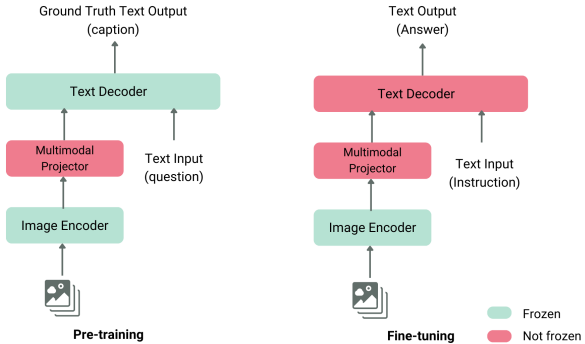
Technische Universität Berlin

Internal MLLM Representations for Efficient Fine-grained Visual Question Answering and Hallucination Detection

Hanno Gottschalk
Institute of Mathematics, TU Berlin and Math+

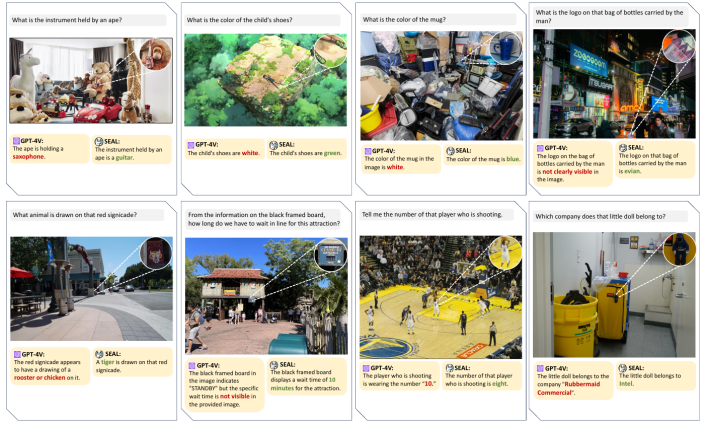
With Liangyu Zhong, et al, Laura Fieback et al. | 10. März 2026

The task of visual quationing and answering (VQA)

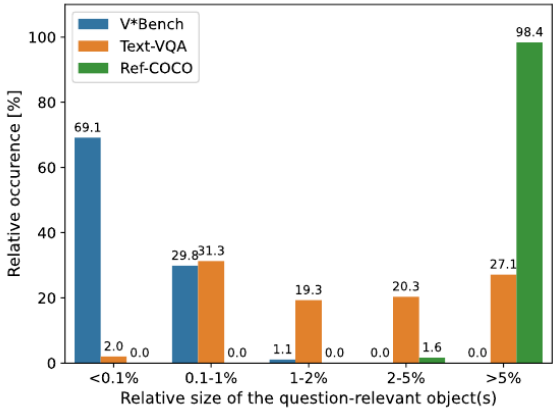


from: Mercity AI

Viaual question and answering



Size of objects in Data Sets



Foreground Bias...

Technische Universität Berlin

Problem description

- Most VLM like InstructBLIP or LLava-1.5 downsize images 224×224 / 336×336 and work on coarse resolutions (taylored for image captioning with foregrond bias).
- This makes it difficult to answer fine grained VQA on small image details.

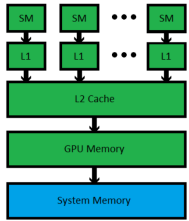
Solution strategies

- LLaVA-OneVision and Gemma-3 process the downsapled image along with cops in a multi-image token cross attention strategy
- Effectiveness on fine grained VQA remains limited as the language branch has problems to select from which image (crop) to decode.
- ZoomEye zooms in a fixed number of crops and probes them with an ordinary VLM, but requires several decodings (inefficient)

List of challenges in fine grained VQA

Method	Training-free	Efficient search algo.	Compatible w/ efficient attention
SEAL [33]	✗	✗	✓
DC^2 [32]	✓	✗	✓
ZoomEye [27]	✓	✗	✓
ViCrop [36]	✓	✓	✗
FOCUS (Ours)	✓	✓	✓

