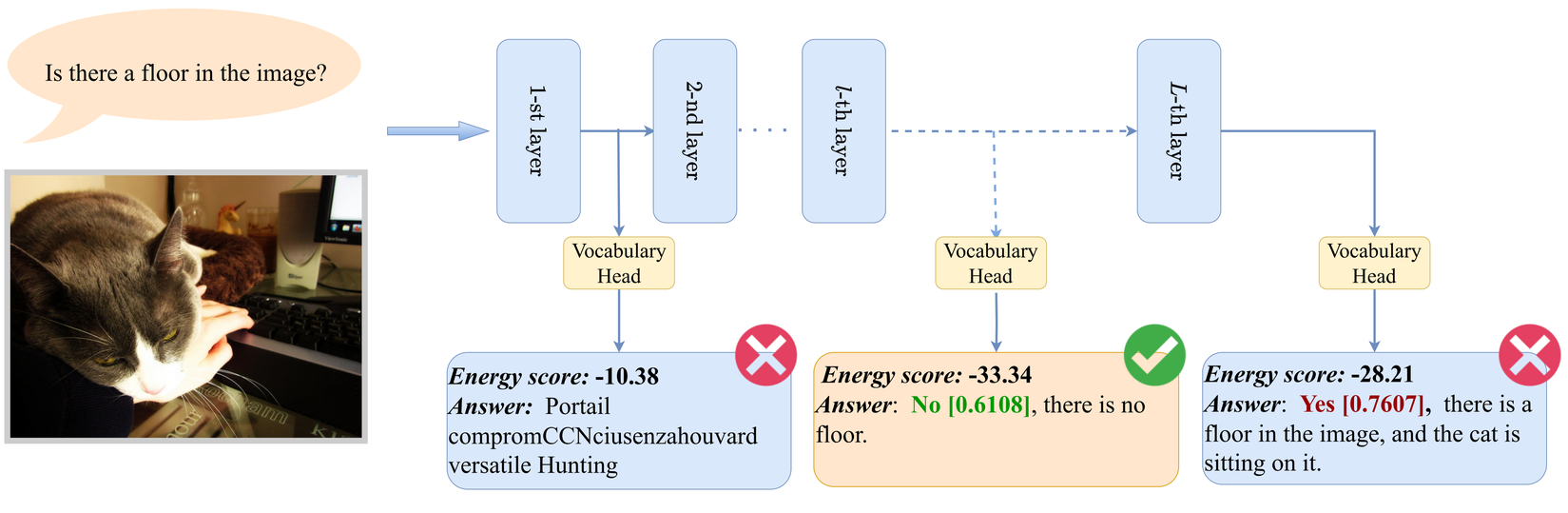
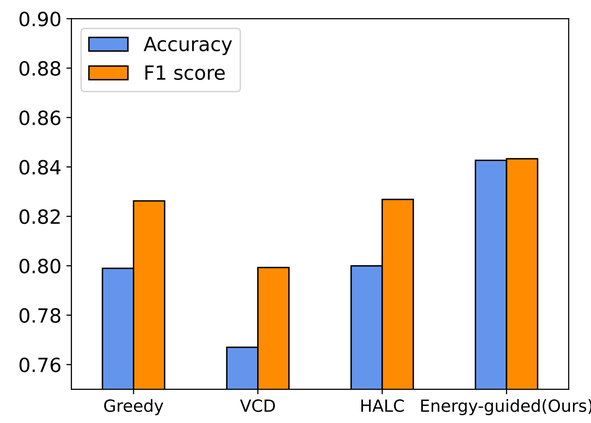
Energy-Guided Decoding for Object Hallucination Mitigation

Xixi Liu[1], Ailin Deng[2], Christopher Zach[1]
[1] Chalmers University of Technology
[2]National University of Singapore

Summary





Unlike existing hallucination mitigation methods such as VCD [1] (which necessitates generating a sophisticated noisy version of the original visual inputs), OPERA [2](which relies on the beam-searching decoding mechanism), HALC [4] (requiring a pre-defined layer bucket and an external detector), and MMVP [3](relies on additional fine-tuning), our method is derived through the lens of internal states of a language decoder. It avoids the need of visual distortion, or fine-tuning visual encoders, or external detectors making it free from contrastive decoding. More importantly, the energy score at each layer can be computed with a single forward pass, making our method significantly less computationally demanding compared to OPERA [2] and HALC [4].

Vision-Language Model Generation

The input tokens consisting of the visual tokens and language tokens is denoted with $\mathbf{x}$ with the length of T. VLMs are commonly trained in an autoregressive manner with a causal attention mask meaning that the prediction of the current token x_t only depends on the previous tokens, formally,

$$\mathbf{h} = \mathsf{VLM}(\mathbf{x}) = \{h_0, h_1, \cdots, h_{T-1}\}, \tag{1}$$

where $\mathbf{h}$ is the output state of the final layer of language decoder, and the size of h_t is f_{dim}. The next token predictive distribution is defined as

$$p(x_t|x_{<t}) = \mathsf{Softmax}[\mathcal{H}(h_t)], \tag{2}$$

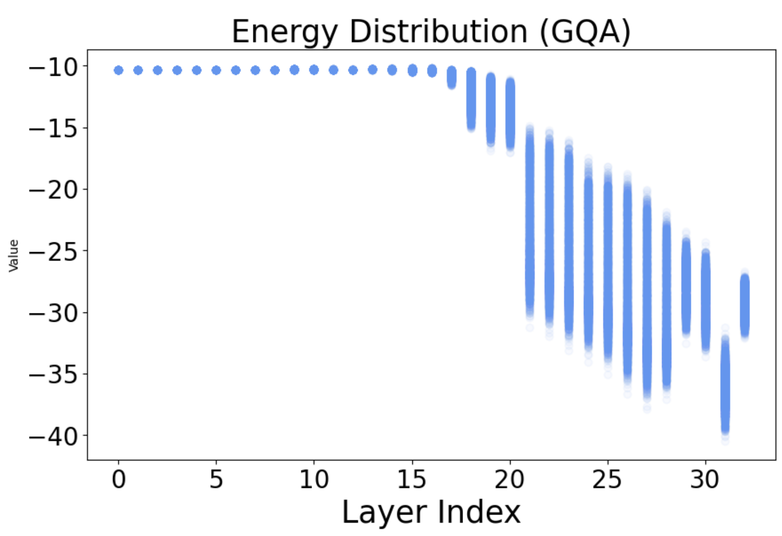
where $x_{<t}$ denotes the sequence of tokens before t-th position $\{x_i\}_{i=0}^{t-1}$ and $\mathcal{H} \in R^{f_{\mathsf{dim}} \times V_{\mathsf{size}}}$ is the learned vocabualry head.

Methods

We use this energy score to identify the layer whose hidden state provides the most reliable representation of the input. In detail, the energy score is given by

$$\textbf{Energy}(h_t^k) = -\operatorname{LogSumExp}[\mathcal{H}(h_t^k)] \tag{3}$$

where $\mathcal{H}(h_t^k)$ denote the logits calculated at layer k for predicting token t. The layer $k^* = \arg\min_k \textbf{Energy}(h_t^k)$ with the lowest score is consequently selected for decoding.



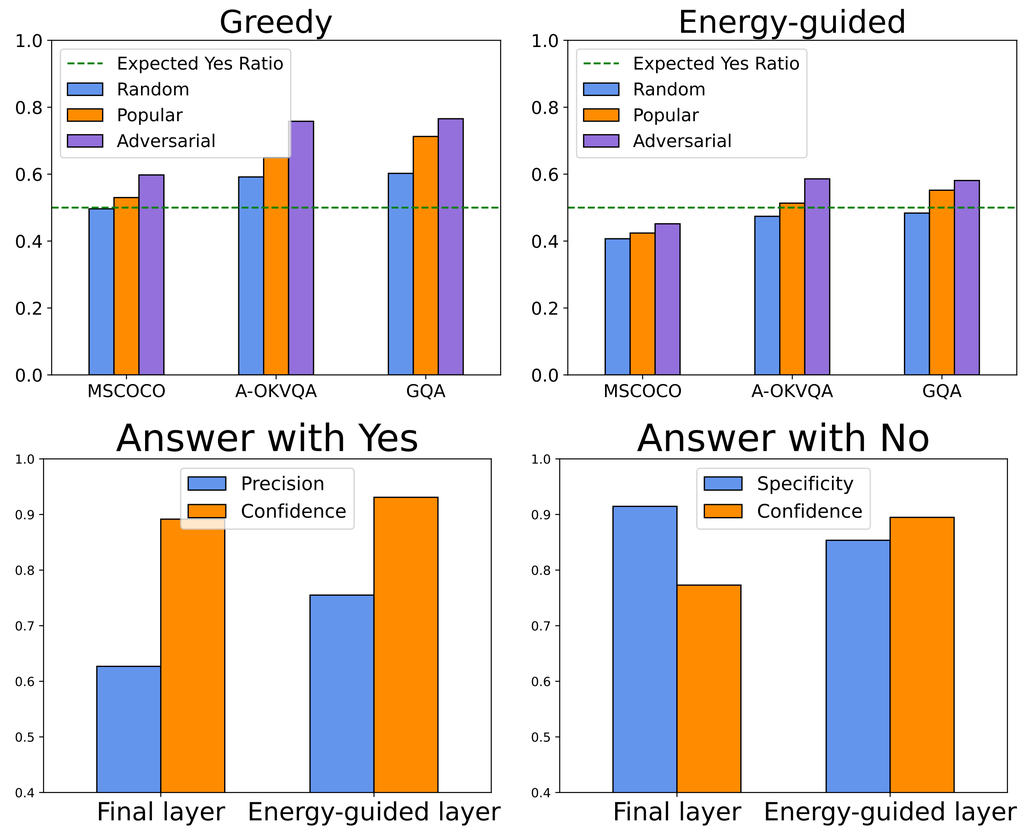
Selected Results

Our method consistently improves accuracy and F1 score on POPE benchmark with GQA dataset over three baseline methods with LLaVa-1.5.

Datasets	Settings	Decoding	LLaVA-1.5 Accuracy↑	Precision	Recall	F1 Score↑	Yes ratio	Δ_{gap} ↓	InstructBLIP Accuracy↑	Precision	Recall	F1 Score↑	Yes ratio	Δ_{gap} ↓
GQA	*Random*	Greedy	85.77	79.69	96.00	87.09	60.23	10.23	<u>86.90</u>	84.70	90.07	<u>87.30</u>	53.17	<u>3.17</u>
		VCD [1]	81.33	75.24	93.40	83.34	62.07	12.07	80.90	77.35	87.40	82.07	56.50	6.50
		HALC [3]	85.90	79.87	96.00	87.19	60.10	10.10	85.97	83.08	90.33	86.55	54.37	4.37
		OPERA [2]	<u>88.57</u>	85.47	92.93	<u>89.05</u>	54.37	<u>4.37</u>	**87.33**	87.23	87.47	**87.35**	50.13	**0.13**
		Energy (Ours)	**89.37**	90.70	87.73	**89.19**	48.37	**1.63**	86.53	92.35	79.67	85.54	43.13	6.87
	Popular	Greedy	74.73	67.35	96.00	79.16	71.27	21.27	76.37	70.70	90.07	79.21	63.70	13.70
		VCD [1]	71.53	64.79	94.33	76.82	72.8	22.80	73.00	67.97	87.00	76.32	64.00	14.00
		HALC [3]	74.87	67.48	96.00	79.25	71.13	<u>21.13</u>	74.50	68.61	90.33	77.99	65.83	15.83
		OPERA [2]	<u>79.83</u>	73.64	92.93	<u>82.17</u>	63.10	23.10	<u>79.77</u>	75.79	87.47	**81.21**	57.70	<u>7.70</u>
		Energy (Ours)	**82.53**	79.47	87.73	**83.40**	55.20	**5.20**	**80.27**	80.63	79.67	<u>80.15</u>	49.40	**0.60**
	Adversarial	Greedy	69.43	62.69	96.00	75.85	76.57	26.57	71.50	65.68	90.07	75.96	68.57	18.57
		VCD [1]	68.97	62.67	93.80	75.14	74.83	24.83	69.10	64.01	87.27	73.85	68.17	18.17
		HALC [3]	69.53	62.77	96.00	75.91	76.47	26.47	69.70	63.95	90.33	74.88	70.63	20.63
		OPERA [2]	<u>75.00</u>	68.40	92.93	<u>78.80</u>	67.93	<u>17.93</u>	<u>74.00</u>	68.91	87.47	<u>77.09</u>	63.47	<u>13.47</u>
		Energy (Ours)	**79.63**	75.50	87.73	**81.16**	58.09	**8.09**	**76.57**	75.02	79.67	**77.27**	53.10	**3.10**

The proposed method is less biased in terms of yes ratio, *i.e.*,

$$\Delta_{\text{gap}} = \left| \frac{\#\text{ of answers with yes}}{\#\text{ of total questions}} - 0.5 \right|. \tag{4}$$



References

[1] Mitigating object hal- 603 lucinations in large vision-language models through visual 604 contrastive decoding. CVPR, 2024.
[2] Opera: Alleviating hallucination in multi- 588 modal large language models via over-trust penalty and 589 retrospection-allocation. CVPR, 2024
[3] Eyes wide shut? exploring the661 visual shortcomings of multimodal llms. CVPR, 2024.
[4] Halc: Object hallucination reduc-522 tion via adaptive focal-contrast decoding. ICML, 2024.

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM