VEGAS: Mitigating Hallucinations in Large Vision-Language Models via Vision-Encoder Attention Guided Adaptive Steering

Supplementary Material

1. Experiment Setups

In VEGAS, we introduce adaptive logits steering based on the following vision attention block entropy (VABE):

$$\mathrm{VABE}_4^l = \frac{1}{H}\sum_{h=1}^{H} \mathrm{BE}_4\left(\tilde{\mathbf{A}}_v^{(l,h)}\right), \tag{1}$$

where H is the number of attention heads and $\tilde{\mathbf{A}}_v^{(l,h)}$ denotes the pre-softmax attention over image tokens at head h of layer l.

The weighting coefficient α in the above logits steering is adaptively determined as:

$$\alpha = \begin{cases} \alpha_1, & \text{if } \mathrm{VABE}_4^l > \eta, \\ \alpha_2, & \text{otherwise}, \end{cases} \tag{2}$$

In all experiments, we use VABE_4^{15} as the hallucination indicator. We choose $\eta = 0.31$ for all LLAVA-1.5 [4] and Shikra [1]. For LLAVA-1.5 and MINIGPT-4 [6], we use $\alpha_1 = 1.0$ and $\alpha_2 = 0.8$. And we set $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$ for Shikra.

In MINIGPT-4, instead of using image tokens, query tokens are provided to the LLM as the vision inputs. Calculated from the self-attention and image token cross-attention, each query represents information from multiple image patches [2, 3, 5]. Thus for MINIGPT-4, we use query token attention entropy instead of VABE as the indicator. Specifically we choose $\eta = 2.1$ as the threshold for query token attention entropy.

2. Additional Experiments

2.1. Impact of the logits weight α

In this ablation study, we vary the weight parameter α to assess its impact on overall performance. We evaluate three LVLMs on the CHAIR benchmark using different fixed $\alpha \in [0, 1]$. Fig. 1 presents the results. Generally, higher α values lead to better reductions in hallucinations. However, for Shikra, when α approaches 1, we observe extremely low object-hallucination rates but a tendency for the model to generate incomplete or truncated sentences. Accordingly, we adopt large α values for LLaVA-1.5 and MiniGPT-4, but a relatively smaller α for Shikra.

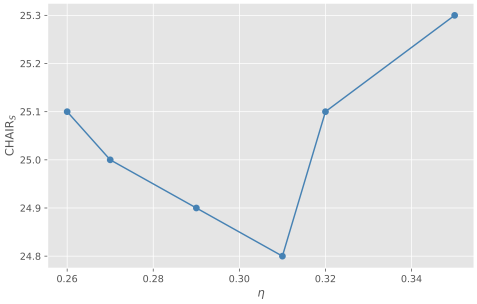

Figure 2. Ablation study on the threshold η in VEGAS using greedy decoding on LLaVA-1.5-7B for the CHAIR benchmark. We vary η across a range of values and observe how it affects performance: the optimal value is $\eta = 0.31$. When η is set too low or too high, VEGAS effectively behaves like a fixed α configuration ($\alpha = 0.8$ or $\alpha = 1.0$, respectively), which yields inferior results.

2.2. Ablation study on threshold η

As described in Sec. 1, we apply different values of α depending on whether the current token's VABE_4^{15} exceeds a threshold η. This technique, which we term Adaptive Logits Steering, enables dynamic weighting of the original and attention-replaced logits. In our ablation study, we vary η and evaluate performance on the CHAIR task. Fig. 2 shows that the optimal value is $\eta = 0.31$. When η is set much lower, the method defaults to $\alpha = 0.8$; when η is too high, it effectively behaves like fixed $\alpha = 1.0$, in both cases yielding inferior performance.

