NEURAL INFORMATION PROCESSING SYSTEMS

FOCUS: Internal MLLM Representations for Efficient Fine-Grained VQA

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

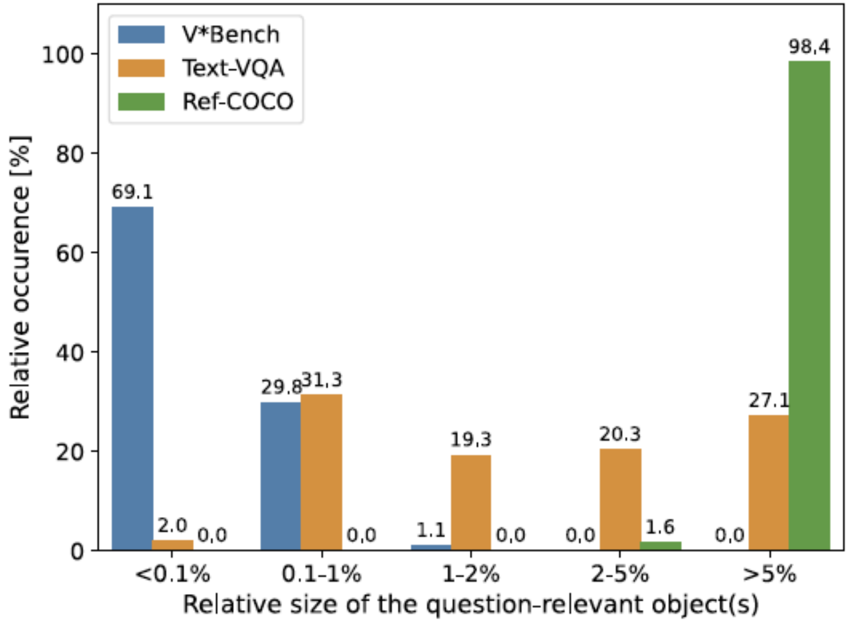
* Equal contribution

Technische Universität Berlin

CARIAD

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

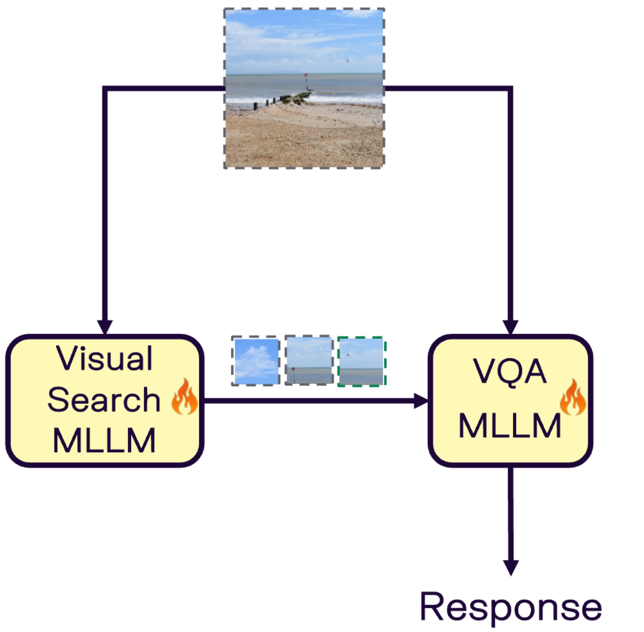


Model	Accuracy on V*Bench [%]
Random guessing	35.99
LLaVA-1.5 (full image)	48.60
LLaVA-1.5 (only GT region)	87.20 (+38.6 pp.)