GPU memory organization



Memory	**latency (cycles)**	**location**	**typical size**
Registry	∼1	On-Chip (Kern)	∼256 KB pro SM
L1 Cache / Shared Mem	28 – 60	On-Chip (SM)	128 – 256 KB
L2 Cache	150 – 250	On-Chip (Shared)	4 – 96 MB
VRAM (GDDR6/HBM)	400 – 800+	Off-Chip (Grafikkarte)	8 – 24 GB
System-RAM	1.000 – 2.000+	Mainboard (PCIe)	16 – 128 GB

FlashAttention

$$O = \texttt{softmax}(QK^\top)V$$

Problem

- In efficient implementations of VLM, FlashAttention is used and the desired attention matrix $A = \texttt{softmax}(QK^\top)$ is nowhere stored.
- FlashAttention uses tiling, i.e. single blocks $K^{(i)}$, $Q^{(i)}$ and $V^{(i)}$ are processed separately in the GPU
- This is combined with the `OnlineSoftmax` algorithm.

The problem with global softmax

- Standard softmax for input sequence $x = (z_1, \ldots, z_n)$:

$$\text{softmax}(z)_i = \frac{e^{z_i - m}}{\sum_{j=1}^{n} e^{z_j - m}} \quad \text{with} \quad m = \max_j x_j$$

Problem with tiling:

- We only see one block $x^{(i)}$ at a time.
- The global maximum m and the complete sum are only known at the end.

Online softmax recursion

- For reach new block we $z^{(i)}$ update statistics in fast SRAM memory:

$$m_i = \max(m_{i-1}, \max(z^{(i)}))$$

- Update the old exponential to the new maximum value

$$d_i = d_{i-1} \cdot e^{m_{i-1} - m_i} + \sum_{j=1}^{i} e^{z^{(j)} - m_j}$$

- Update the output

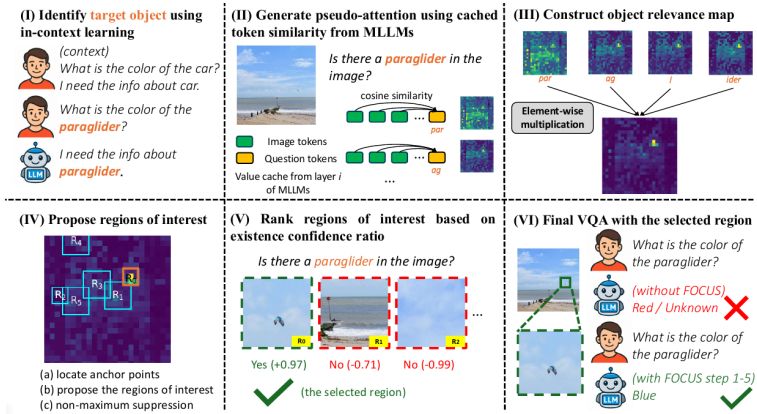
$$O_i = O_{i-1} \cdot \frac{d_{i-1} \cdot e^{m_{i-1} - m_i}}{d_i} + \frac{e^{z^{(i)} - m_i}}{d_i} \cdot V^{(i)}$$

V-V pseudoattention

- Fast access to V is possible as it is stored for reuse in the VRAM.
- For VLM, the same is true for $\hat{V}$, the value matrix for visual tokens.
- Instead of the standard attention matrix, we may thus choose the pseudo-attention matrix at layer l

$$A_{i,j}^{(l)} = \cos(V_i^{(l)}, \hat{V}_j^{(l)})$$

FOCUS general strategy



Computation of relevance maps

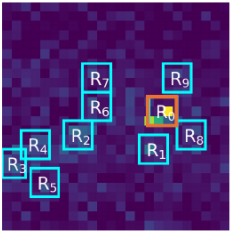
- Aggregation over layers for head r

$$A_r = \frac{1}{2} \sum_{l=1}^{L} (A_r^{(l)} + \mathbb{1})$$

- Removal of only partially important tokens

$$A = A_0 \odot \ldots \odot A_s \quad \text{(Schur product)}$$

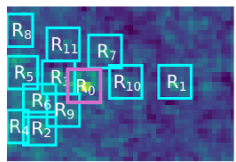
Example of a relavence map

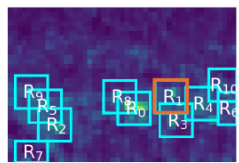


- Q: What is the color of the tissue box?

Two selection strategies

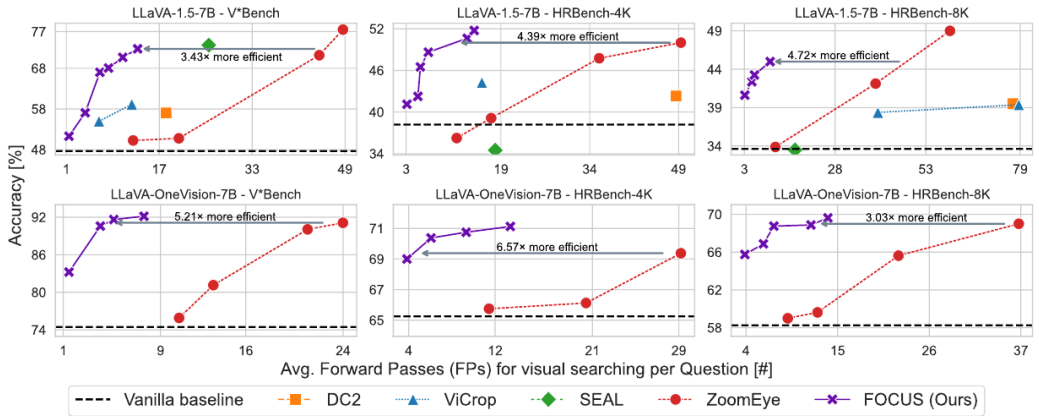
- Q: Is the soccer ball left or right of the water dispenser?





- Questions on properties of one object → only retain most relevant patch (according to VLM confidence in VQA)
- Questions on relative position of objects → retain multiple patches

Comparison with the SOTA in fine grained VQA



Examples

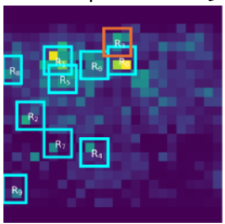
(I) **Question:** What is the color of the **candles**? (A) red (B) yellow (C) gray (D) white
Label: B | **Answer** (LLaVA-1.5): D ✗ | **Answer** (LLaVA-1.5 ***w/ FOCUS***): B ✓

Original image

GT region

Selected ROI

Object relevance map
candles | Selected: R_3

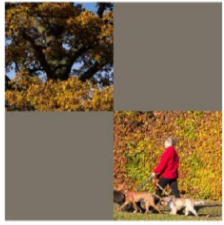


Examples

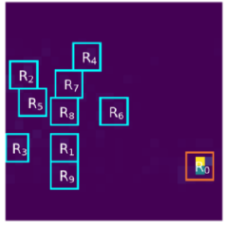
(II) **Question:** What is the relative position of the person in the red jacket compared to the large tree? (A) Behind the large tree (B) Right of the large tree (C) In front of the large tree (D) Left of the large tree
Label: B | **Answer** (LLaVA-1.5): D ✗ | **Answer** (LLaVA-1.5 ***w/ FOCUS***):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

Original image

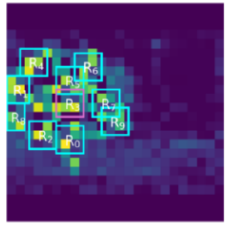
Combined region



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



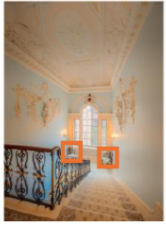
Object relevance map
large tree | Selected: R_3



