Zooming without Zooming: Region-to-Image Distillation for Fine-Grained Multimodal Perception

Lai Wei, *Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, Weiran Huang*

ICML 2026

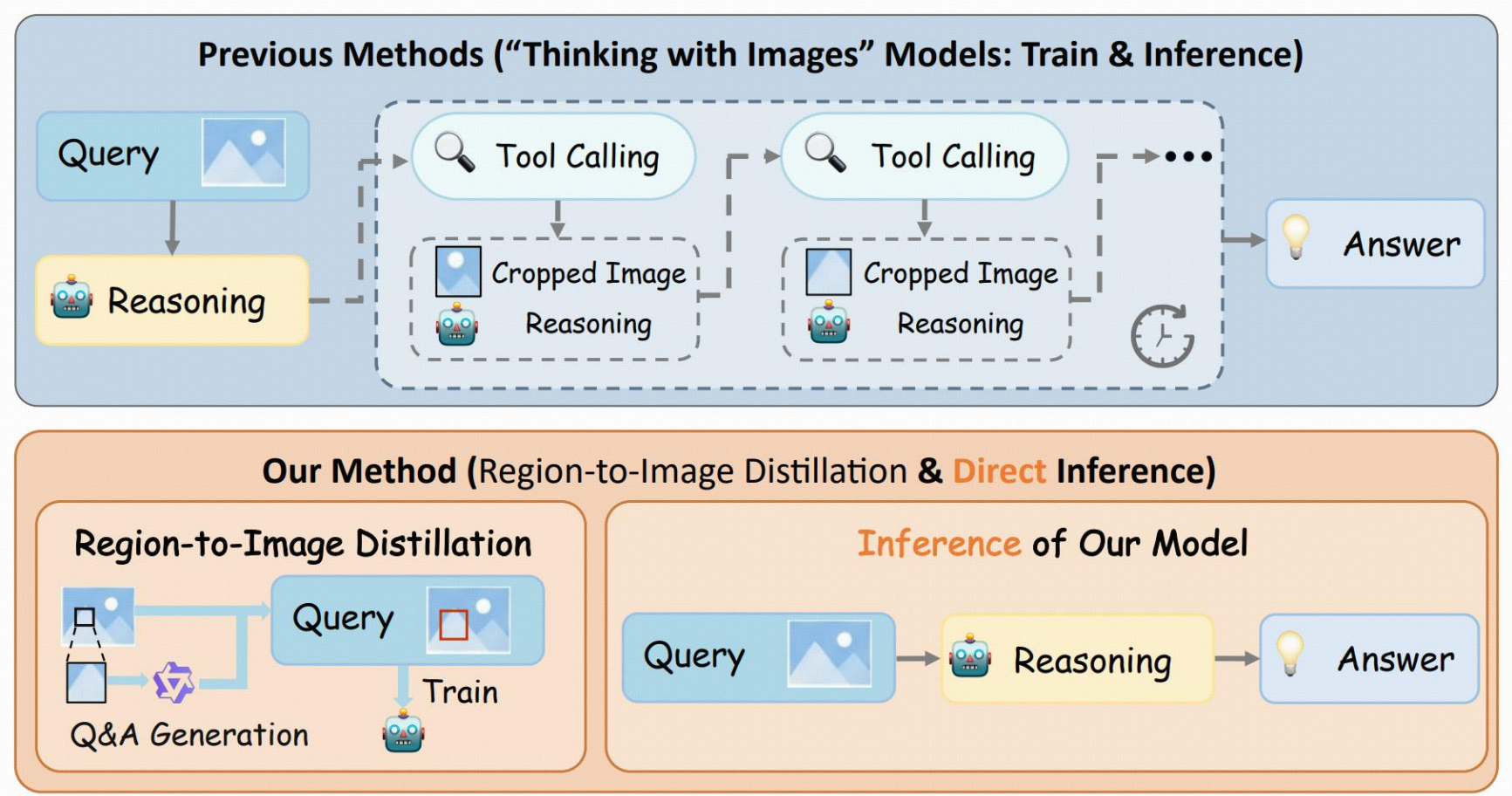
MIFA LAB

ANT GROUP

北京中关村学院
Zhongguancun Academy

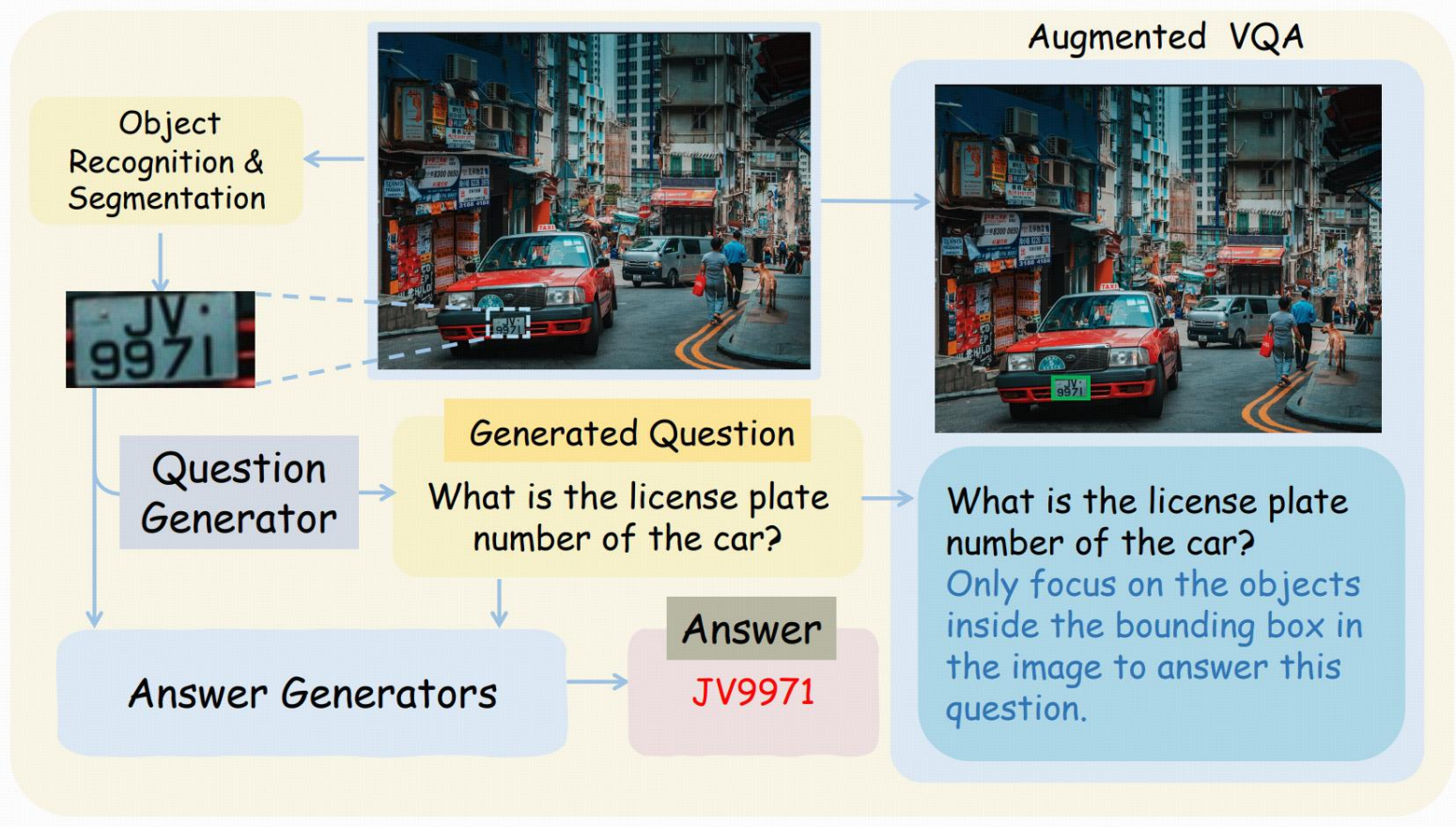
Motivation

Multi-modal Large Language Models (MLLMs) excel at broad visual understanding but still struggle with fine-grained perception, where decisive evidence is small and easily overwhelmed by global context. Recent "**Thinking-with-Images**" methods alleviate this by iteratively zooming in and out regions of interest during inference, but **incur high latency due to repeated tool calls and visual re-encoding**.



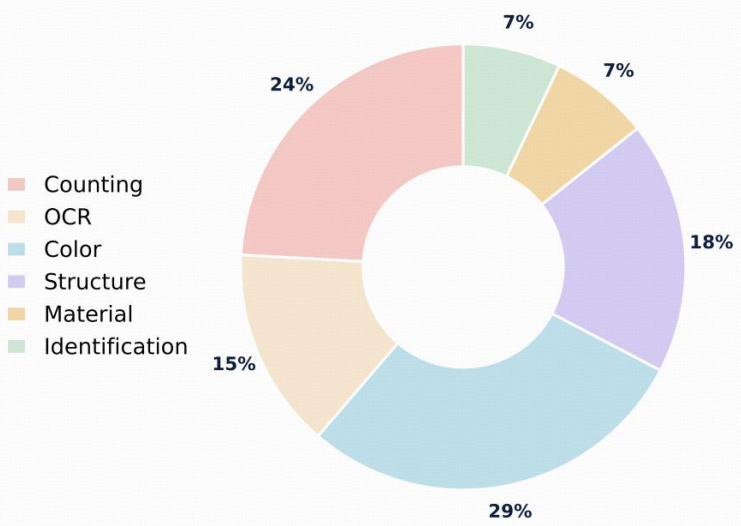
Method

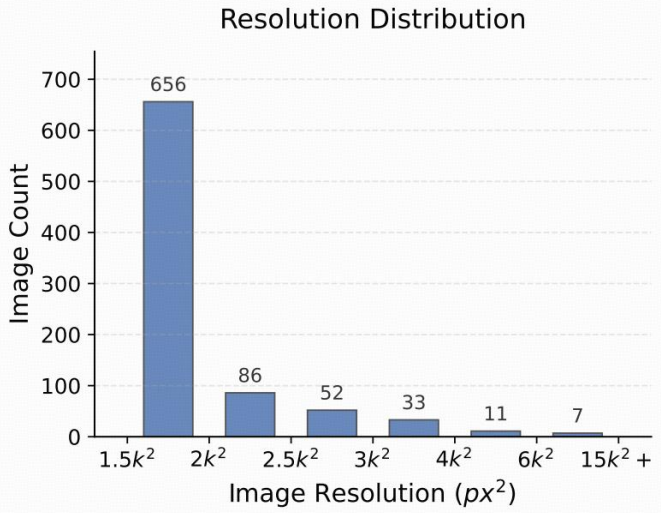
To address this, we propose **Region-to-Image Distillation**, which transforms zooming from an inference-time tool into a training-time primitive, thereby **internalizing the benefits of agentic zooming into a single forward pass** of an MLLM. In particular, we first zoom in to micro-cropped regions to let strong teacher models generate high-quality VQA data, and then distill this region-grounded supervision back to the full image. After training on such data, the smaller student model improves “single-glance” fine-grained perception without tool use.