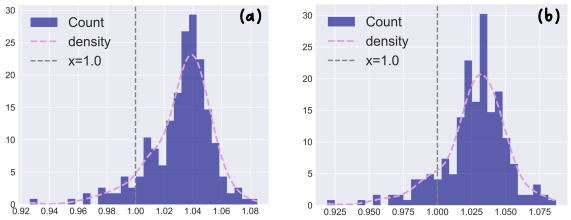

Figure 3. Ratio of hallucinated-token to non-hallucinated-token block entropy (BE_4) for vision-attention maps in LLAVA-1.5. (a) Layer 15, (b) Layer 31. All values are calculated on real-object tokens. Ratios above 1.0 indicate that hallucinated tokens tend to exhibit higher block entropy. This pattern holds consistently across many layers within the LLM, including middle layers and final layers.

2.3. Hallucination indicator at various layers

As described in Sec. 1, we use VABE_4^{15} as our primary hallucination indicator. Fig. 3 shows that many layers within

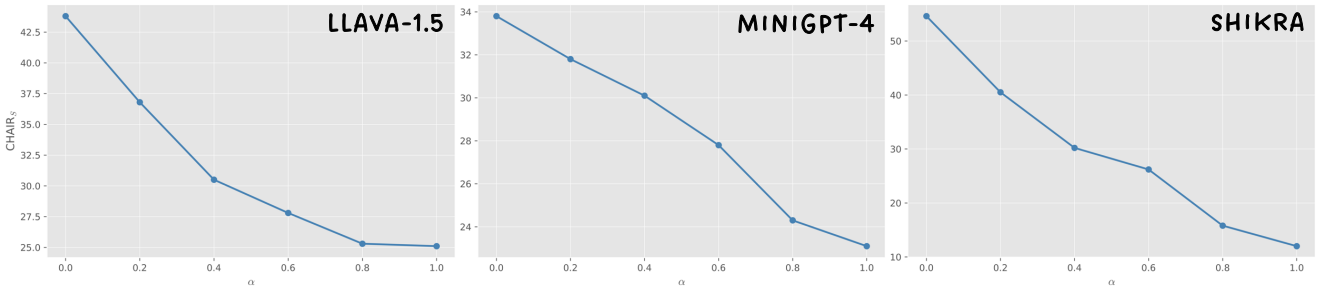

Figure 1. Ablation study on the logits weight α in using greedy decoding across three LVLM models on the CHAIR benchmark. For each model, we report performance when α is fixed to different values in the range $[0, 1]$. The results illustrate how increasing α, i.e., placing greater weight on the attention-replaced logits, impacts hallucination reduction.

the LLM (including both middle and final layers) can serve as effective indicators of hallucination risk. To maintain simplicity and consistency in our framework, we thus adopt VABE_4^{15} as the default indicator throughout.

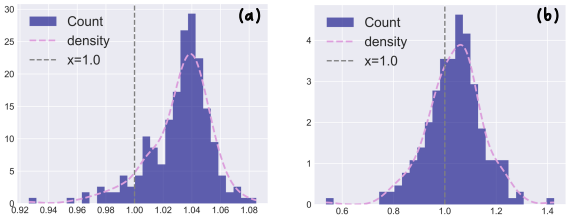

Figure 4. Ratio of hallucinated-token to non-hallucinated-token vision-attention block entropy (VABE) in LLAVA-1.5 at Layer 15: (a) calculated using VABE_4^{15}; (b) calculated using VABE_8^{15}. All values are calculated on real-object tokens. The ratio typically exceeds 1.0 when using VABE_4^{15}, indicating that hallucinated tokens tend to exhibit larger VABE_4^{15}. However, when using a larger block size (e.g., 8), VABE no longer clearly differentiates between hallucinated and non-hallucinated tokens.

2.4. VABE block size in hallucination detection

As defined in Eq. (1), VABE is derived from our introduced Block Entropy metric, where the choice of block size is a critical hyperparameter. Accordingly, we evaluate the effect of varying block size on the effectiveness of the hallucination indicator. As shown in Fig. 4, smaller block sizes (e.g., 4) provide clearer discrimination between hallucinated and non-hallucinated tokens.

References

[1] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[3] Shihong Ling, Yue Wan, Xiaowei Jia, and Na Du. Driveblip2: Attention-guided explanation generation for complex driving scenarios. *arXiv preprint arXiv:2506.22494*, 2025. 1

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[5] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. 1

[6] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1