Examples

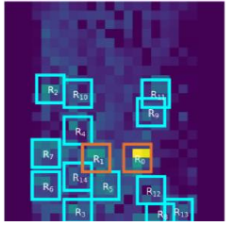
(III) **Question:** How many **chairs** are there in the image? (A) One (B) Four (C) Two (D) Three
Label: C | **Answer** (LLaVA-1.5): A ✗ | **Answer** (LLaVA-1.5 ***w/ FOCUS***): C ✓

Original image



Combined region

Object relevance map chairs | Selected: R_0 & R_1

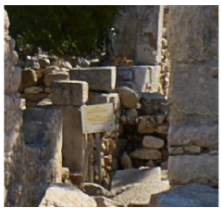


Examples

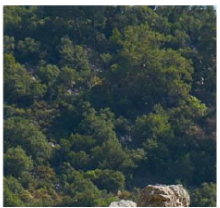
(IV) **Question:** What is the color of the **sign** in the image? (A) Green and white (B) Yellow and white (C) Yellow and green (D) White and red
Label: B | **Answer** (LLaVA-1.5): A ✗ | **Answer** (LLaVA-1.5 ***w/ FOCUS***): C ✗

Original image

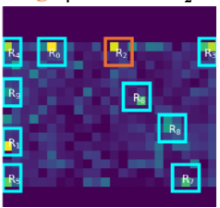
GT region



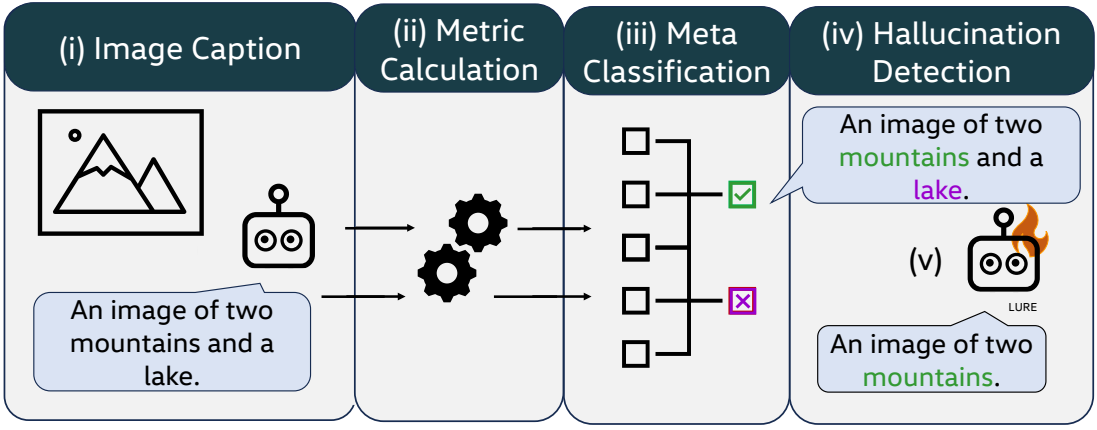
Selected ROI



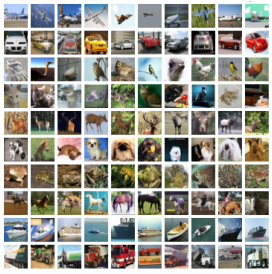
Object relevance map
sign | Selected: R_2



Hallucination and Hallucination Detection



How to Measure Hallucination?



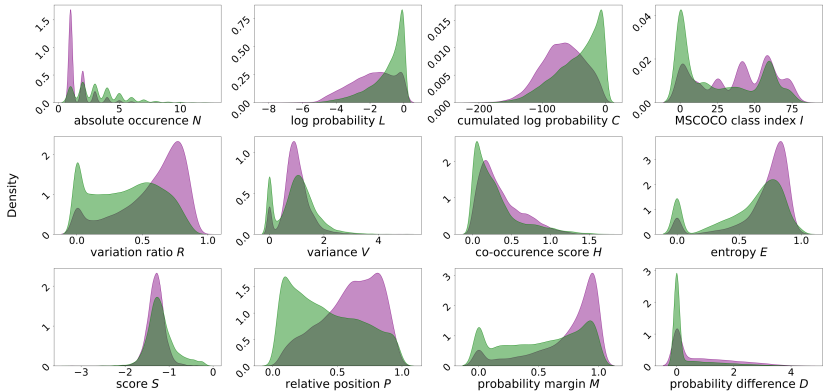
- CHAIR metirc based on MSCOCO

$$\text{CHAIR} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}$$

Which Metics Influence Hallucunation?

- MSCoCo class index
- Co-occurrence score
- Relative position
- Number of occurences
- Mean absolute attention
- Variation ratio
- Clip score
- log-Probability
- cumulated log-probability
- Sequence score
- Variance
- Entropy
- probability margin
- Probability difference

Separation Performance of Single Metrics



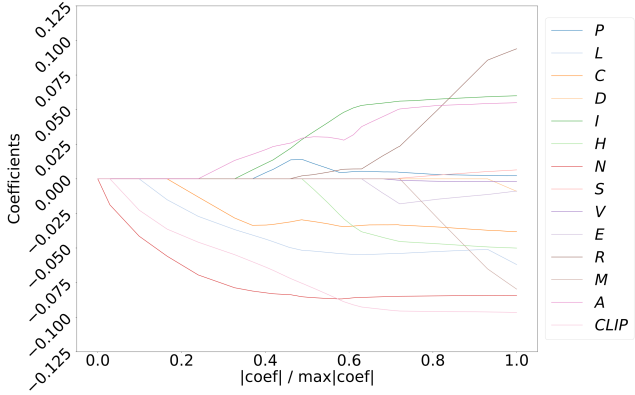
Meta Classification

- The metrics are used to predict the labels y: "hallucinated" $= 1$ and "not hallucinated" $= 2$.
- Logistic regression with L1 penalization $\varphi(z) = \frac{e^z}{1+e^z}$

$$\mathcal{L} = -\frac{1}{n}\sum_{j=1}^{n} y_j \log \varphi(\mathbf{M}_j^\top \theta) + (1 - y_j) \log(1 - \varphi(\mathbf{M}_j^\top \theta)) + \lambda \|\theta\|_1.$$

- Or use GBM for classification

Metrics Importance



Performance

		ACC (in %) ↑		*AUROC* (in %) ↑		*AUPRC* (in %) ↑	
		LR	GB	LR	GB	LR	GB
InstructBLIP	*L*	$89.46^{(\pm 1.4e-3)}$	$89.46^{(\pm 1.3e-3)}$	$73.51^{(\pm 8.7e-3)}$	$73.16^{(\pm 9.0e-3)}$	$27.07^{(\pm 2.1e-2)}$	$25.6^{(\pm 2.5e-2)}$
	E	$89.49^{(\pm 1.6e-3)}$	$89.48^{(\pm 1.6e-3)}$	$65.49^{(\pm 1.3e-2)}$	$66.23^{(\pm 1.5e-2)}$	$15.38^{(\pm 5.7e-3)}$	$17.68^{(\pm 7.0e-3)}$
	Ours	$91.33^{(\pm 1.8e-3)}$	**91.48**$^{(\pm 1.5e-3)}$	$89.94^{(\pm 9.1e-3)}$	**90.39**$^{(\pm 6.8e-3)}$	$56.13^{(\pm 1.2e-2)}$	**57.50**$^{(\pm 1.0e-2)}$
mPLUG-Owl	*L*	$72.42^{(\pm 4.3e-3)}$	$72.48^{(\pm 4.6e-3)}$	$71.75^{(\pm 9.4e-3)}$	$71.86^{(\pm 9.3e-3)}$	$51.21^{(\pm 1.2e-2)}$	$50.65^{(\pm 1.1e-2)}$
	E	$70.06^{(\pm 4.9e-3)}$	$70.77^{(\pm 2.9e-3)}$	$66.01^{(\pm 6.3e-3)}$	$68.33^{(\pm 6.1e-3)}$	$40.09^{(\pm 8.2e-3)}$	$45.54^{(\pm 1.2e-2)}$
	Ours	$82.87^{(\pm 2.7e-3)}$	**83.56**$^{(\pm 3.1e-3)}$	$88.55^{(\pm 3.7e-3)}$	**89.87**$^{(\pm 2.3e-3)}$	$76.11^{(\pm 5.8e-3)}$	**78.61**$^{(\pm 7.9e-3)}$
	Ours^{clip}	$85.05^{(\pm 2.8e-3)}$	**85.84**$^{(\pm 2.9e-3)}$	$91.29^{(\pm 3.2e-3)}$	**92.19**$^{(\pm 2.1e-3)}$	$80.98^{(\pm 4.3e-3)}$	**82.69**$^{(\pm 6.1e-3)}$
MiniGPT-4	*L*	$86.91^{(\pm 3.6e-3)}$	$86.85^{(\pm 3.9e-3)}$	$67.26^{(\pm 2.1e-2)}$	$67.01^{(\pm 2.1e-2)}$	$26.25^{(\pm 1.7e-2)}$	$25.41^{(\pm 1.2e-2)}$
	E	$86.84^{(\pm 3.6e-3)}$	$86.82^{(\pm 3.6e-3)}$	$60.78^{(\pm 1.8e-2)}$	$63.19^{(\pm 1.2e-2)}$	$15.77^{(\pm 6.7e-3)}$	$18.98^{(\pm 1.3e-2)}$
	Ours	$88.92^{(\pm 3.5e-3)}$	**89.27**$^{(\pm 4.9e-3)}$	$88.16^{(\pm 1.5e-2)}$	**89.74**$^{(\pm 1.3e-2)}$	$54.90^{(\pm 6.5e-2)}$	**57.27**$^{(\pm 5.7e-2)}$
LLaVa	*L*	$81.57^{(\pm 1.4e-3)}$	$81.49^{(\pm 1.6e-3)}$	$70.53^{(\pm 8.7e-3)}$	$70.73^{(\pm 6.6e-3)}$	$37.53^{(\pm 2.0e-2)}$	$36.59^{(\pm 1.7e-2)}$
	E	$81.28^{(\pm 2.8e-3)}$	$81.26^{(\pm 2.9e-3)}$	$62.73^{(\pm 9.0e-3)}$	$64.63^{(\pm 7.7e-3)}$	$23.85^{(\pm 6.3e-3)}$	$27.52^{(\pm 4.6e-3)}$
	Ours	$87.25^{(\pm 2.0e-3)}$	**87.78**$^{(\pm 3.0e-3)}$	$90.05^{(\pm 4.0e-3)}$	**91.01**$^{(\pm 4.3e-3)}$	$70.15^{(\pm 1.0e-2)}$	**72.58**$^{(\pm 1.3e-2)}$
	Ours^{clip}	$87.93^{(\pm 1.2e-3)}$	**88.36**$^{(\pm 1.6e-3)}$	$91.73^{(\pm 2.4e-3)}$	**92.50**$^{(\pm 2.5e-3)}$	$73.00^{(\pm 7.8e-3)}$	**75.08**$^{(\pm 8.5e-3)}$

Conclusion and Outlook

- Internal representations based on V-V (flash) attention can be used to efficiently select relevant regions based on similarity with object tokens in common embedding space
- The method is equally performant but considerable more efficient that comparable SOTA approaches like ZoomEye
- Further applications of internal applications is hallucination detection in VLM

Zhong, L., Rosenthal, F., Sicking, J., Hüger, F., Bagdonat, T., Gottschalk, H., and Schwinn, L. (2025). FOCUS: Internal MLLM Representations for Efficient Fine-Grained Visual Question Answering. NeurIPS 2025, arXiv preprint arXiv:2506.21710.
Fieback, Laura, Jakob Spiegelberg, and Hanno Gottschalk. "Metatoken: Detecting hallucination in image descriptions by meta classification."VISAPP 2025 arXiv:2405.19186 (2024).