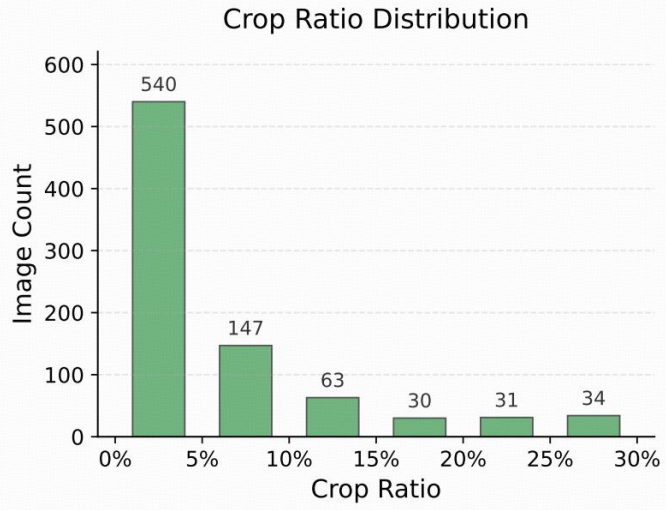


ZoomBench

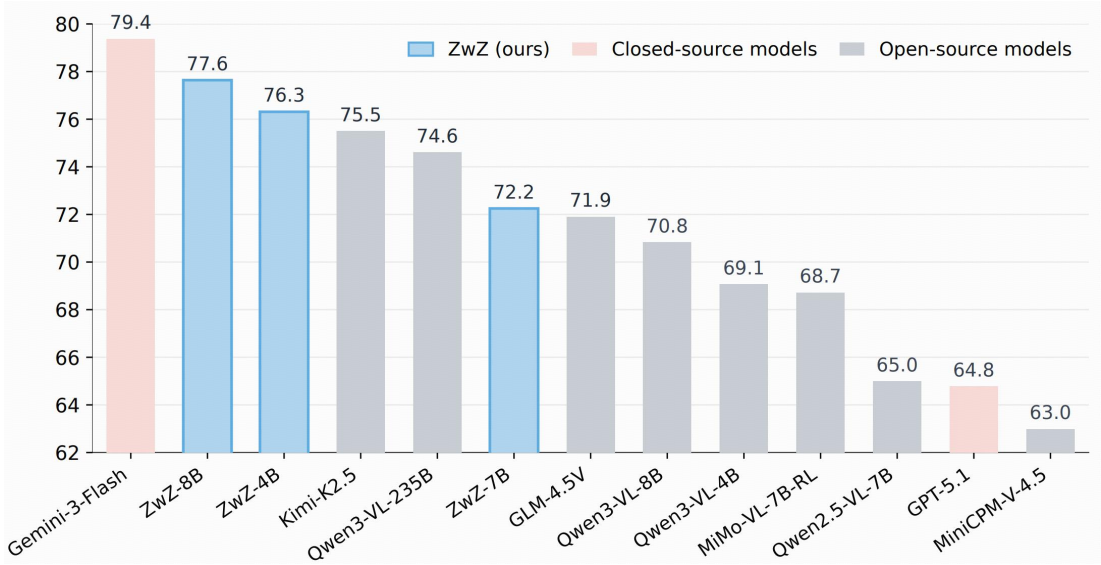
Benchmarks	Data Construction					Evaluation Protocol	
	Question Collection	Question Format	Evidence Annotation	Dimension	Difficulty	Dual-View	Interpretability
CV-Bench	Manual, Templated	MCQ	✗	4	24.7	✗	✗
VStar	Manual, Templated	MCQ	✗	2	19.9	✗	✗
HR-Bench	Hybrid, Templated	MCQ	✗	6	29.6	✗	✗
MME-RealWorld	Manual	MCQ	✗	10+	37.4	✗	✗
TreeBench	Hybrid	MCQ	Manual	10	63.0	✗	✗
FINERS-4k	Manual	MCQ, OQ	✗	4	33.0	✗	✗
ZoomBench (Ours)	**Hybrid (R2I)**	**MCQ, OQ**	**Auto**	6	57.5	✓	✓







Main Experiments



Our ZwZ models achieve substantial performance gains across all evaluated fine-grained perception benchmarks, along with general multi-modal cognition improvements on many benchmarks such as visual reasoning, AIGC detection and GUI agents.

Table 2. Main results on various benchmarks. We report accuracy (%) for each model. Among open-source models (except GPT-5.1 and Gemini-3-Flash), the best results are highlighted in **bold**, and the second-best are underlined. ZwZ consistently improves over the corresponding Qwen-VL baselines, achieving the best overall average among open-source models.

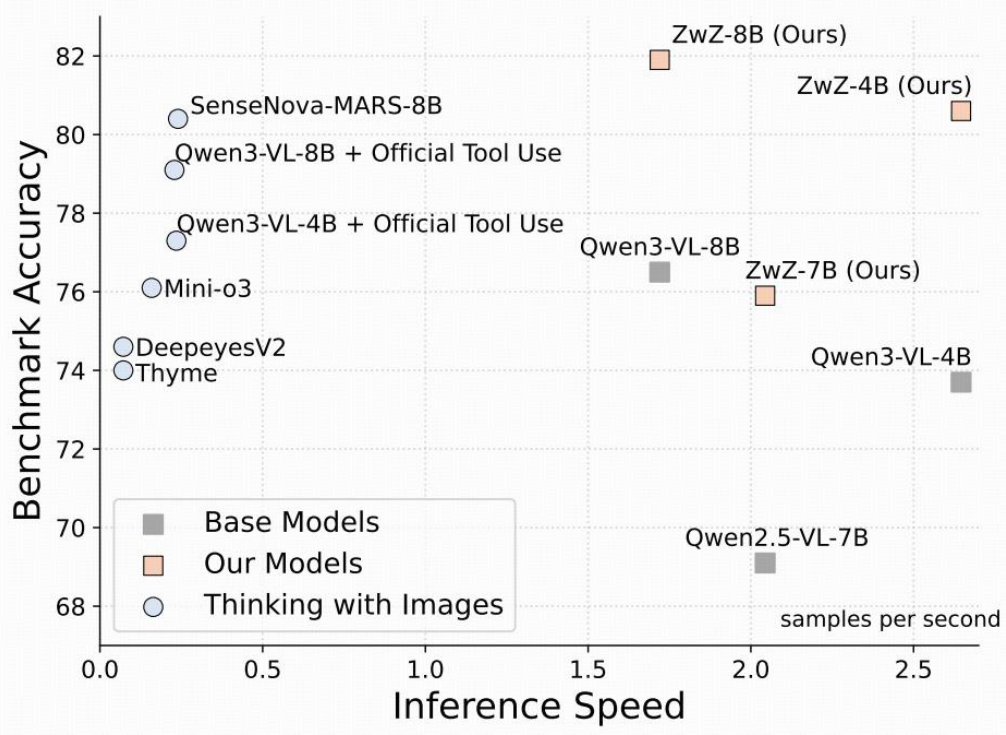
Models	General Perception								Specific Perception		OOD Generalization		Avg
	ZoomBench	HR-4K	HR-8K	VStar	CV-B.	MME-RW-en	MME-RW-cn	GP-Avg	CountQA	ColorB.	MMStar	BabyVision	
Closed-Source Models													
GPT-5.1	47.22	67.00	65.25	70.16	84.22	64.04	55.57	64.78	31.41	83.43	71.60	13.92	59.44
Gemini-3-Flash	59.29	87.88	85.00	86.39	89.57	74.86	72.62	79.37	66.88	85.47	83.60	34.51	75.10
Open-Source Models													
Qwen3-VL-4B	40.24	78.25	72.88	80.10	84.95	63.47	63.63	69.07	28.14	81.63	69.73	13.66	61.52
Qwen2.5-VL-7B	42.49	71.62	67.88	78.53	75.34	60.80	58.30	64.99	18.91	76.36	61.93	12.89	56.82
Qwen3-VL-8B	37.87	78.88	74.63	86.39	85.44	65.96	66.67	70.83	28.99	82.77	70.93	12.89	62.86
MiMo-VL-7B-RL	45.09	74.38	72.88	81.15	84.31	63.40	59.78	68.71	28.27	82.80	73.53	16.24	61.98
MiniCPM-V-4.5 (9B)	42.60	69.88	63.62	70.16	80.25	58.16	56.23	62.99	23.43	79.75	67.87	14.95	56.99
GLM-4.5V (108B)	49.23	81.63	74.88	83.25	87.59	66.04	60.71	71.90	35.93	84.59	75.87	15.72	65.04
Qwen3-VL-235B-A22B	49.11	**84.50**	81.62	87.96	86.72	67.07	65.29	74.61	40.58	85.62	76.33	18.30	67.55
Kimi-K2.5 (1T)	56.33	81.87	75.38	85.86	**89.18**	**71.51**	68.40	75.50	**52.81**	**86.61**	**81.80**	**33.25**	**71.18**
Our Models													
ZwZ-4B (Ours)	55.74	81.75	79.50	**92.67**	87.90	68.52	68.09	76.31	30.82	83.08	71.13	16.24	66.86
ZwZ-7B (Ours)	55.62	75.38	73.25	88.48	79.83	66.21	66.96	72.25	20.72	80.82	63.40	15.98	62.42
ZwZ-8B (Ours)	**58.11**	84.38	**82.00**	91.10	87.40	69.87	**70.59**	**77.64**	32.40	83.59	73.13	16.75	68.12

Table 3. Comparison of different datasets. The best results are highlighted in **bold**, and the second-best are underlined. Model trained on our synthetic dataset achieves superior performance.

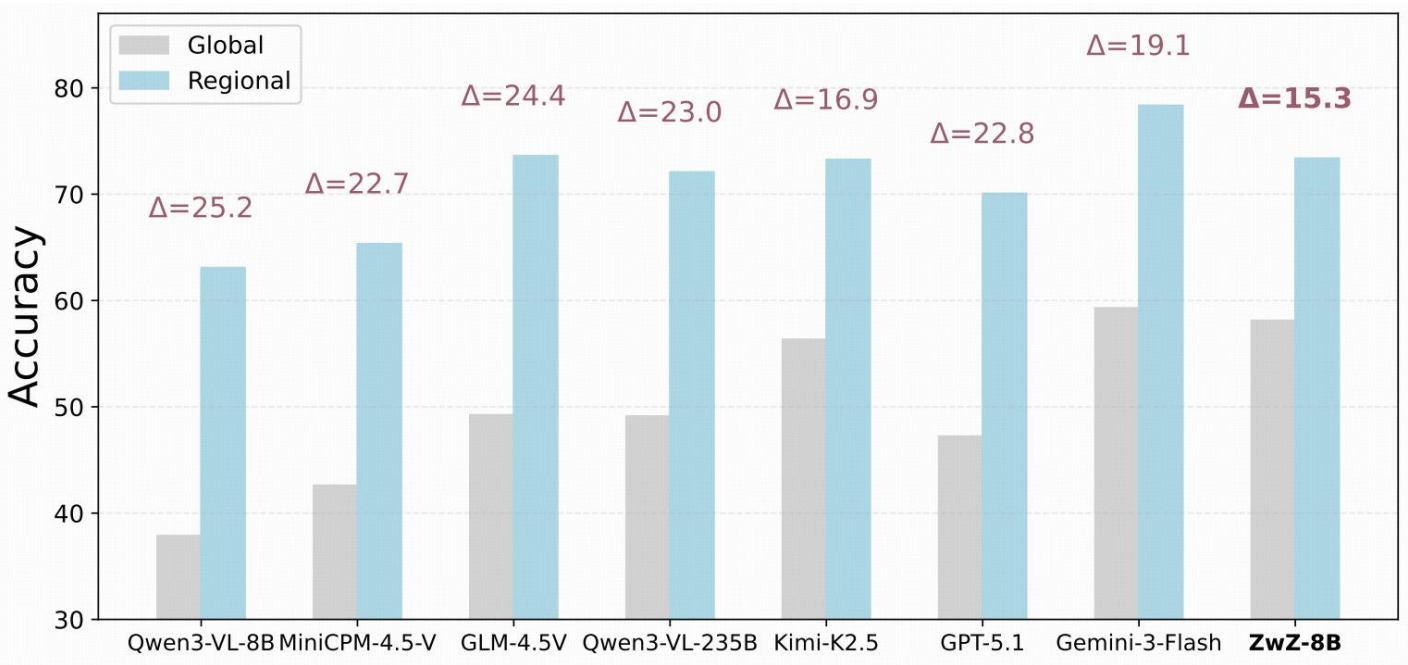
Training Data	Size	Synthetic?	General Perception							Specific Perception		OOD Generalization		Avg
			ZoomBench	HR-4K	HR-8K	VStar	CV-B.	MME-RW-en	MME-RW-cn	CountQA	ColorB.	MMStar	BabyVision	
Qwen3-VL-8B	-	-	37.87	78.88	74.63	86.39	85.44	65.96	66.67	28.99	82.77	70.93	12.89	62.86
+ DeepEyes data	47K	✗	45.80	**84.75**	80.12	88.48	**88.62**	68.71	71.03	30.56	82.94	72.67	10.57	65.84
+ Thyme-RL data	55K	✗	40.93	82.75	78.05	**91.62**	88.29	**71.14**	**71.17**	30.96	**83.98**	**74.87**	13.66	66.13
+ Oasis data	500K	✓	37.51	81.50	77.62	83.77	87.07	64.96	70.26	27.23	81.78	70.53	13.40	63.24
+ MM-Self-Instruct data	65K	✓	37.04	81.38	78.75	82.72	86.32	68.02	67.48	28.73	83.13	72.00	15.21	63.71
+ our data	10K	✓	52.90	82.88	81.38	**91.62**	87.78	68.43	69.63	**33.97**	83.19	72.53	**16.75**	67.37
+ our data	74K	✓	**58.11**	84.38	**82.00**	91.10	87.40	69.87	70.59	32.40	83.59	73.13	**16.75**	**68.12**

Further Experiments

ZwZ attains higher accuracy with substantially lower inference cost (around 10× faster inference speed) than agentic and tool-use baselines.



