

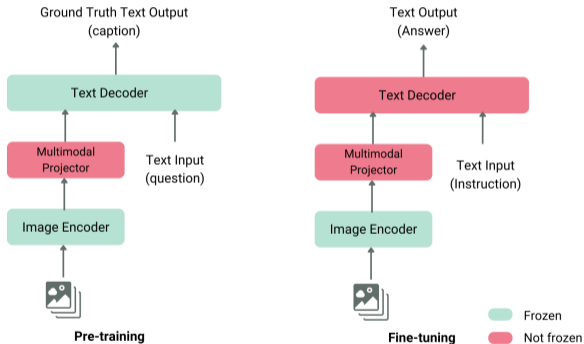
Internal MLLM Representations for Efficient Fine-grained Visual Question Answering and Hallucination Detection

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)

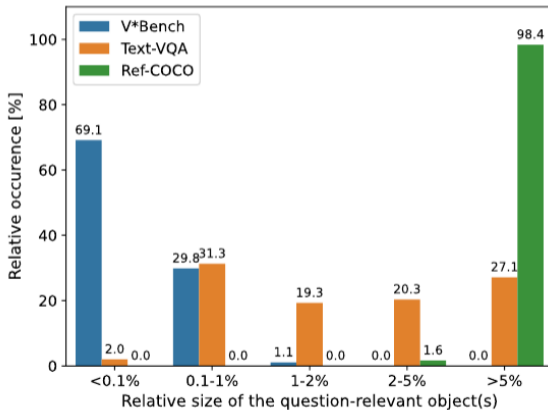


from: Mercy AI

Visual question and answering

<p>What is the instrument held by an ape?</p>  <p>GPT-4V: The ape is holding a saxophone.</p> <p>SEAL: The instrument held by an ape is a guitar.</p>	<p>What is the color of the child's shoes?</p>  <p>GPT-4V: The child's shoes are white.</p> <p>SEAL: The child's shoes are green.</p>	<p>What is the color of the mug?</p>  <p>GPT-4V: The color of the mug in the image is white.</p> <p>SEAL: The color of the mug is blue.</p>	<p>What is the logo on that bag of bottles carried by the man?</p>  <p>GPT-4V: The logo on the bag of bottles carried by the man is not clearly visible in the image.</p> <p>SEAL: The logo on that bag of bottles carried by the man is Intel.</p>
<p>What animal is drawn on that red signcade?</p>  <p>GPT-4V: The red signcade appears to have a drawing of a rooster or chicken on it.</p> <p>SEAL: A liger is drawn on that red signcade.</p>	<p>From the information on the black framed board, how long do we have to wait in line for this attraction?</p>  <p>GPT-4V: The black framed board in the image indicates "STANDBY" but the specific wait time is not visible in the provided image.</p> <p>SEAL: The black framed board displays a wait time of 10 minutes for the attraction.</p>	<p>Tell me the number of that player who is shooting.</p>  <p>GPT-4V: The player who is shooting is wearing the number "10."</p> <p>SEAL: The number of that player who is shooting is eight.</p>	<p>Which company does that little doll belong to?</p>  <p>GPT-4V: The little doll belongs to the company "Rubbermaid Commercial".</p> <p>SEAL: The little doll belongs to Intel.</p>

Size of objects in Data Sets



Foreground Bias...

Problem description

- ▶ Most VLM like InstructBLIP or LLaVA-1.5 downsize images $224 \times 224 / 336 \times 336$ and work on coarse resolutions (tailored for image captioning with foreground bias).
- ▶ This makes it difficult to answer fine grained VQA on small image details.

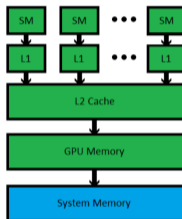
Solution strategies

- ▶ LLaVA-OneVision and Gemma-3 process the downscaled image along with crops 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

$$O = \text{softmax}(QK^T)V$$

Problem

- ▶ In efficient implementations of VLM, FlashAttention is used and the desired attention matrix $A = \text{softmax}(QK^T)$ 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 $x = (z_1, \dots, z_n)$:

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

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 each 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

(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

Image tokens (green) Question tokens (yellow)

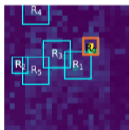
Value cache from layer i of MLLMs

(III) Construct object relevance map

Element-wise multiplication

par ag l ider

(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 ratio

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 1-5)
 Blue ✅

Computation of relevance maps

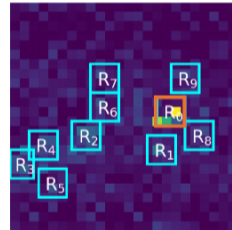
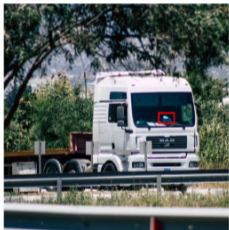
- ▶ Aggregation over layers for head r

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

- ▶ Removal of only partially important tokens

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

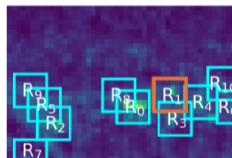
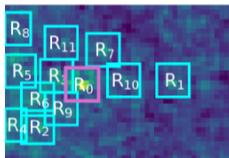
Example of a relevance map



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

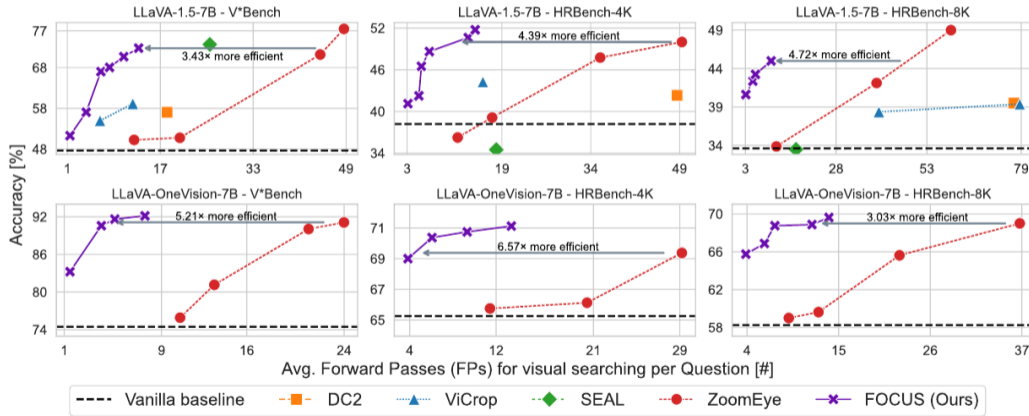
Two selection strategies

- ▶ 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



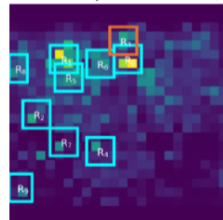
GT region



Selected ROI



Object relevance map
candles | Selected: R_3



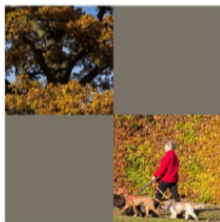
Examples

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 ❌

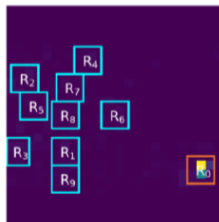
Original image



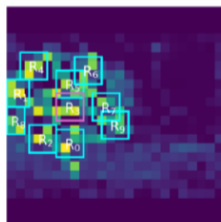
Combined region



Object relevance map
 person in the red jacket | Selected: R₀



Object relevance map
 large tree | Selected: R₃



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 ✅

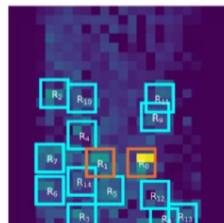
Original image



Combined region



Object relevance map **chairs** | Selected: R_0 & R_1



Examples

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

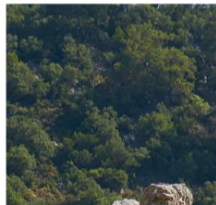
Original image



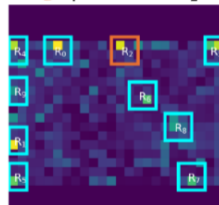
GT region



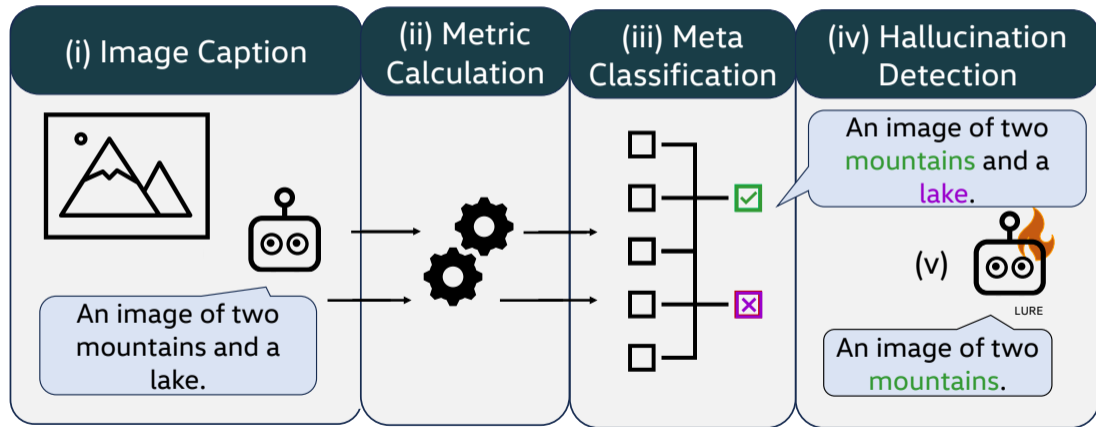
Selected ROI



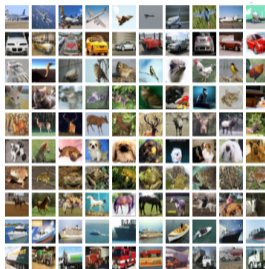
Object relevance map
sign | Selected: R_2



Hallucination and Hallucination Detection



How to Measure Hallucination?



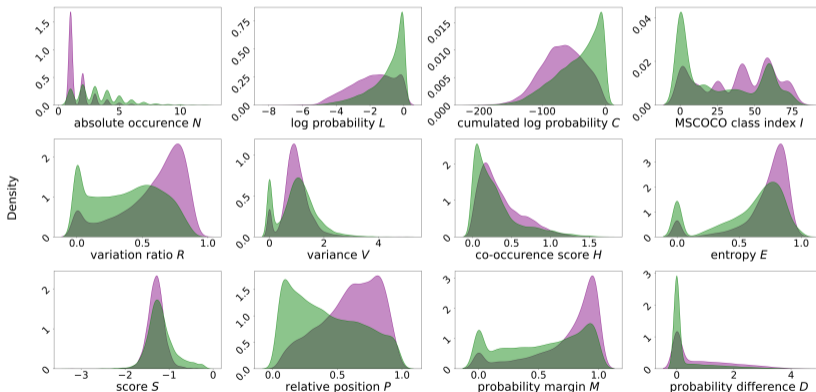
- ▶ CHAIR metric based on MSCOCO

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

Which Metrics Influence Hallucination?