Recent Visual Cropping Approaches

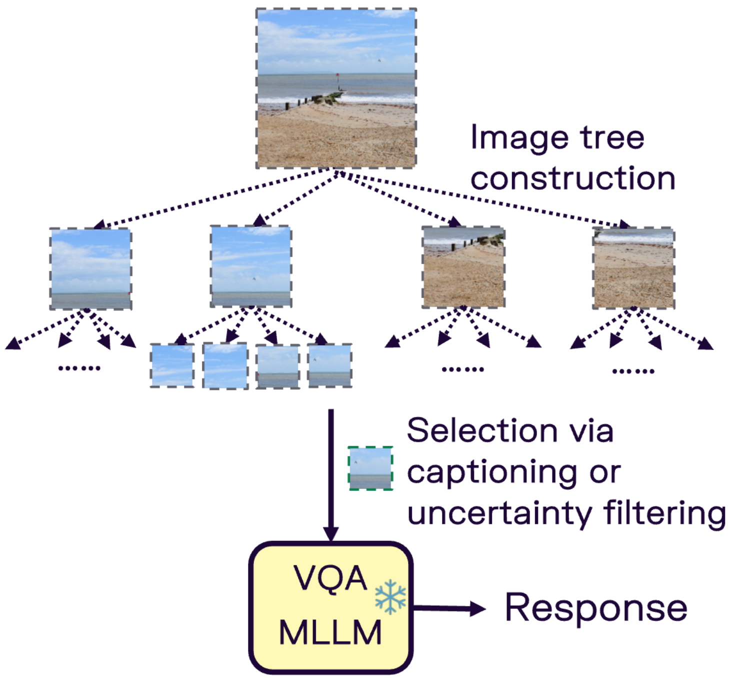
Prior methods suffer from different limitations.

Task-specific fine-tuning



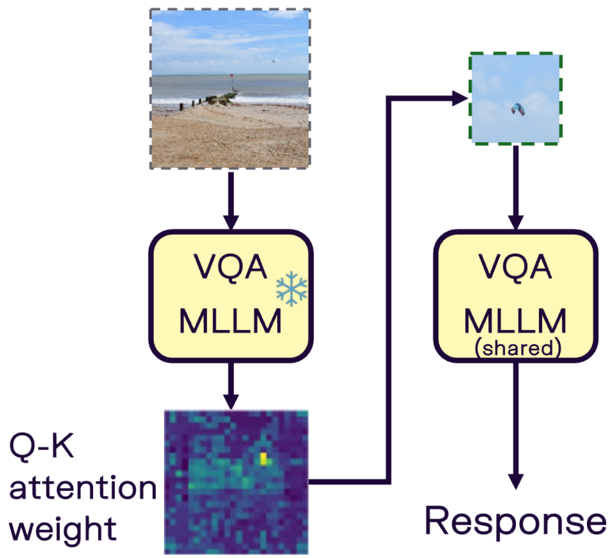
SEAL (CVPR 2024)

Exhaustive tree search



DC² (AAAI 2025), ZoomEye (EMNLP 2025)

Importance map



ViCrop (ICLR 2025)

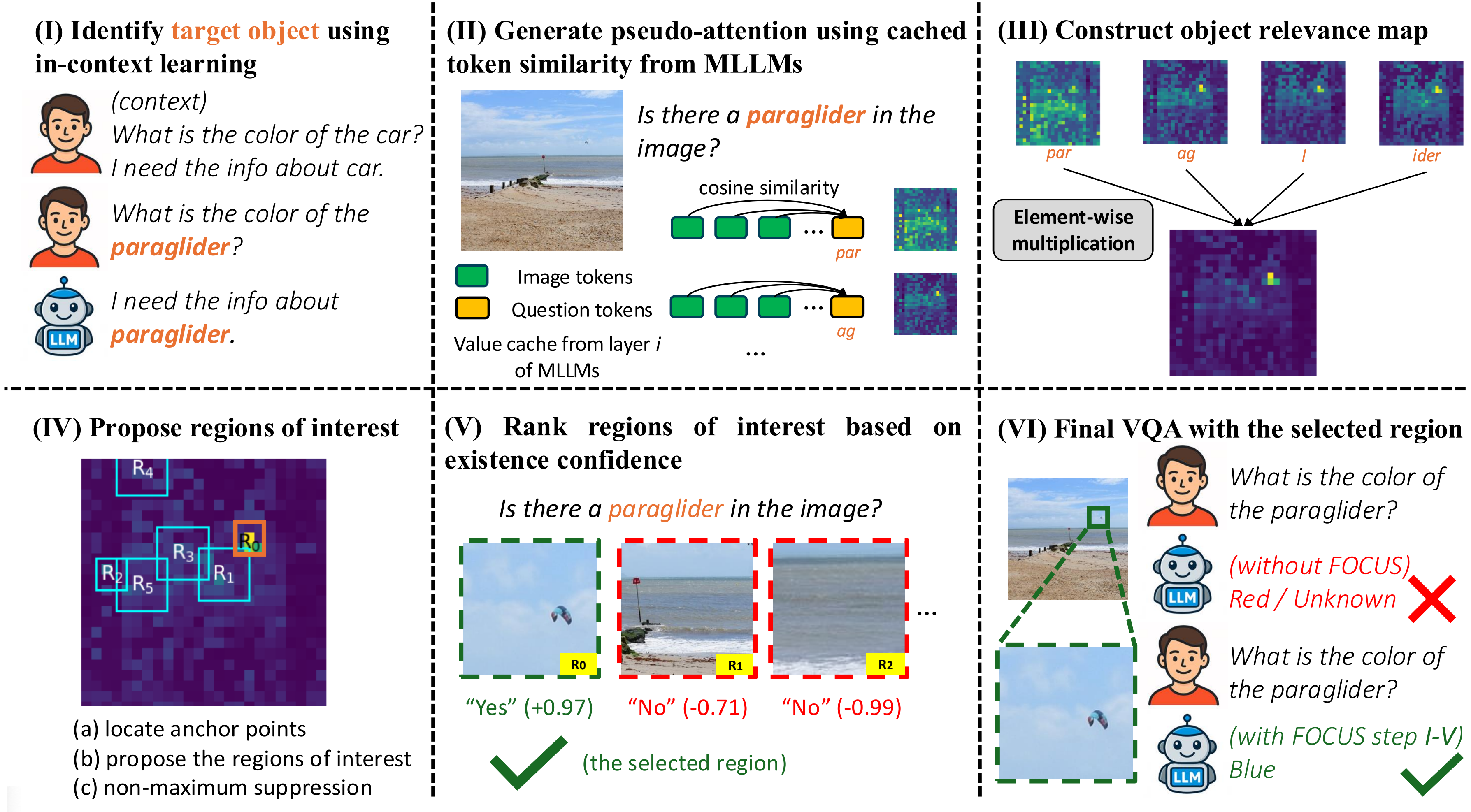
Task-specific fine-tuning and multiple MLLMs needed

Uninformed search strategies

Incompatibility with FlashAttention

FOCUS for Fine-Grained VQA

Fine-Grained Visual Object Cropping Using Cached Token Similarity



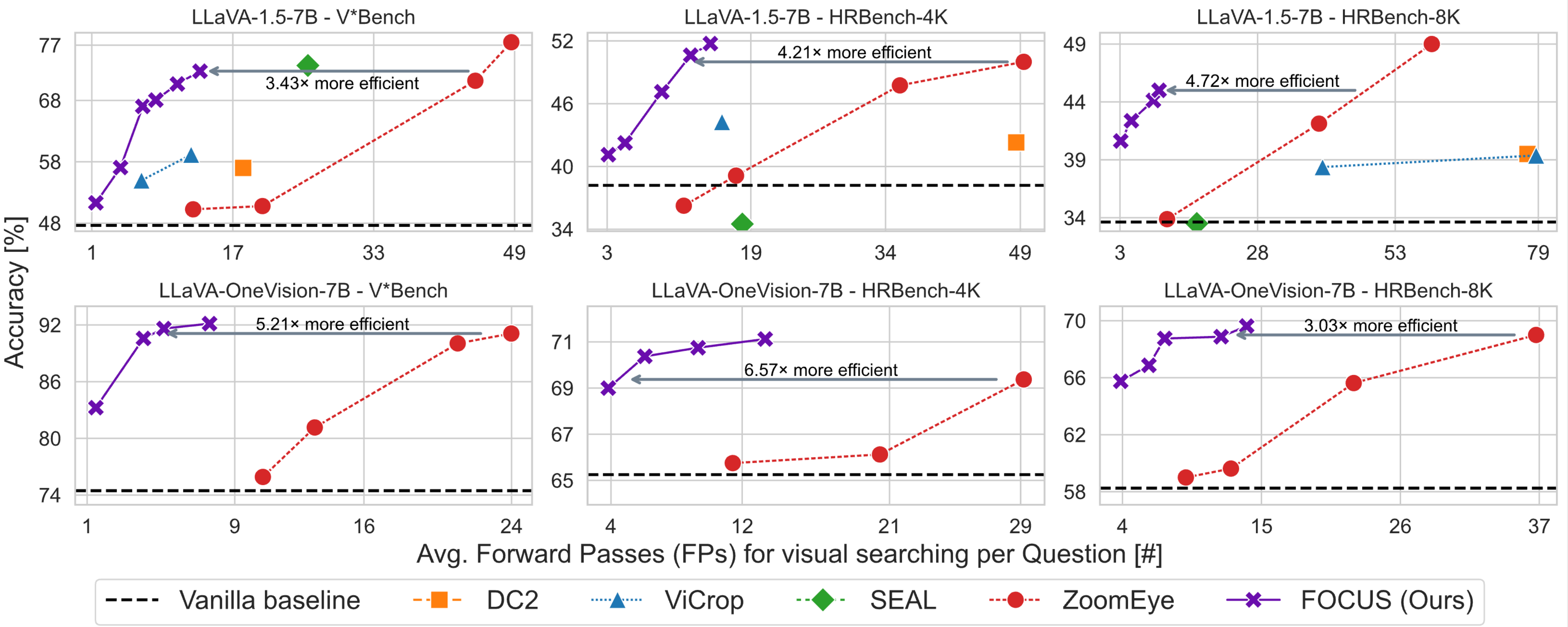
How FOCUS addresses existing key limitations?

