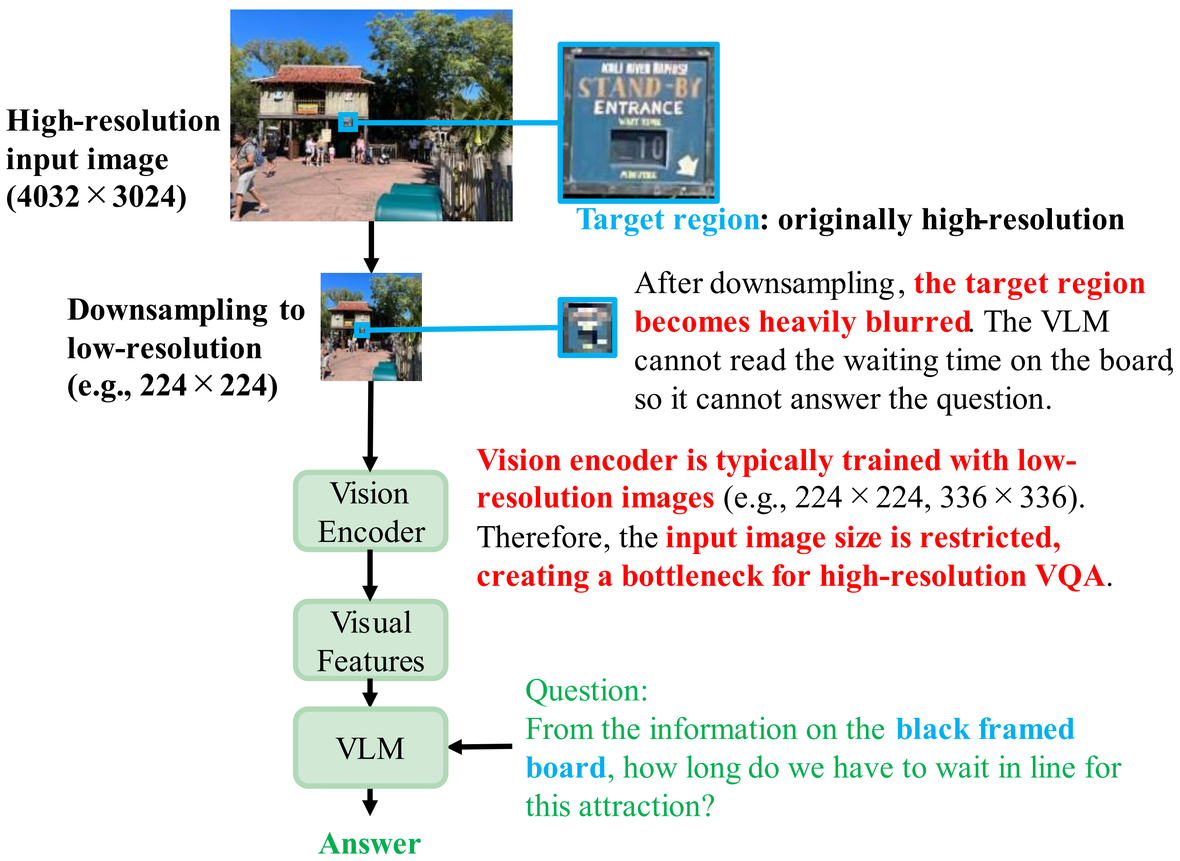
Zero-shot Small-Object VQA by CLIP-Based Patch Extraction

Yuki Kasahara
Artificial Intelligence R&D Dept
Mitsubishi Electric Corporation
Amagasaki, Japan
Kasahara.Yuki@da.MitsubishiElectric.co.jp

Ken Miyamoto
Artificial Intelligence R&D Dept
Mitsubishi Electric Corporation
Amagasaki, Japan
Miyamoto.Ken@bc.MitsubishiElectric.co.jp

1. Introduction

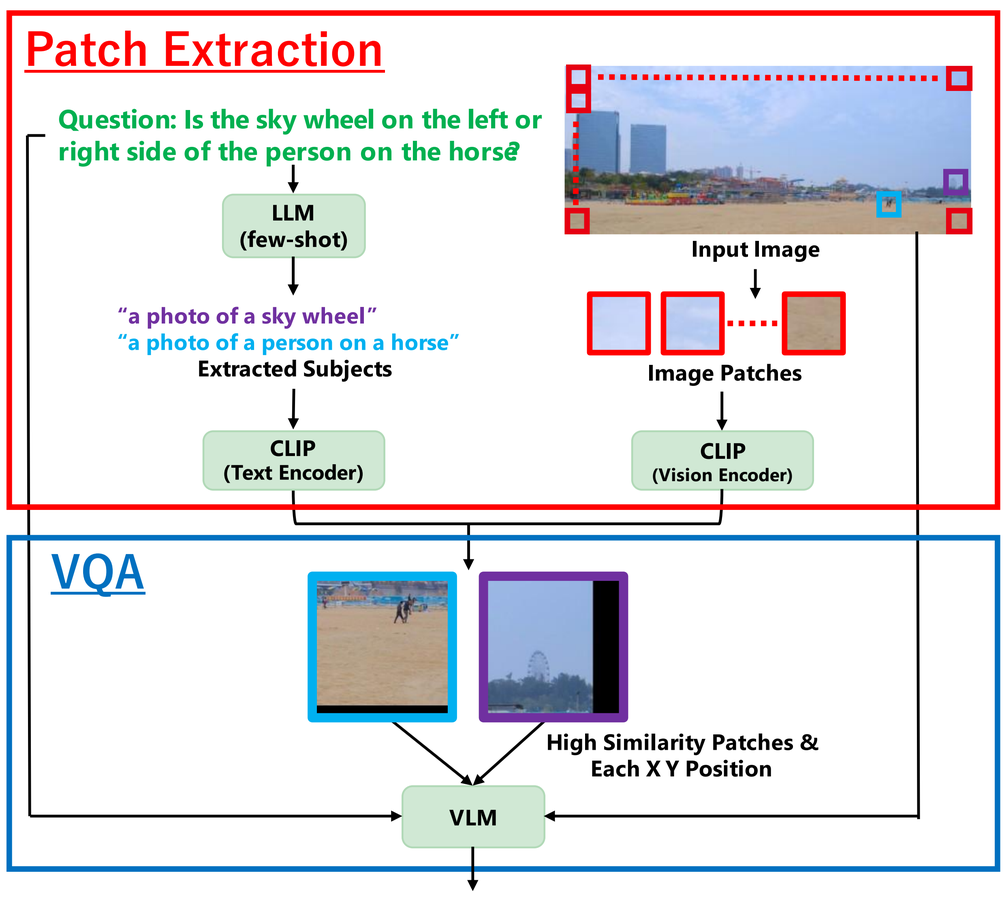
- **Target:** Visual Question Answering (VQA) for small object



Example of VQA that illustrates the resolution bottleneck of conventional models

2. Architecture of the proposed method

- **The points:**
  - **No training is required (Zero-shot learning)**



CLIP-Based Patch Extraction Pipeline

3. Experiments on V* Bench

- **Contents:** high-resolution images (avg 2246 × 1582) and corresponding 191 small-object VQA tasks
- **Task:** Quantitative questions for images
  - Attribute Recognition (115 tasks: color, material, state, etc.)
  - Spatial Relationship Reasoning (76 tasks: positional relations among multiple objects).
- **Metrics**: Accuracy on four-option multiple-choice questions

Attribute Recognition
Question: What is the color of the van ?

2250 × 1500

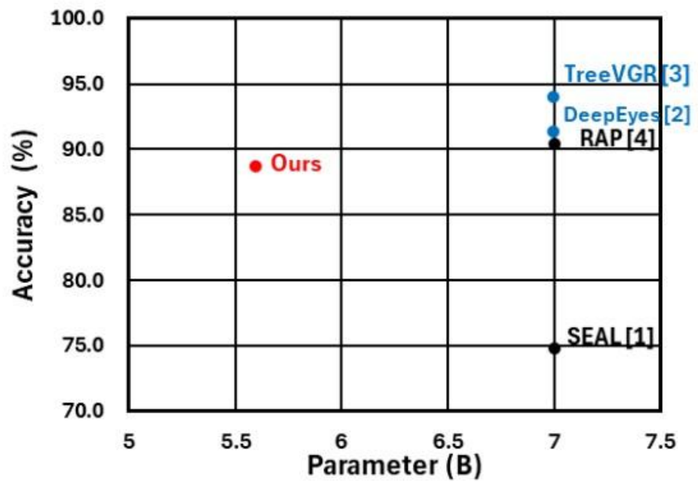
Spatial Relationship Reasoning
Question: Is the vehicle on the left or right side of the blue umbrella ?

2250 × 1500

Images and the corresponding questions in V* Bench

4. Results

- Our method achieves **close to the best attribute accuracy in zero-shot approaches**, though the **parameters of our method are less than others.**



Results on Attribute Recognition of V* Bench. The blue, black and red dots represent RL, zero-shot, and our zero-shot method, respectively.

Results on Attribute Recognition Tasks

- Although our method is **zero-shot**, it approaches the performance of RL-based methods and even achieves a **higher score than ViGoRL**.

Results on V* Bench

Method	Leaning Method	BaseModel	Attibute (%)	Spatial (%)	Overall (%)
SEAL [1]	-	LLaVA-1.5-7B	74.8	76.3	75.4
DeepEyes [2]	RL	Qwen2.5-VL-7B	91.3	88.2	90.1
ViGoRL [3]	RL	Qwen2.5-VL-7B	-	-	86.4
TreeVGR [4]	RL	Qwen2.5-VL-7B	94.0	87.0	91.1
RAP [5]	-	LLaVA-1.5-7B	90.4	96.1	91.1
Ours	**-**	**Phi-4-multimodal-5.6B**	**88.7**	**75.0**	**83.3**
Ours	**-**	**Qwen2.5-VL-7B**	**87.0**	**86.8**	**86.9**

5. References

[1] P. Wu and S. Xie, ``V*: Guided visual search as a core mechanism in multimodal llms," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13084--13094.
[2] Z. Zheng, M. Yang, J. Hong, C. Zhao, G. Xu, L. Yang, C. Shen, and X. Yu, ``DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning," arXiv preprint arXiv:2505.14362, 2025.
[3] G. Sarch, S. Saha, N. Khandelwal, A. Jain, M. J. Tarr, A. Kumar, and K. Fragkiadaki, ``Grounded Reinforcement Learning for Visual Reasoning," arXiv preprint arXiv:2505.23678, 2025.
[4] H. Wang, X. Li, Z. Huang, A. Wang, J. Wang, T. Zhang, J. Zheng, S. Bai, Z. Kang, J. Feng, and others, ``Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology," arXiv preprint arXiv:2507.07999, 2025.
[5] W. Wang, Y. Jing, L. Ding, Y. Wang, L. Shen, Y. Luo, B. Du, and D. Tao, ``Retrieval-augmented perception: High-resolution image perception meets visual rag," arXiv preprint arXiv:2503.01222, 2025.