- ▶ MSCoCo class index
- ▶ Co-occurrence score
- ▶ Relative position
- ▶ Number of occurrences
- ▶ 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



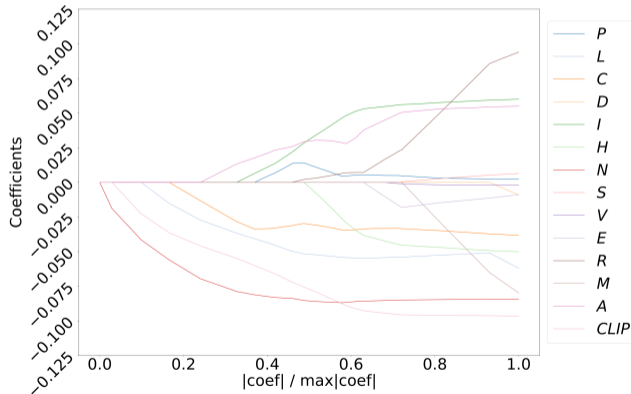
Meta Classification

- ▶ 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	<i>L</i>	89.46(±1.4e-3)	89.46(±1.3e-3)	73.51(±8.7e-3)	73.16(±9.0e-3)	27.07(±2.1e-2)	25.6(±2.5e-2)
	<i>E</i>	89.49(±1.6e-3)	89.48(±1.6e-3)	65.49(±1.3e-2)	66.23(±1.5e-2)	15.38(±5.7e-3)	17.68(±7.0e-3)
	Ours	91.33(±1.8e-3)	91.48 (±1.5e-3)	89.94(±9.1e-3)	90.39 (±6.8e-3)	56.13(±1.2e-2)	57.50 (±1.0e-2)
mPLUG-Owl	<i>L</i>	72.42(±4.3e-3)	72.48(±4.6e-3)	71.75(±9.4e-3)	71.86(±9.3e-3)	51.21(±1.2e-2)	50.65(±1.1e-2)
	<i>E</i>	70.06(±4.9e-3)	70.77(±2.9e-3)	66.01(±6.3e-3)	68.33(±6.1e-3)	40.09(±8.2e-3)	45.54(±1.2e-2)
	Ours	82.87(±2.7e-3)	83.56 (±3.1e-3)	88.55(±3.7e-3)	89.87 (±2.3e-3)	76.11(±5.8e-3)	78.61 (±7.9e-3)
	Ours ^{clip}	85.05(±2.8e-3)	85.84 (±2.9e-3)	91.29(±3.2e-3)	92.19 (±2.1e-3)	80.98(±4.3e-3)	82.69 (±6.1e-3)
MiniGPT-4	<i>L</i>	86.91(±3.6e-3)	86.85(±3.9e-3)	67.26(±2.1e-2)	67.01(±2.1e-2)	26.25(±1.7e-2)	25.41(±1.2e-2)
	<i>E</i>	86.84(±3.6e-3)	86.82(±3.6e-3)	60.78(±1.8e-2)	63.19(±1.2e-2)	15.77(±6.7e-3)	18.98(±1.3e-2)
	Ours	88.92(±3.5e-3)	89.27 (±4.9e-3)	88.16(±1.5e-2)	89.74 (±1.3e-2)	54.90(±6.5e-2)	57.27 (±5.7e-2)
FlanV2	<i>L</i>	81.57(±1.4e-3)	81.49(±1.6e-3)	70.53(±8.7e-3)	70.73(±6.6e-3)	37.53(±2.0e-2)	36.59(±1.7e-2)
	<i>E</i>	81.28(±2.8e-3)	81.26(±2.9e-3)	62.73(±9.0e-3)	64.63(±7.7e-3)	23.85(±6.3e-3)	27.52(±4.6e-3)
	Ours	87.25(±2.0e-3)	87.78 (±3.0e-3)	90.05(±4.0e-3)	91.01 (±4.3e-3)	70.15(±1.0e-2)	72.58 (±1.3e-2)
	Ours ^{clip}	87.93(±1.2e-3)	88.36 (±1.6e-3)	91.73(±2.4e-3)	92.50 (±2.5e-3)	73.00(±7.8e-3)	75.08 (±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 considerably more efficient than comparable SOTA approaches like ZoomEye
- ▶ Further applications of internal representations 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).