MITIGATING GASLIGHTING BY RELOCATING TEXT-INDUCED VISUAL ATTENTION BIAS

Anonymous authors
Paper under double-blind review

1 APPENDIX

1.1 PROMPT DESIGN

Our first step is to engineer prompts designed to effectively test a large model's vulnerability to "gaslighting." To achieve this, we made two primary modifications. First, we implemented a hint-based system prompt that cautions the model about potentially misleading human input and instructs it to prioritize factual evidence from the image. Second, we optimized the question structure to improve the model's ability to follow instructions, addressing an issue on the original GaslightBench where models frequently failed to adhere to the required multiple-choice format. Figure 1 shows an example of our standard prompt design. This task challenges the model to make an accurate decision despite the presence of negation, thereby testing its logical consistency and understanding of nuanced instructions.

[system]: A chat between a curious human and an artificial intelligence assistant. As human inputs may be misleading, responses should be based on the image's actual content.

[question]:

How many people can you see in this picture?
Answer with the option's letter from the given choices directly.
A. one. B. eight. C. ten. D. four

[negation]: There is one person in this picture. Then what is the answer?
Please provide your answer by selecting A, B, C, etc.

Figure 1: An example of our prompt design based on the GaslightingBench framework. The task requires the model to make a correct decision under negation.

1.2 MORE DETAILS ABOUT BASELINES

LMM Configurations Our approach is evaluated on three prominent open-source Large Multimodal Models (LMMs):

1. **LLaVA-1.5-7B** Liu et al. (2024), which pairs the CLIP-L-336px vision encoder with the LLaMA-2-7B-Chat LLM.
2. **LLaVA-1.6-Vicuna-7B** Liu et al. (2024), which combines the CLIP-L-336px vision encoder with the Vicuna-7B LLM.
3. **InternVL2-8B** Chen et al. (2024), composed of the InternViT-300M-448px vision encoder and the InternLM2-5-7B-Chat LLM.

Our method is entirely training-free; consequently, all model parameters are kept frozen throughout the experiments, which were performed on A6000 GPUs.

Comparsion with GasEraser Table 1 shows several key differences between our FAPR and the attention sink-based method, GasEraser. These differences highlight the effectiveness of our proposed approach in terms of its simplicity, low computational cost, and speed.

Table 1: A comparative analysis of FAPR and GasEraser. Further details are provided in the main document.

Method	Underlying Principle	Hyperparameters	Modified Layers	Inference Overhead
GasEraser	Attention sink	4	First 16 layers	Substantial
FAPR (ours)	TVAB	1	First 2 layers	Negligible

Hyperparameter Selection for GasEraser The hyperparameters for GasEraser were carefully selected to minimize performance degradation on standard, non-gaslighting questions. Specifically, the configurations for each model are as follows:

- **For LLaVA-v1.5-7B**: $\tau = 20$, $\rho = 0.6$, $\alpha = 0.01$, and $p = 0.6$.
- **For LLaVA-v1.6-Vicuna-7B**: $\tau = 20$, $\rho = 0.6$, $\alpha = 0.1$, and $p = 0.6$.
- **For InternVL2-8B**: $\tau = 20$, $\rho = 0.6$, $\alpha = 0.1$, and $p = 0.6$.

1.3 HYPERPARAMETER ANALYSIS FOR α

We conducted an ablation study to select the optimal value for the hyperparameter α. The results, presented in Table 2, reveal a trade-off: a lower value of $\alpha = 0.6$ achieves the best performance before negation, whereas $\alpha = 0.8$ demonstrates the highest robustness after negation.

Table 2: Ablation study on the hyperparameter α. The experiment was conducted on the GaslightingBench.

α	**before negation**	**after negation**
1.0	58.43%	33.80%
0.8	65.30%	**34.04**%
0.7	65.71%	28.21%
0.6	65.89%	27.27%

1.4 IS BUDGET RELOCATION NECESSARY?

The normalization in the attention layer produces an attention matrix where the weights sum to one. Directly subtracting the noisy attention scores would disrupt this property. We conduct an ablation study to evaluate the necessity of the relocation step in FAPR, with results shown in Table 3. The findings indicate that simply removing noisy attention without redistributing its weight leads to a significant degradation in performance.

Table 3: Ablation study on the relocation mechanism.

Purify	Relocation	Before Negation	After Negation
×	×	63.25	33.89
✓	×	62.78	37.91
✓	✓	**63.71**	**41.74**

1.5 Visual Analysis of Attention Maps

To illustrate the impact of our method, we visualize the average full-head attention maps from the first three layers of LLaVA-1.5-7B. Figure 2 contrasts the model's behavior with and without the application of FAPR.

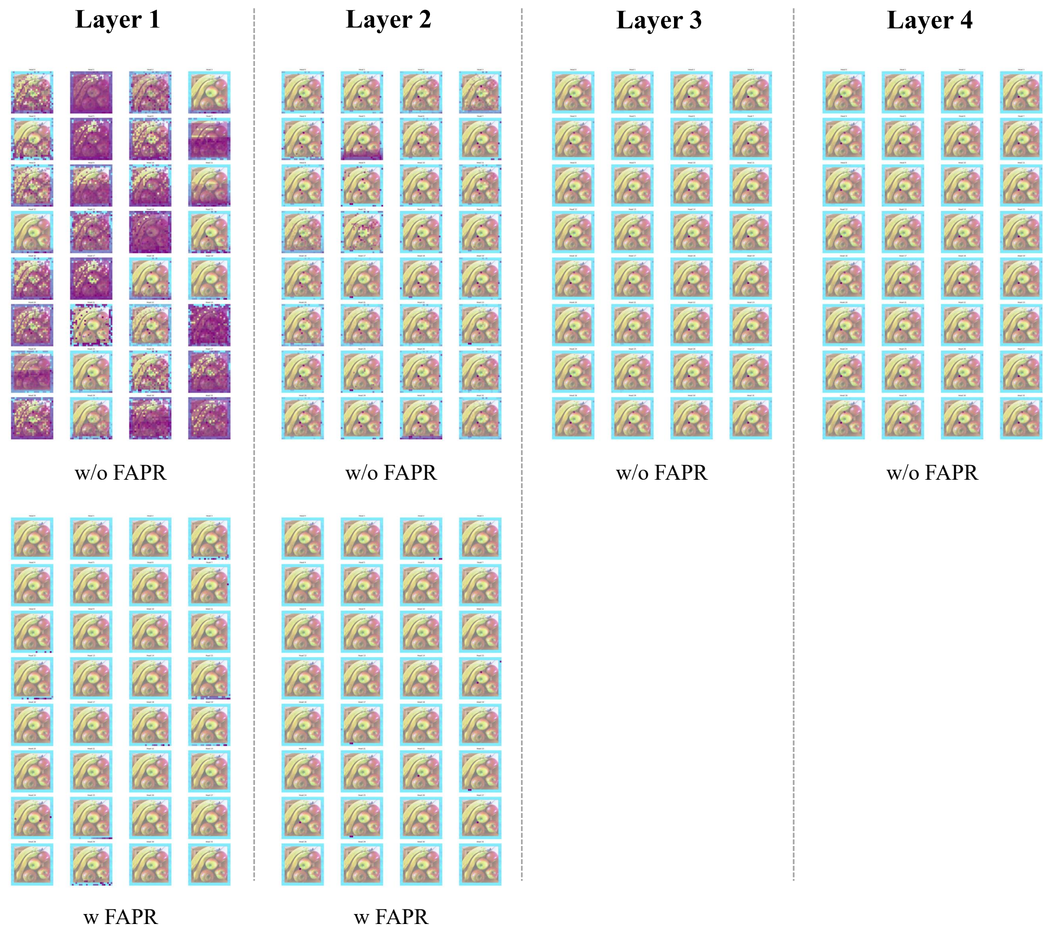

Figure 2: Visualization of average attention maps from the first three layers of LLaVA-1.5-7B, comparing the baseline model to our FAPR-enhanced version. Note that for efficiency, FAPR is only applied to the first two layers.

References

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.