49.0	41.0	17.3	49.4	
ERGO	83.8	73.0	69.9	52.6	41.7	42.7	58.4			
700×700	640×28×28	Qwen2.5-VL-7B-Inst.	56.5	57.0	49.9	40.3	40.2	10.9	40.3	
		<i>Non-efficiency-oriented Post Training Methods</i>								
		PixelReasoner (Su et al., 2025a)	67.2	66.5	59.9	47.7	42.3	23.9	48.9	
		DeepEyes (Zheng et al., 2025b)	64.9	64.4	58.3	48.9	38.5	26.8	48.1	
		TreeVGR [‡] (Wang et al., 2025a)	67.0	62.4	54.4	47.5	49.1	24.1	49.2	
		MiniO3 (Lai et al., 2025)	74.9	62.8	57.3	34.4	36.0	31.7	48.2	
		<i>Efficiency-oriented Post Training Methods</i>								
		MGPO [†] (Huang et al., 2025b)	67.5	62.8	57.3	44.4	42.0	29.7	48.7	
VisionThink [‡] (Yang et al., 2025)	61.8	66.9	60.1	46.6	39.5	17.9	45.9			
ERGO	81.7	67.1	66.1	49.6	43.2	35.0	55.2			

Accuracy-Efficiency Trade-off

- ERGO: **fewer** vision tokens, **higher** score & **practical** latency improvement



Pixel Const.	Model	# of vision tokens	V*
16384×28×28	Qwen2.5-VL-7B	4,471	77.0
1280×28×28	PixelReasoner	1,563	74.3
	DeepEyes	1,374	75.9
	TreeVGR	1,244	76.4
	MGPO	1,315	77.5
	VisionThink	1,749	73.8
	MiniO3	1,981	81.2
	ERGO	1,632	83.8
640×28×28	ERGO	1,025	81.7

Pixel Const.	Model	Max. tool cnt	V*	Latency (s)
16384×28×28	Qwen2.5-VL-7B	–	77.0	4.89
640×28×28	DeepEyes	4	64.9	3.42
		2	63.9	3.07
		1	64.4	2.18
640×28×28	MiniO3	4	74.9	5.35
		2	61.8	3.87
		1	41.4	2.03
640×28×28	ERGO	1	81.7	1.61

vLLM on 1×H100 (batch=16)

Ablation Studies

No.	Method	r_{acc}	r_{region}	r_{box}	RW	Avg.
	Qwen2.5-VL-7B					52.4
Ⓐ	r_{acc} only	✓				53.5
Ⓑ	r_{region} only		✓			51.4
Ⓒ	+box adj. reward		✓	✓		54.9
Ⓓ	+reward weighting (RW)		✓	✓	✓	55.3
Ⓔ	ERGO	✓	✓	✓	✓	58.4

(a) Reward design

No.	(α, β)	Avg.
(i)	(1.0, 1.0)	56.2
(ii)	(1.0, 2.0)	54.3
(iii)	(0.5, 0.25)	56.8
(iv)	(1.0, 0.5)	58.4

(b) TCE reward weight

Parameter size	Average
3B	55.6
7B	56.4
72B	58.4

(c) Parameter size

Reward model	Average
GLM4.5V-108B	58.2
InternVL3-78B	57.2
Qwen2.5-VL-72B	58.4

(d) Reward model

γ	Function	Average
0.4	step	57.8
0.8	step	56.2
0.6	step	58.4
0.6	linear anneal.	51.0

(e) Box adjustment reward

ERGO (Efficient Reasoning & Guided Observation)

- Contributions to High-Resolution VQA
 - **Coarse-to-fine reasoning** pipeline
 - → Identify task-relevant regions at low resolution, then refine in high resolution
 - **Efficiency-oriented RL training** (without reasoning SFT data or additional labeling)
 - → Learn to select informative regions using contextual cues
- Results
 - High accuracy on high-resolution benchmarks
 - Fewer vision tokens & Faster latency