In ZoomBench, we evaluate each instance under two conditions: Global-View (full image) and Regional-View (micro-crop). The zooming gap is defined as the performance difference between the two. It measures how often a model fails to attend to solvable fine-grained evidence under realistic global inputs, capturing the perception bottleneck rather than recognition ability.

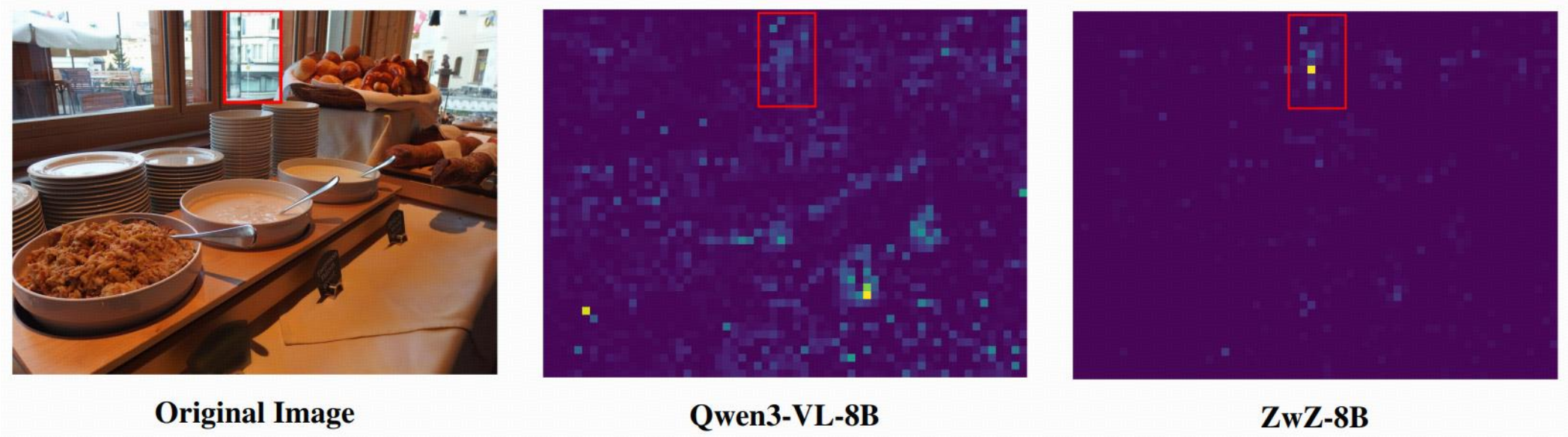


Ablation Study

We ablate three grounding strategies (including our final chosen strategy):
(a) R2I + no-bbox: Directly distilling the crop-based VQA pairs to the full image without any spatial guidance.
(b) R2I + bbox-in-question: Providing the target bounding box coordinates as text (e.g., [x1, y2, x2, y2]) within the prompt.
(c) R2I + bbox-in-image (ours): Overlaying the bounding box directly onto the image.

Strategies	Zoom-Bench	HR-4K	HR-8K	VStar	CountQA	ColorBench	Avg
Direct Synthesis	40.95	82.12	78.12	84.82	32.20	83.45	66.94
R2I							
+ bbox-in-image	52.90	82.28	81.38	91.62	33.97	83.19	**70.89**
+ bbox-in-question	46.98	79.25	77.12	89.53	31.53	82.62	67.84
+ no-bbox	46.27	81.50	80.62	88.48	28.27	83.09	68.04

Attention Map Analysis



Question: What is the shape of the blue sign on the pole? A. circle B. square C. inverted triangle D. rectangle Answer: C

Rethinking “Thinking with Images”

TwI is most valuable when the action produces unpredictable information gain.

Category	Tool/Action for Images (examples)	Predictable?
Information-gain actions *(unpredictable external information)*		
Web search / retrieval	Bring in new, unseen images from external sources (Wu et al., 2025a)	✗
Information-neutral actions *(reformat/reveal existing evidence)*		
Zoom in (crop)[1]	Zooming in key regions for easier perception (Zheng et al., 2025b)	✓
Zoom out	Zooming out for detecting adversarial AIGC (Xu et al., 2025)	✓
Flip	Horizontal/vertical flipping (Xu et al., 2025)	✓
Rotate	Rotation for viewpoint normalization (Zhang et al., 2025h)	✓
Denoise	Remove Gaussian noise (Xu et al., 2025)	✓
2D grounding	2D object detection and segmentation (Lu et al., 2025; Zhou et al., 2025)	✓
3D grounding	Depth estimator; 3D object detection and segmentation (Chen et al., 2025)	✓
Draw images	Draw sketches or auxiliary lines (Wei et al., 2025a; Zhao et al., 2025)	✓

Rethinking “Thinking with Images”

Our framework can be extended to distill many other information-neutral actions.

Algorithm 1 *Region-to-Image Distillation* (R2I)

Require: Raw image pool $\mathcal{D}_{\text{raw}}$; proposal function $\mathcal{P}(\cdot)$; question generator $\mathcal{T}_{\text{gen}}$; teacher ensemble $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$; sparsity threshold τ; max questions per crop K
Ensure: Distilled dataset $\mathcal{D}_{\text{syn}}$

$\mathcal{D}_{\text{syn}} \leftarrow \emptyset$ // initialize
for each image $I \in \mathcal{D}_{\text{raw}}$ **do**
  $\mathcal{B} \leftarrow \mathcal{P}(I)$ // propose candidate boxes via detection/segmentation
  for each box $B \in \mathcal{B}$ and $\text{AREA}(B)/\text{AREA}(I) \leq \tau$ **do**
    $R \leftarrow \text{CROP}(I, B)$ // zoom in: crop region R from I
    $\mathcal{Q}_R \leftarrow \mathcal{T}_{\text{gen}}(R; K)$ // generate K region-answerable questions
    for each question $Q \in \mathcal{Q}_R$ **do**
      $\mathbf{a} \leftarrow [\,]$ // collect teacher answers
      for each teacher $\mathcal{T}_j \in \{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ **do**
        $a_j \leftarrow \mathcal{T}_j(R, Q)$; $\mathbf{a} \leftarrow \mathbf{a} \cup \{a_j\}$ // answer on the crop to reduce hallucination
      end for
      if $\text{HIGHCONSENSUS}(\mathbf{a})$ **then**
        $A \leftarrow \text{CONSENSUSMAP}(\mathbf{a})$ // e.g., majority vote
        $(I', Q') \leftarrow \text{GROUNDTOFULL}(I, B, Q)$ // zoom out: overlay B on I to form I' and add a spatial constraint to form Q'
        $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}_{\text{syn}} \cup \{(I', Q', A)\}$ // store full-image training triplet
      end if
    end for
  end for
end for
$\mathcal{D}_{\text{syn}} \leftarrow \text{REJECTIONSAMPLING}(\mathcal{D}_{\text{syn}})$
return $\mathcal{D}_{\text{syn}}$

Algorithm 2 *A general view of our method.*

Require: Raw image pool $\mathcal{D}_{\text{raw}}$; a tool-call action $f(\cdot)$; question-answer generator $\mathcal{T}$
Ensure: Distilled dataset $\mathcal{D}_{\text{syn}} = \{(I, \widehat{Q}, A)\}$

$\mathcal{D}_{\text{syn}} \leftarrow \emptyset$
for each image $I \in \mathcal{D}_{\text{raw}}$ **do**
  $\widehat{I} = f(I)$
  $(Q, A) \sim \mathcal{T}(\widehat{I})$
  $I, \widehat{Q} = f^{-1}(\widehat{I}, Q)$
  $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}_{\text{syn}} \cup \{(I, \widehat{Q}, A)\}$
end for
return $\mathcal{D}_{\text{syn}}$

Future Works

A promising direction is to develop a unified and dynamic agent policy for token effiency. Based on our analysis, this agent can (i) default to single-pass inference with enhanced perceptual skills, (ii) decide when and how to invoke tools, and (iii) prioritize information-gain TwI actions, while using information-neutral operations only sparingly. This would combine the efficiency of our method with the open-world capability of agentic TwI.

Thank you

Paper Github WeChat