FOCUS: Internal MLLM Representations for Efficient Fine-Grained Visual Question Answering

Liangyu Zhong*, Fabio Rosenthal*, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, Leo Schwinn

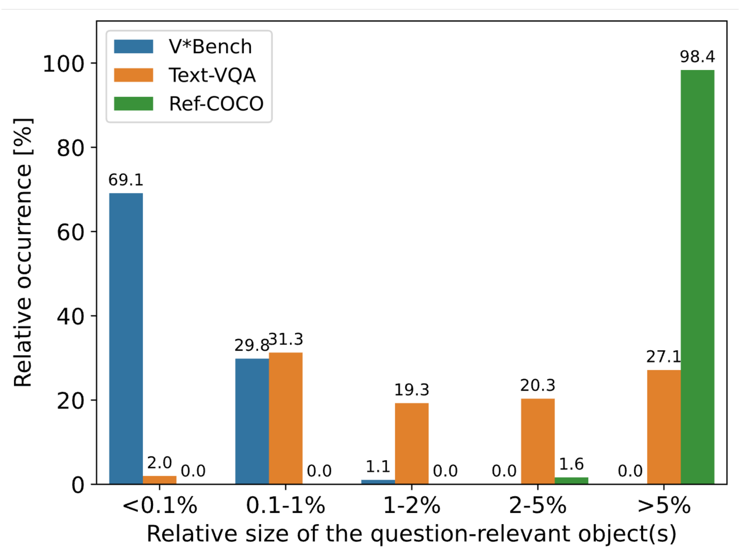
*equal contribution

TL;DR:

We propose a training-free visual cropping method that leverages MLLM-internal representations for VQA tasks focusing on small details, achieving strong performance with 3 - 6.5x higher efficiency than prior methods.

Motivation

- Most VQA datasets contain images with large objects
- On datasets with small relevant objects, **MLLM performance drops significantly**
- **Providing the relevant image region** substantially improves MLLM accuracy
- **Visual cropping** methods can identify relevant regions at test time

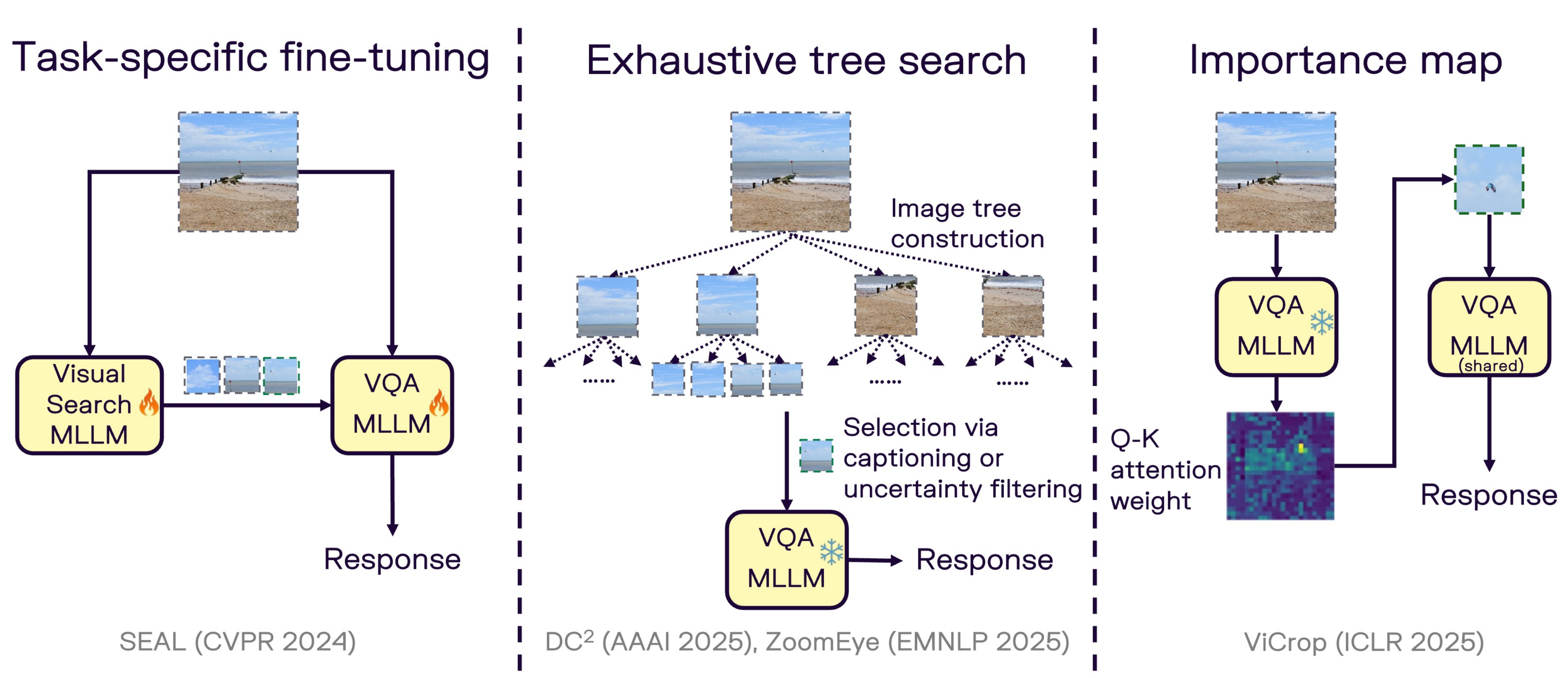


Model	Accuracy on V*Bench [%]
Random guessing	35.99
LLaVA-1.5 (full image)	48.60
LLaVA-1.5 (only GT region)	87.20 (+38.6 pp.)

Recent Visual Cropping Approaches

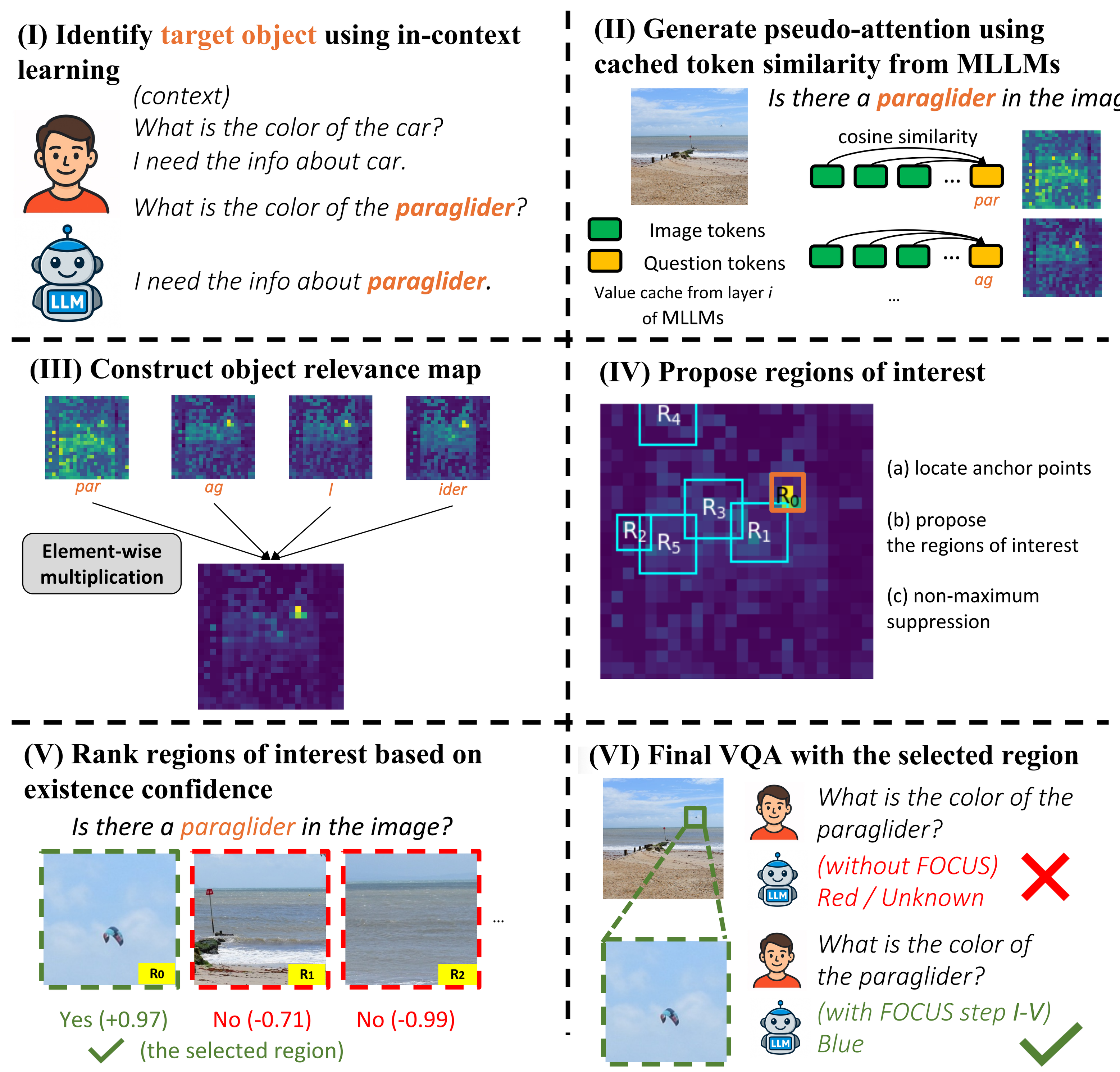
Previously proposed techniques have one of the following key limitations:

1. Task-specific fine-tuning and multiple MLLMs needed (SEAL [1])
2. Exhaustive tree searches due to uninformed search strategies (DC², ZoomEye [2-3])
3. Incompatibility with modern attention mechanisms like FlashAttention (ViCrop [4])



FOCUS for Fine-Grained VQA

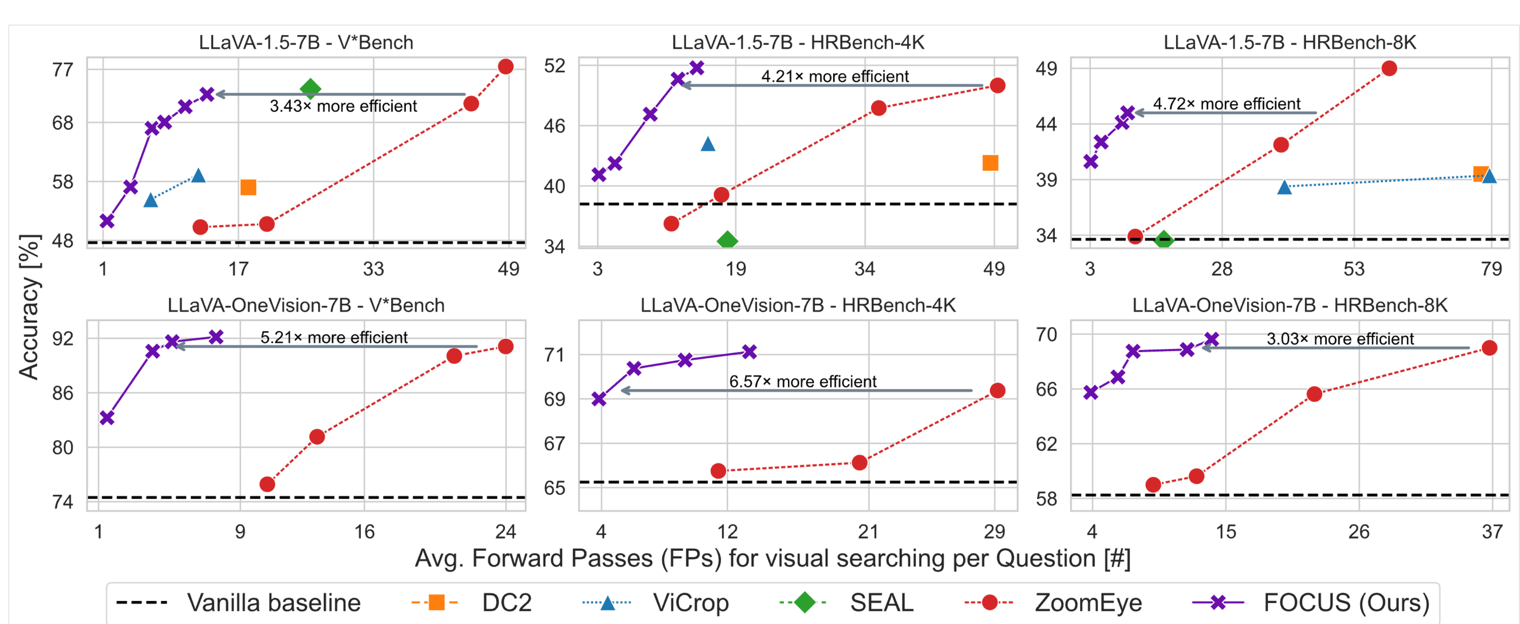
Fine-Grained Visual **O**bject **C**ropping **U**sing Cached Token **S**imilarity



How FOCUS addresses existing key limitations?

1. Training-free localization using MLLMs' KV cache for question-relevant regions
2. Text-guided, object-aware cropping without exhaustive search
3. V-V pseudo-attention replaces Q-K weights for compatibility with efficient attention

Results



Key Message: FOCUS outperforms three baselines and matches ZoomEye on fine-grained VQA with 3 - 6.5x less compute.

Project Page:

Paper:

Key Message: FOCUS achieves SOTA accuracy with Qwen-2.5-VL [5] and generalizes to VQA with larger objects.

Model	V*Bench [%]	HRBench-4K [%]	HRBench-8K [%]
Qwen-2.5-VL	79.06	71.62	68.62
w/ FOCUS	**90.58**	**79.25**	**76.25**

Model	A-OKVQA Acc. [%]	A-OKVQA Δ	GQA Acc. [%]	GQA Δ
LLaVA-1.5	77.99	-	61.97	-
w/ ViCrop	60.66	-17.33	60.98	-0.99
w/ FOCUS	74.76	-3.23	60.34	-1.63
LLaVA-OV	91.44	-	62.01	-
w/ FOCUS	91.00	-0.44	51.02	-10.99

Ablation: Component	Object rel. map	Proposal ranking	V*Bench Acc. [%] ↑	V*Bench Recall [%] ↑	HRBench-4K Acc. [%] ↑
Component	✗	✓	48.68	18.37	36.13
	✓	✗	51.30	38.48	41.13
Pseudo-attn.	K-K (w/o RoPE)		69.10	63.47	45.63
Layers	0 – 14		66.49	76.17	47.38
	0 – 32		71.20	75.56	49.38
Original design choice			**72.77**	**77.49**	**51.75**
Vanilla baseline			47.64	-	36.13
Random guess			35.99	-	25.00

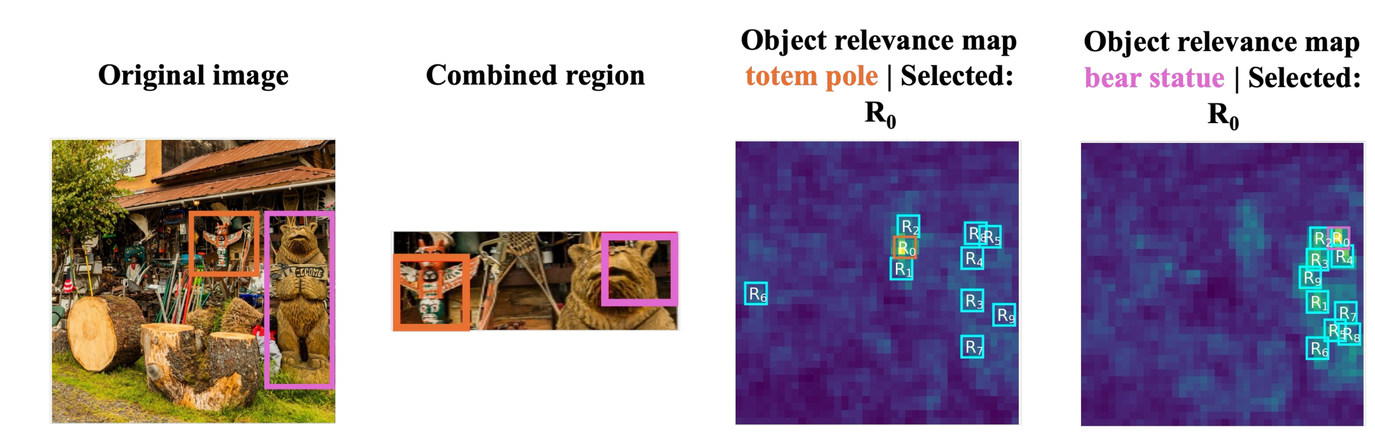
Insights:

- Cached tokens are **object-aware** and encode **spatial cues**
- **Deeper layers** yield stronger localization
- **V-V** pseudo-attention outperforms K-K (w/o RoPE)

Qualitative Examples

(I) **Question:** What is the color of the candles? (A) red (B) yellow (C) gray (D) white
Label: B | **Answer** (LLaVA -1.5): D ✗ | **Answer** (LLaVA -1.5 *w/ FOCUS*): B ✓

(II) **Question:** What is the position of the totem pole in relation to the bear statue?
(A) To the left (B) To the right (C) Behind the bear statue (D) In front
Label: A | **Answer** (LLaVA -OneVision): D ✗ | **Answer** (LLaVA -OneVision *w/ FOCUS*): A ✓



References

[1] "V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs." In: CVPR 2025 by Wu & Xi
[2] "Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal LLMs." In: AAAI 2024 by Wang et al.
[3] "ZoomEye: Enhancing Multimodal LLMs with Human-Like Zooming Capabilities through Tree-Based Image Exploration." In: EMNLP 2025 by Shen et al.
[4] "MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs." In: ICLR 2025 by Zhang et al.
[5] "Qwen2.5-VL Technical Report." In: arXiv by Bai et al.: https://arxiv.org /abs/2502.13923