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

CARIAD
A VOLKSWAGEN GROUP COMPANY

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 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

(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

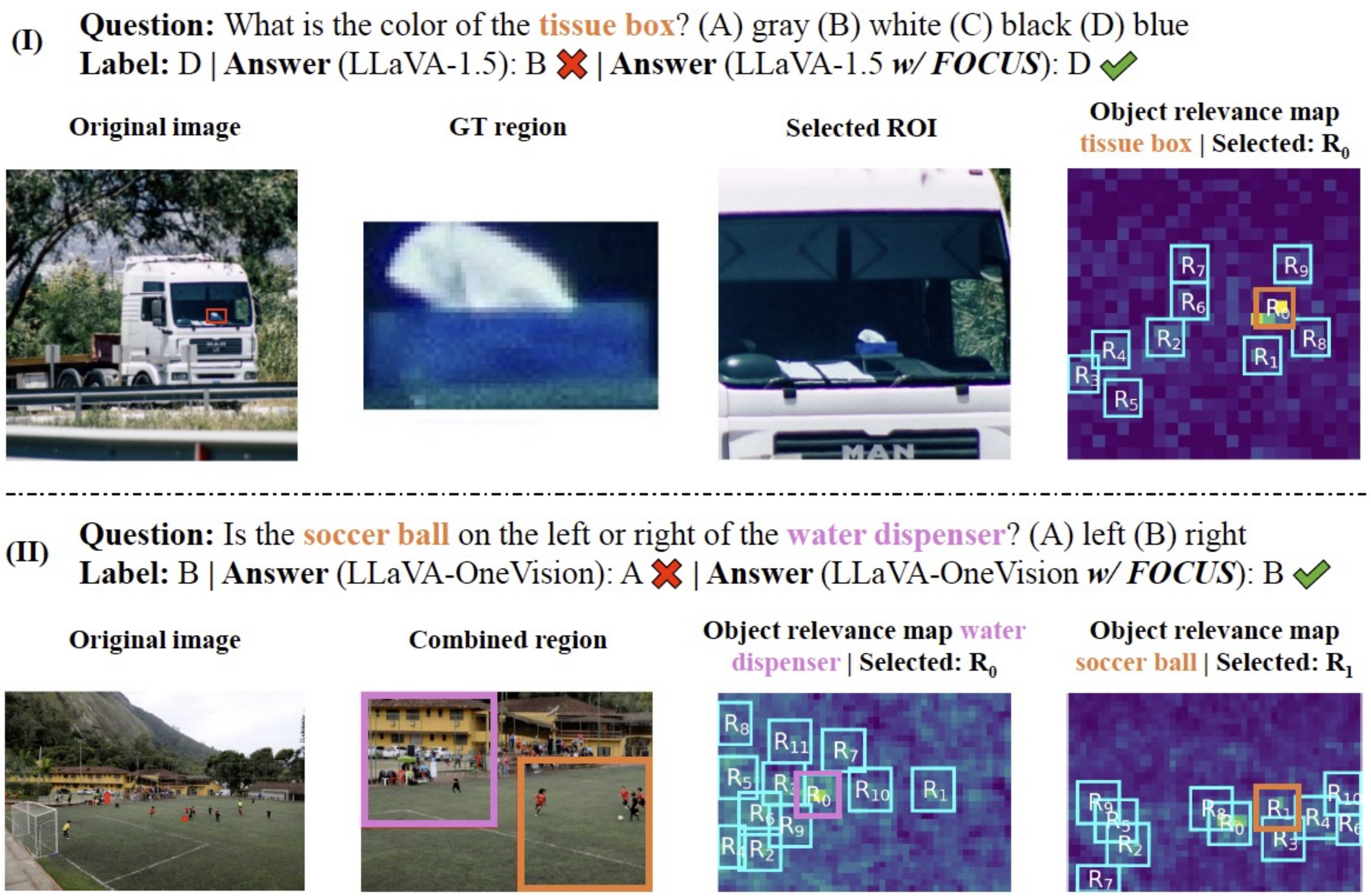
Ablation			V*Bench		HRBench-4K
Component	Object rel. map	Proposal ranking	Acc. [%] ↑	Recall [%] ↑	Acc. [%] ↑
	✗	✓	48.68	18.37	36.13
	✓	✗	51.30	38.48	41.13
Pseudo-attn.	K-K (w/o RoPE [29])		69.10	63.47	45.63
Layers	0 – 14		66.49	76.17	47.38
	0 – 32		71.20	75.56	49.38
Original design choice			**72.77**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

Insights:

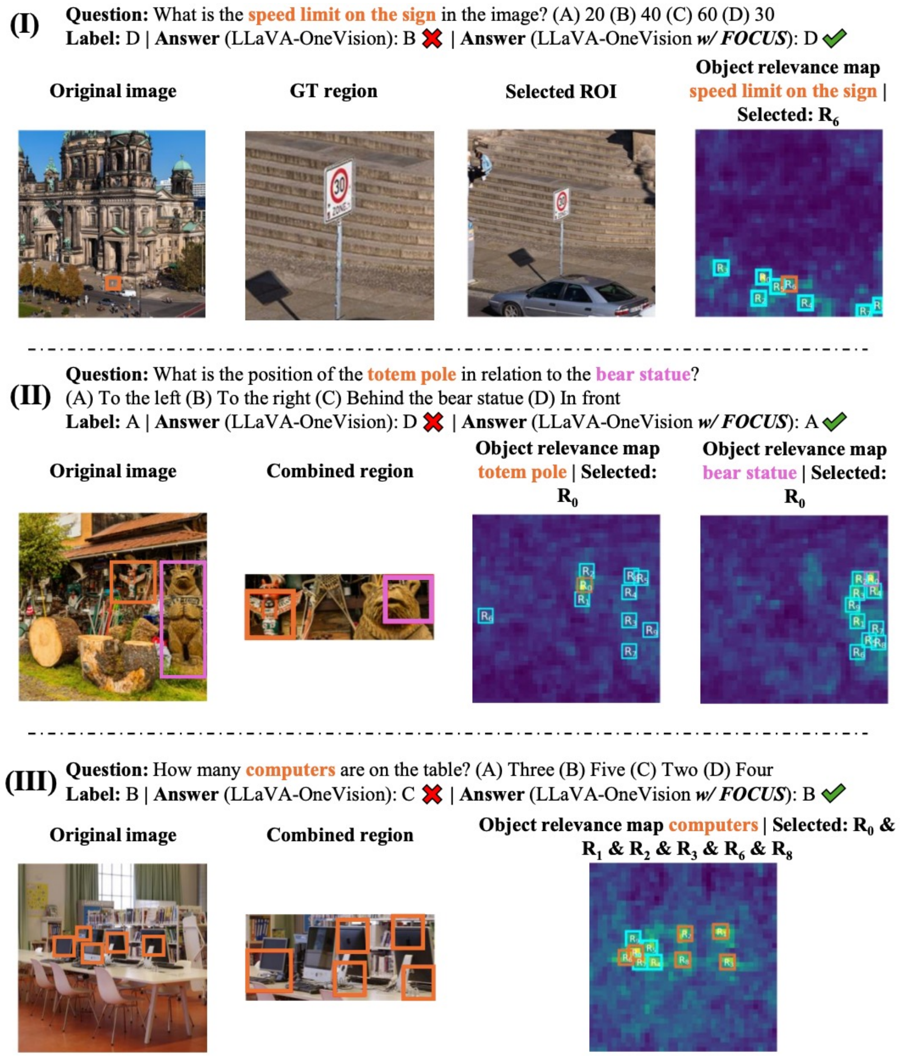
- Cached tokens are **object-aware** and encode **spatial cues**
- **Deeper layers** yield stronger localization
- **V-V pseudo-attention** outperforms K-K (w/o RoPE) pseudo-attention

CARIAD
A VOLKSWAGEN GROUP COMPANY

Qualitative Examples



Qualitative Examples



CARIAD
A VOLKSWAGEN GROUP COMPANY

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: