



FOCUS: Internal MLLM Representations for Efficient Fine-Grained VQA

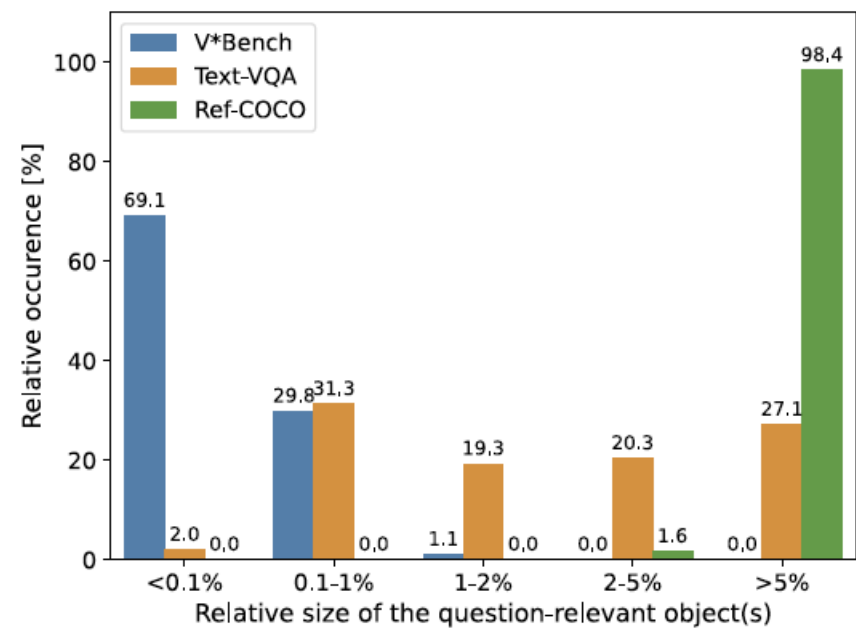
Liangyu Zhong*, Fabio Rosenthal*, Joachim Sicking, Fabian Hümer,
Thorsten Bagdonat, Hanno Gottschalk, Leo Schwinn

* Equal contribution



Motivation

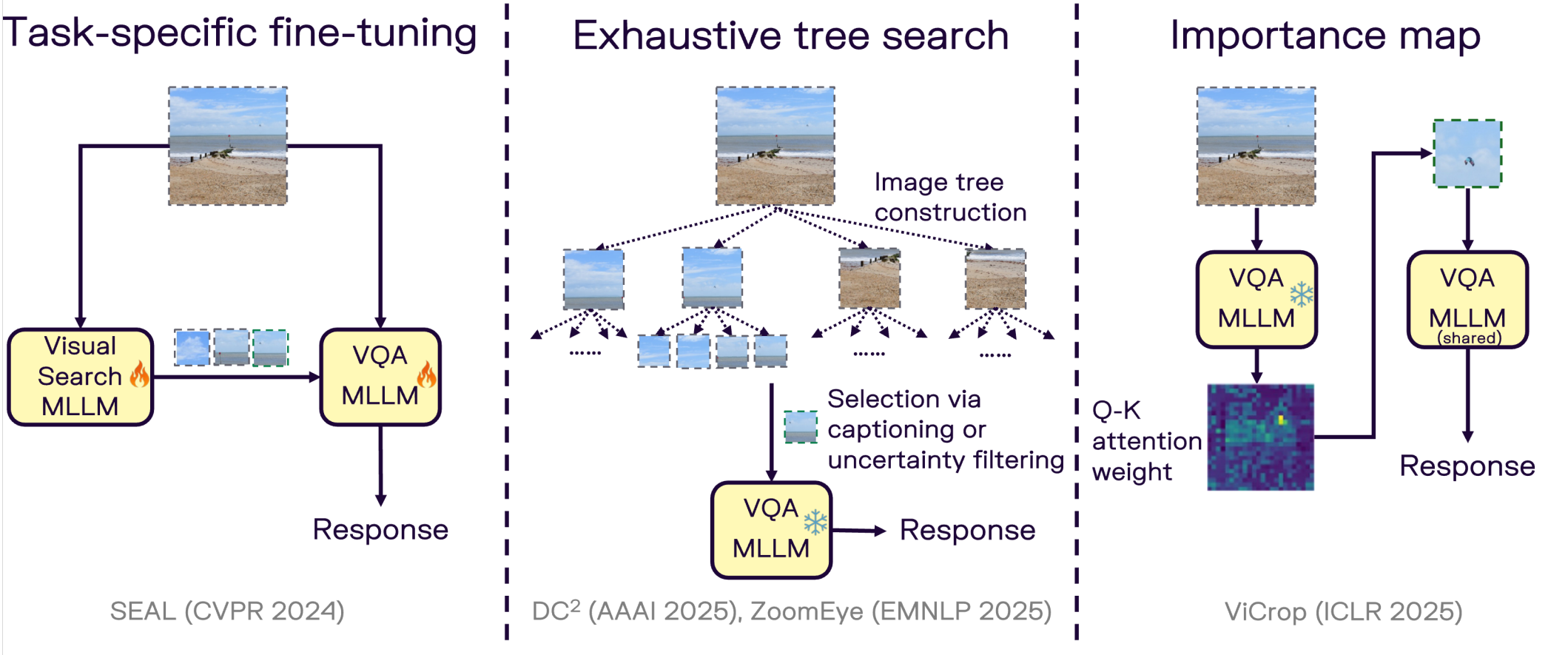
- Most VQA datasets contain images with large objects
- On datasets with small relevant objects, MLLM performance drops significantly
- Providing the relevant image region substantially improves MLLM accuracy



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

Prior methods suffer from different limitations.



Task-specific fine-tuning and multiple MLLMs needed


Uninformed search strategies

Incompatibility with FlashAttention


FOCUS for Fine-Grained VQA

Fine-Grained Visual Object Cropping Using Cached Token Similarity


(I) Identify target object using in-context learning



(context)
What is the color of the car?
I need the info about car.




What is the color of the **paraglider**?



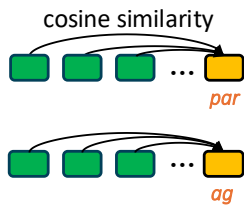
I need the info about **paraglider**.

(II) Generate pseudo-attention using cached token similarity from MLLMs



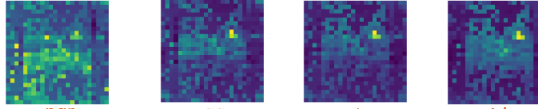
Is there a **paraglider** in the image?

cosine similarity

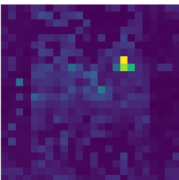


Value cache from layer i of MLLMs

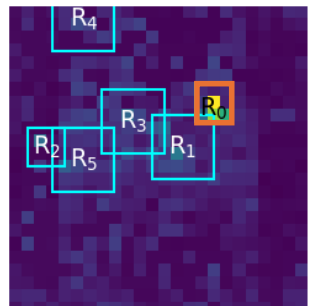
(III) Construct object relevance map



Element-wise multiplication




(IV) Propose regions of interest



- (a) locate anchor points
- (b) propose the regions of interest
- (c) non-maximum suppression

(V) Rank regions of interest based on existence confidence


Is there a **paraglider** in the image?




"Yes" (+0.97) "No" (-0.71) "No" (-0.99)

✓ (the selected region)


(VI) Final VQA with the selected region




What is the color of the paraglider?



(without FOCUS)
Red / Unknown ❌



What is the color of the paraglider?



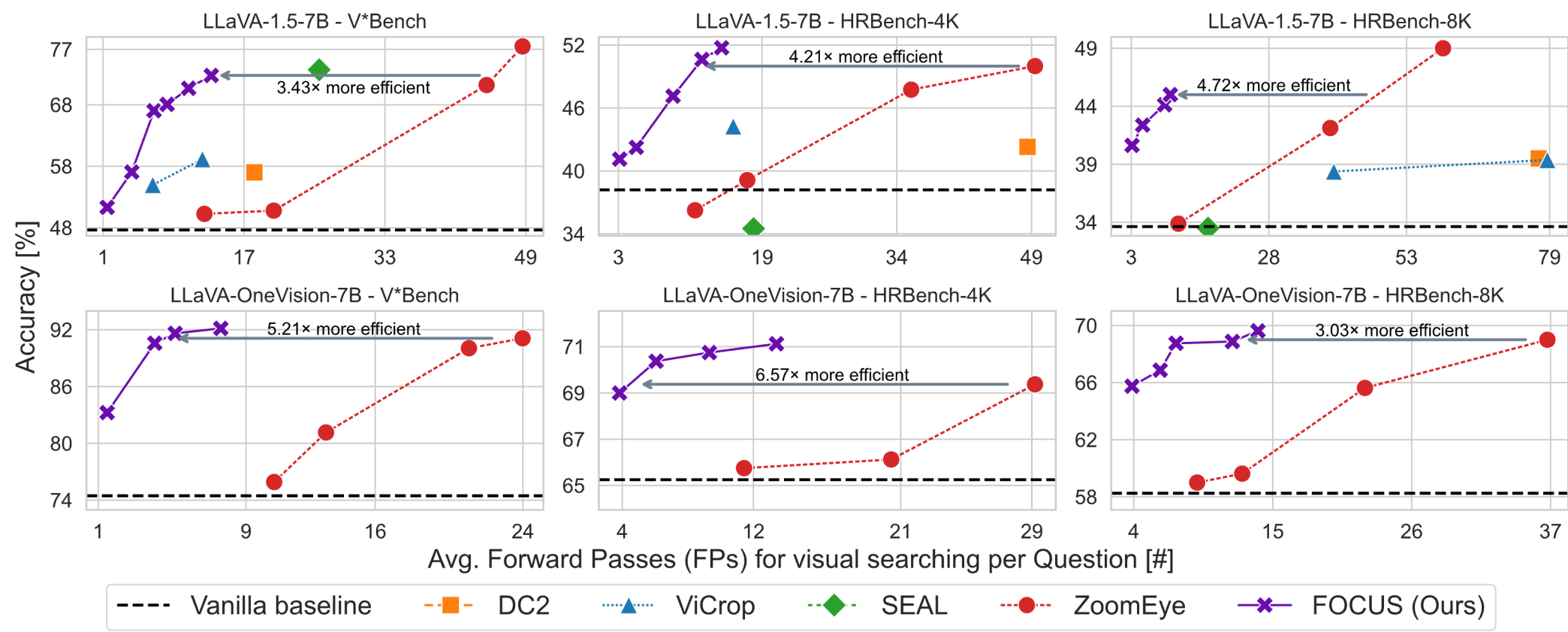
(with FOCUS step I-V)
Blue ✓

How FOCUS addresses existing key limitations?

1. Training-free localization using MLLMs' KV cache
2. No exhaustive tree search due to text-guided, object-aware cropping
3. V-V pseudo-attention replaces Q-K weights for compatibility with efficient attention

Experiments

Evaluation on fine-grained VQA benchmarks



Key message: FOCUS outperforms three baselines and matches ZoomEye on fine-grained VQA with 3 - 6.5x less compute, when using LLaVA-1.5 and LLaVA-OneVision.

Experiments

Additional results

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

(a) FOCUS with Qwen-2.5-VL

Model	A-OKVQA		GQA	
	Acc. [%]	Δ	Acc. [%]	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

(b) FOCUS on VQA with larger objects

Key message: FOCUS achieves SOTA accuracy with Qwen-2.5-VL and generalizes to VQA with larger objects.

Experiments

Ablation studies

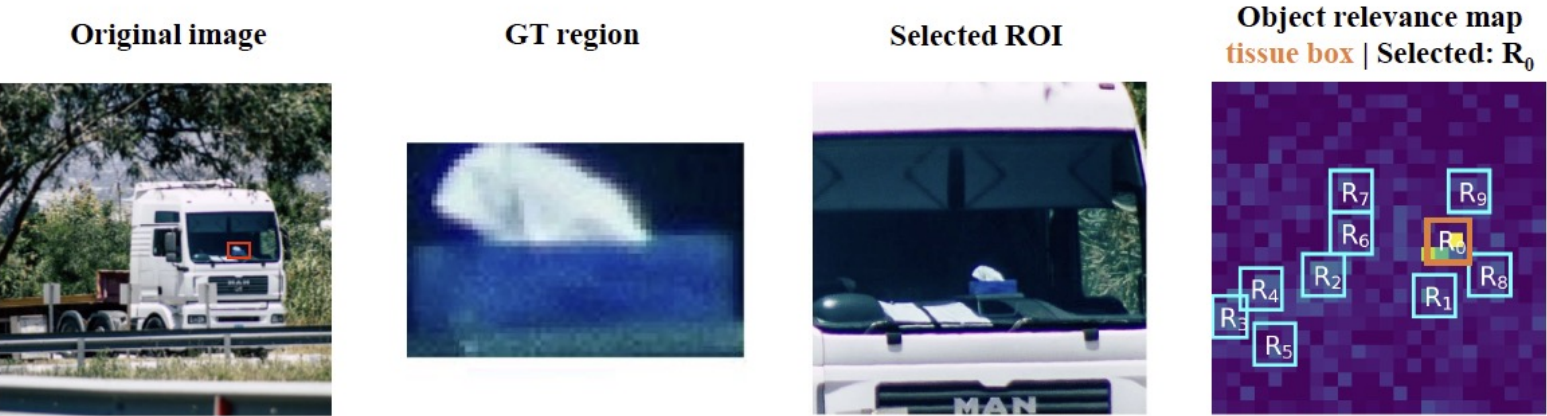
Ablation			V*Bench		HRBench-4K
Component	Object rel. map	Proposal ranking	Acc. [%] ↑	Recall [%] ↑	Acc. [%] ↑
	✗	✓	48.68	18.37	36.13
	✓	✗	51.30	38.48	41.13
Pseudo-attn.	K-K (w/o RoPE [29])		69.10	63.47	45.63
Layers	0 — 14		66.49	76.17	47.38
	0 — 32		71.20	75.56	49.38
Original design choice			72.77	77.49	51.75
Vanilla baseline			47.64	-	36.13
Random guess			35.99	-	25.00

Insights:

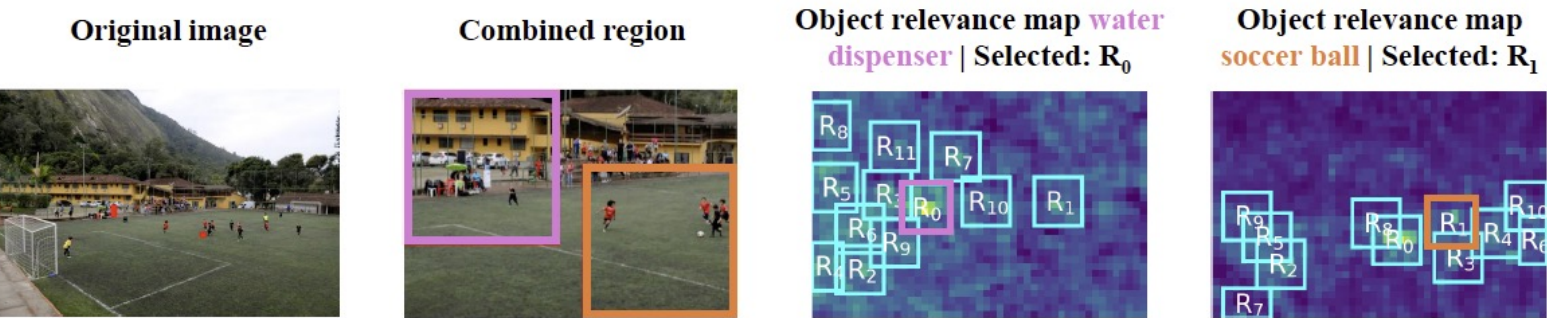
- Cached tokens are **object-aware** and encode **spatial cues**
- **Deeper layers** yield stronger localization
- **V-V pseudo-attention** outperforms K-K (w/o RoPE) pseudo-attention

Qualitative Examples

(I) **Question:** What is the color of the **tissue box**? (A) gray (B) white (C) black (D) blue
Label: D | **Answer** (LLaVA-1.5): B ❌ | **Answer** (LLaVA-1.5 w/ *FOCUS*): D ✔️



(II) **Question:** Is the **soccer ball** on the left or right of the **water dispenser**? (A) left (B) right
Label: B | **Answer** (LLaVA-OneVision): A ❌ | **Answer** (LLaVA-OneVision w/ *FOCUS*): B ✔️



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 ✔️

Original image

GT region

Selected ROI

Object relevance map
candles | Selected: R_3

(II) Question: What is the relative position of the person in the red jacket compared to the large tree? (A) Behind the large tree (B) Right of the large tree (C) In front of the large tree (D) Left of the large tree
Label: B | Answer (LLaVA-1.5): D ❌ | Answer (LLaVA-1.5 w/ *FOCUS*): D ❌

Original image

Combined region

Object relevance map
person in the red jacket | Selected: R_0

Object relevance map
large tree | Selected: R_3

(III) Question: How many **chairs** are there in the image? (A) One (B) Four (C) Two (D) Three
Label: C | Answer (LLaVA-1.5): A ❌ | Answer (LLaVA-1.5 w/ *FOCUS*): C ✔️

Original image

Combined region

Object relevance map
chairs | Selected: R_0 & R_1

(I) Question: What is the **speed limit on the sign** in the image? (A) 20 (B) 40 (C) 60 (D) 30
Label: D | Answer (LLaVA-OneVision): B ❌ | Answer (LLaVA-OneVision w/ *FOCUS*): D ✔️

Original image

GT region

Selected ROI

Object relevance map
speed limit on the sign | Selected: R_6

(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 ✔️

Original image

Combined region

Object relevance map
totem pole | Selected: R_0

Object relevance map
bear statue | Selected: R_0

(III) Question: How many **computers** are on the table? (A) Three (B) Five (C) Two (D) Four
Label: B | Answer (LLaVA-OneVision): C ❌ | Answer (LLaVA-OneVision w/ *FOCUS*): B ✔️

Original image

Combined region

Object relevance map
computers | Selected: R_0 & R_1 & R_2 & R_3 & R_6 & R_8

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.

Project Page:



Paper:

