VIHD: Visual Intervention-based Hallucination Detection for Medical Visual Question Answering

Jiayi Chen[1(✉)], Benteng Ma[4], Zehui Liao[1], Winston Chong[2,3], Yasmeen George[1], and Jianfei Cai[1]

[1] Department of Data Science & AI, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
[2] Alfred Health Radiology, Alfred Health, Melbourne, VIC 3004, Australia
[3] School of Translational Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, VIC 3800, Australia
[4] Hong Kong Polytechnic University, Hong Kong SAR, China
jiayi.chen@monash.edu

Abstract. While medical Multimodal Large Language Models (MLLMs) have shown promise in assisting diagnosis, they still frequently generate hallucinated responses that appear linguistically plausible but lack visual evidence. Such hallucinations pose risks to clinical decision-making and necessitate effective detection. Existing introspective detection methods primarily perform uncertainty estimation or logical verification by analyzing model responses conditioned on original or perturbed inputs. However, such external perturbations are often heuristic and context-agnostic, which overlooks the internal cross-modal dependency between generated tokens and related visual tokens during decoding. To address this issue, we propose VIHD, a **V**isual **I**ntervention-based **H**allucination **D**etection method that leverages targeted visual token masking to calibrate semantic entropy for more effective hallucination detection. VIHD locates visually dominant decoder layers via *Visual Dependency Probing (VDP)*, executes *Visual Intervention Decoding (VID)* via token masking to calibrate the semantic distribution, and quantifies the resulting *Calibrated Semantic Entropy (CSE)* as a reliable hallucination signal. Extensive experiments on three medical VQA benchmarks with two medical MLLMs demonstrate that VIHD consistently outperforms state-of-the-art methods, underscoring the importance of fine-grained visual dependency for hallucination detection. The code will be available at https://github.com/Jiayi-Chen-AU/VIHD.

Keywords: Vision-Language Model · Hallucination Detection · Visual Question Answering

1 Introduction

Medical multimodal large language models (MLLMs) have shown remarkable efficacy in assisting medical diagnosis, supporting visual question answering

Corresponding author: J. Chen.

(VQA) and automated report generation [16,23,11,27]. However, these models are susceptible to *hallucinated responses* that are linguistically coherent yet contradict visual evidence [9,18,12]. This issue poses risks to clinical decision-making and underscores the need for effective hallucination detection.

Existing hallucination detection methods can be broadly categorized into three paradigms: supervised detection, external verification, and introspective detection. **Supervised methods** [6,2,26,22] typically train auxiliary detectors on annotated hallucination datasets. While effective in-distribution, they hinge on labor-intensive annotations and generalize poorly to novel domains or architectures. **External verification** [3,29,28,24] leverages auxiliary LLMs, MLLMs, or knowledge bases to validate model outputs, but introduces substantial computational overhead and relies heavily on external tools. In contrast, **introspective approaches** [17,5,18,25,12] are entirely self-contained, requiring neither extra training nor external resources, making them efficient and deployment-friendly. Existing introspective approaches typically estimate token uncertainty [17,20], semantic uncertainty [5,30,18], or logical consistency [25,12] by analyzing model responses to original or perturbed inputs. However, such external perturbations are inherently heuristic and context-agnostic [10]. They treat MLLMs as black boxes, failing to probe the internal causal dependencies between generated tokens and visual evidence during decoding.

To mitigate this issue, we propose VIHD, a training-free **V**isual **I**ntervention-based **H**allucination **D**etection method, that probes the fidelity of cross-modal reasoning through targeted token-level interventions. VIHD operates in three stages. First, we identify visually dominant layers in the LLM decoder where cross-modal attention is most prominent. Secondly, we introduce a visual intervention decoding scheme that masks high-attention visual tokens to produce intervened responses, thereby revealing the model's dependency on specific visual evidence. Third, we design an adaptive strategy that integrates original and intervened semantic distributions via complementary or contrastive fusion, yielding a calibrated semantic entropy metric for hallucination detection.

Our main contributions are threefold: (1) We introduce VIHD, a training-free framework that harnesses fine-grained visual interventions to probe internal causal dependencies for hallucination detection. (2) We design an adaptive pipeline that identifies critical decoder layers, performs attention-guided visual token masking, and derives calibrated semantic entropy to quantify hallucination severity. (3) Extensive experiments on three medical VQA benchmarks and two medical MLLMs demonstrate the superiority of VIHD in hallucination detection.

2 Methodology

As illustrated in Fig. 1, VIHD detects hallucinations via fine-grained and targeted visual interventions. It comprises three components: (1) *Visual Dependency Probing*, which pinpoints visually dominant decoder layers for targeted intervention; (2) *Visual Intervention Decoding*, which performs attention-guided token masking to produce intervened responses; (3) *Calibrated Semantic En-*

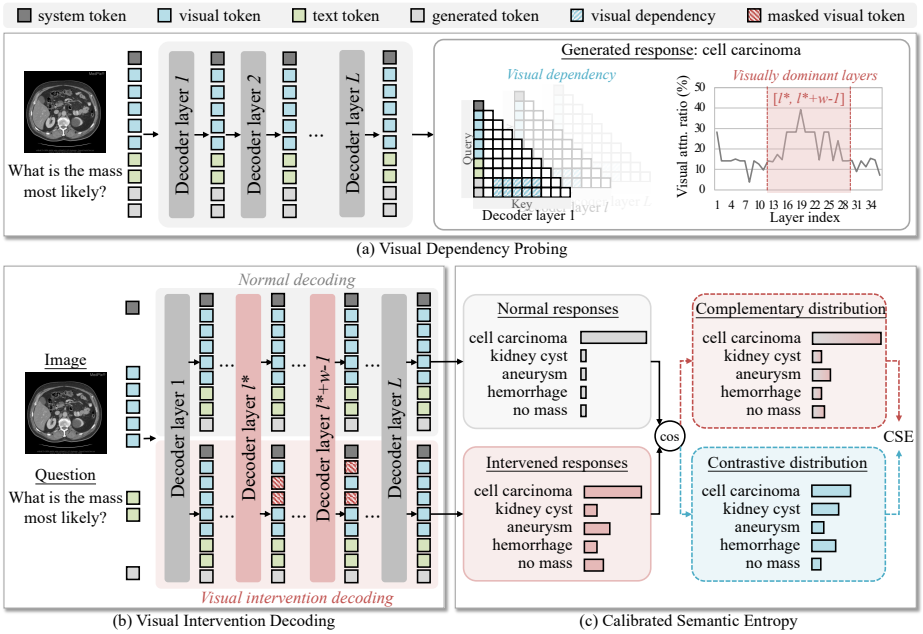

Fig. 1: Framework of VIHD. VIHD comprises (a) visual dependency probing, (b) visual intervention decoding, and (c) calibrated semantic entropy.

tropy, which calibrates semantic distribution based on normal and intervened responses and quantifies semantic entropy as a hallucination signal.

2.1 Visual Dependency Probing (VDP)

Given an input image x_v and a textual query x_q, a medical MLLM parametrized with θ generates a response $r = (r_1, \ldots, r_T)$ autoregressively, where T is the sequence length. The probability of generating response r [15] is factorized as:

$$p_\theta(r|x_v, x_q) = \prod_{t=1}^{T} p_\theta(r_t \mid x_v, x_q, r_{<t}). \tag{1}$$

where r_t is the t-th generated token and $r_{<t}$ is the preceding sequence.

Hallucinations often arise when tokens are generated primarily from linguistic priors without sufficient visual evidence. To quantify visual dependency during decoding, we first analyze the attention interactions between generated tokens and the input sequence. Let $\mathbf{Q}_r^{(l)} \in \mathbb{R}^{T\times d}$ denote the query matrix of generated tokens and $\mathbf{K}^{(l)} \in \mathbb{R}^{N\times d}$ represent the key matrix of the input sequence. The attention matrix $\mathbf{A}^{(l,h)} \in \mathbb{R}^{T\times N}$ for the h-th head is computed as:

$$\mathbf{A}^{(l,h)} = \text{softmax}\left(\frac{\mathbf{Q}_r^{(l,h)}\mathbf{K}^{(l,h)\top}}{\sqrt{d_h}}\right), \quad h \in \{1, \ldots, H\}, \tag{2}$$

where $d_h = \frac{d}{H}$ denotes the head dimension. We then quantify the layer-wise visual dependency $s^{(l)}$ by aggregating the attention weights assigned to visual tokens (indexed by $\mathcal{V}$) across all heads and generation tokens:

$$s^{(l)} = \frac{1}{H \cdot T} \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{v \in \mathcal{V}} A_{t,v}^{(l,h)}. \tag{3}$$

Finally, we localize a window of w consecutive layers with the strongest visual dependency $[l^*,\ l^*{+}w{-}1]$. The optimal starting layer l^* is defined as:

$$l^* = \underset{l}{\operatorname{argmax}} \left(\frac{1}{w} \sum_{i=0}^{w-1} s^{(l+i)} \right). \tag{4}$$

2.2 Visual Intervention Decoding (VID)

After identifing visually dominant layers via VDP, we propose the visual intervention decoding algorithm to generate the intervened response r'. Given a visual input x_v and a textual query x_q, VID dynamically intervens the decoding process by applying fine-grained and targeted visual token masking. For each visually dominant layer l, we calculate the attention weights between the generated token and visual tokens at each generation step t.

$$\mathbf{A}_{t,v}^{(l)} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{A}_{t,v}^{(l,h)}, \quad (l \in [l^*, l^*{+}w{-}1]). \tag{5}$$

We then mask the top-ρ fraction of visual tokens with the highest attention weight $\mathbf{A}_{t,v}^{(l)}$ to suppress their contribution. The intervened response r' is then generated conditioned on the perturbed visual representation. The probability of generating r' is formulated as:

$$p_\theta(r') = \prod_{t=1}^{T} p_\theta(r'_t | x_v, x_q, r'_{<t}). \tag{6}$$

2.3 Calibrated Semantic Entropy (CSE)

To enhance hallucination detection, we compute semantic entropy [5] over a calibrated semantic distribution. Specifically, we perform M independent sampling runs to obtain two sets of responses: M normal responses $\{r^{(i)}\}_{i=1}^{M}$ and M visual-intervened responses $\{r'^{(i)}\}_{i=1}^{M}$. Then, we utilize DeBERTa-Large-MNLI [7] to cluster them into N semantic-equivalence classes [13], yielding a normal semantic distribution $P(s|x_v, x_q)$ and a visual-intervened one $P(s'|x_v, x_q)$.

Crucially, visual intervention impacts vary by query granularity. General queries (*e.g.*, imaging modalities) remain robust to token masking, while detailed ones are relatively susceptible to semantic shifts. Accordingly, we propose

an adaptive fusion strategy based on distribution alignment. When the intervened distribution aligns closely with the original, we treat token masking as a weak perturbation and apply *complementary fusion* to reinforce shared semantics. Conversely, we adopt *contrastive fusion* to suppress masking-induced hallucinations. The calibrated semantic distribution $P_c(s, s')$ is formulated as:

$$P_c(s,s') = \begin{cases} \sigma\left(\frac{1}{1+\alpha}P(s|x_v,x_q) + \frac{\alpha}{1+\alpha}P(s'|x_v,x_q)\right), & \text{if } \cos(P(s),P(s')) \geq \tau, \\ \sigma\left((1+\alpha)P(s|x_v,x_q) - \alpha P(s'|x_v,x_q)\right), & \text{otherwise.} \end{cases} \tag{7}$$

where α denotes the fusion factor, $\sigma(\cdot)$ is the softmax function, and τ is the similarity threshold. Finally, we calculate the calibrated semantic entropy CSE to quantify hallucination severity.

$$CSE = -\sum P_c(s,s') \cdot \log P_c(s,s') \tag{8}$$

A higher calibrated semantic entropy indicates significant semantic dispersion, serving as a robust indicator of hallucination.

3 Experiments

3.1 Experimental Settings

Datasets. We conduct hallucination detection on the test split of three medical VQA datasets covering multiple imaging modalities (CT, MRI, and X-ray): VQA-RAD [14], SLAKE [19], and VQA-Med-2019 [1]. **VQA-RAD** contains 451 test samples with 200 open-ended and 251 close-ended questions. **SLAKE** contains 1061 test samples in English with 645 open-ended and 416 close-ended questions. **VQA-Med-2019** comprises 500 test samples with 436 open-ended and 64 close-ended questions.

Backbones. We perform hallucination detection on two advanced medical MLLMs, Hulu-4B [11] and Lingshu-7B [27], for a robust assessment across diverse architectures.

Evaluation Metrics. Following [18], we employ the GREEN model [21] to generate hallucination labels. The GREEN model identifies matched findings and clinical errors in generated responses and calculates the GREEN score as $\frac{\#matched}{\#matched+\#errors}$. Samples with GREEN < 1.0 are labeled as hallucinated, while others (GREEN $= 1.0$) are non-hallucinated. Performance is quantified using two metrics: (1) *Area Under the ROC Curve (AUC)*, which measures the ranking capability to differentiate hallucinated from non-hallucinated responses; and (2) *Area under the GREEN Curve (AUG)* [18], which evaluates the alignment between estimated hallucination scores and factual correctness.

Implementation Details. During hallucination label construction, we set the LLM sampling temperature to 0.1 to produce stable primary responses. During hallucination detection, we increase the temperature to 1.0 and perform $M = 10$ independent sampling runs following [18]. We adopt Nucleus sampling [8],

Table 1: Performance comparison of VIHD and seven competing methods.

Method	VQA-RAD				SLAKE				VQA-Med-2019			
	Open-Ended		All		Open-Ended		All		Open-Ended		All	
	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG
Hulu-4B												
AvgProb	62.01	53.68	60.13	67.02	60.30	68.24	59.13	72.59	70.48	51.56	67.90	55.00
MaxProb	62.52	53.54	61.17	66.98	60.37	68.21	59.26	72.62	71.45	51.86	69.45	55.39
AvgEnt	61.15	53.56	59.32	66.88	60.39	68.26	59.03	72.56	70.43	51.68	67.72	55.00
MaxEnt	62.49	53.71	61.12	66.98	60.45	68.22	59.24	72.61	71.67	52.74	69.64	55.45
RadFlag	50.00	49.30	50.00	61.26	50.00	59.02	50.00	68.12	50.00	21.74	50.00	22.84
SE	50.89	51.77	49.48	62.24	49.29	59.02	49.30	66.89	47.53	19.58	47.42	21.43
VASE	76.98	69.09	66.23	71.51	74.41	79.12	70.31	81.27	75.09	54.30	72.99	57.41
VIHD	**83.13**	**72.01**	**78.23**	**80.82**	**79.94**	**81.11**	**78.91**	**84.96**	**78.65**	**59.64**	**78.10**	**60.39**
Δ	+6.15	+2.92	+12.00	+9.31	+5.53	+1.99	+8.60	+3.69	+3.56	+5.34	+5.11	+2.98
Lingshu-7B												
AvgProb	56.23	39.63	54.72	59.29	53.94	75.76	53.30	77.39	55.61	35.54	55.11	36.81
MaxProb	56.21	39.67	54.71	59.30	53.93	75.75	53.30	77.38	55.59	35.54	55.12	36.78
AvgEnt	56.28	39.73	54.80	59.34	53.97	75.82	53.34	77.42	55.68	35.63	55.11	36.90
MaxEnt	56.27	39.74	54.74	59.33	53.99	75.82	53.37	77.44	55.64	35.59	55.15	36.82
RadFlag	53.27	36.33	52.94	58.94	57.07	77.14	55.46	78.25	57.30	35.91	56.04	37.85
SE	50.37	35.93	50.92	55.27	52.72	74.13	52.45	76.44	54.27	34.69	53.96	35.96
VASE	71.45	48.02	69.62	69.35	72.37	85.06	70.56	84.52	69.57	56.77	68.85	60.80
VIHD	**73.13**	**55.84**	**71.58**	**70.86**	**75.36**	**88.61**	**72.95**	**86.49**	**74.71**	**64.11**	**74.18**	**64.54**
Δ	+1.68	+7.82	+1.96	+1.51	+2.99	+3.55	+2.39	+1.97	+5.14	+7.34	+5.33	+3.74

employ a sliding window of width $w = \frac{L}{2}$ for visual dependency probing, and mask $\rho = 10\%$ of high-attention visual tokens in visual intervention decoding. For semantic distribution calibration, we set the cosine similarity threshold to $\tau = 0.95$ and the calibration coefficient to $\alpha = 1.0$. All experiments are conducted on a single NVIDIA A100 GPU with 80 GB of memory.

Comparison Methods. We compared VIHD with seven SOTA methods: (1) four token uncertainty-based methods: AvgProb, MaxProb, AvgEnt, and MaxEnt [17], (2) one consistency-based method: RadFlag [31], and (3) two semantic uncertainty-based methods: SE [5] and VASE [18]. VCD [15] and SID [10] are excluded for focusing on hallucination mitigation rather than detection.

3.2 Results

Table 1 presents the hallucination detection performance of VIHD against seven competing methods across three medical VQA benchmarks. The best performances are highlighted in **bold** and improvements over the second-best one are shown in blue. As reported in Table 1, VIHD consistently outperforms competing methods by a notable margin across all VQA datasets and medical MLLMs. On **Hulu-4B**, it improves the average AUC by 8.57% for all questions and 5.08% for open-ended questions. Notably, the most substantial gain is observed on VQA-RAD, where VIHD achieves a 12% AUC increase across all questions. On

Table 2: Ablation study on VQA-RAD and VQA-Med-2019 using Hulu-4B.

Module			VQA-RAD				VQA-Med-2019			
VDP	VID	CSE	Open-Ended		All		Open-Ended		All	
			AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG
✗	✗	✗	50.89	51.77	49.48	62.24	47.53	19.58	47.42	21.43
✗	✓	✗	79.66	71.79	67.85	73.96	76.23	56.54	74.77	59.23
✗	✓	✓	79.97	71.65	76.30	80.35	77.64	58.80	76.56	59.73
✓	✓	✓	**83.13**	**72.01**	**78.23**	**80.82**	**78.65**	**59.64**	**78.10**	**60.39**

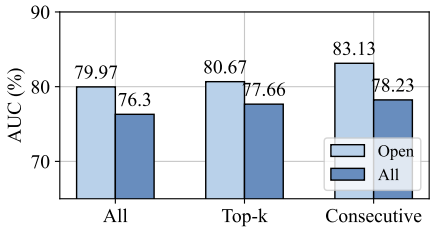

Fig. 2: Variants of VDP.

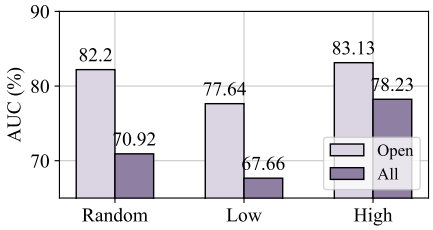

Fig. 3: Variants of visual token masking.

the more capable **Lingshu-7B** backbone, VIHD maintains its advantage with average AUC improvements of 3.22% overall and 3.27% on open-ended questions, with a substantial 5.33% gain on VQA-Med-2019. These results validate VIHD's effectiveness in detecting hallucinations across diverse medical imaging modalities and question types. The consistent improvements over VASE across both model capacities and dataset characteristics underscore the efficacy of internal fine-grained visual intervention in hallucination detection. Furthermore, VIHD requires 11.59s per sample at inference, comparable to VASE (11.55s), demonstrating performance gains without extra computational cost.

3.3 Ablation Study and Further Analysis

Ablation Study. Table 2 presents the component-wise ablation study on VQA-RAD and VQA-Med-2019 using Hulu-4B. Without any component, the baseline, Semantic Entropy [5], yields overall AUC scores of 49.48% and 47.42% on VQA-RAD and VQA-Med-2019, respectively. The introduction of visual intervention decoding (VID) substantially improves AUC by 18.37% and 27.35%, demonstrating the effectiveness of targeted visual intervention. Augmenting VID with calibrated semantic entropy (CSE) further boosts performance by 8.45% and 1.79%. Finally, incorporating visual dependency probing (VDP) to focus interventions at visually dominant layers leads to the best overall performance. These results validate the effectiveness of each proposed component.

Variants of Visual Dependency Probing (VDP). As shown in Fig. 2, we evaluate three layer selection strategies for visual dependency probing: (1) *all-layer selection*, which applies token masking across all decoder layers; (2)

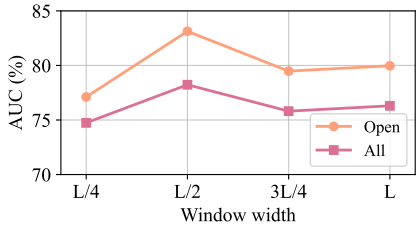

Fig. 4: Effect of sliding window width.

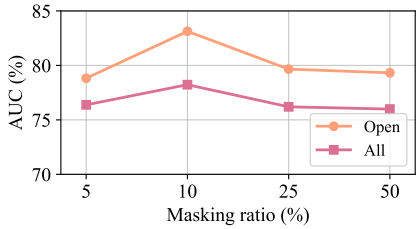

Fig. 5: Effect of masking ratio.

Table 3: Performance comparison on VQA-RAD using Hulu-4B with Top-K sampling.

Method	VQA-RAD			
	Open-Ended		All	
	AUC	AUG	AUC	AUG
AvgProb	65.06	53.51	60.68	69.58
MaxProb	65.89	53.24	61.73	69.79
AvgEnt	63.61	51.83	59.88	69.30
MaxEnt	65.06	53.41	61.73	69.85
RadFlag	50.00	44.77	50.00	62.54
SE	48.21	42.30	48.56	58.64
VASE	76.84	67.50	70.40	75.48
VIHD (Ours)	80.56	71.85	77.07	80.25
Δ	+3.72	+4.35	+6.67	+4.77

attention-based selection, which identifies isolated layers with peak visual dependency; and (3) *consecutive attention-based selection*, which selects a sequential span of high cumulative dependency to enforce continuity. Compared to the all-layer selection, the attention-based strategy yields a 1.36% AUC gain on VQA-RAD, highlighting the necessity of targeting visually dominant layers. Enforcing layer continuity further improves AUC by 0.57%. Therefore, we adopt the consecutive attention-based strategy as our default.

Variants of Visual Token Masking. We evaluate three token masking strategies within visual intervention decoding: (1) *random masking*, (2) *low-attention masking*, and (3) *high-attention masking*. As shown in Fig. 3, high-attention masking outperforms random masking by 7.41% AUC on VQA-RAD, which demonstrates that targeted perturbations of salient visual tokens facilitate effective semantic calibration. Conversely, low-attention masking acts merely as a subtle augmentation with negligible impact, thus leading to limited detection. Therefore, we adopt high-attention masking in visual intervention decoding.

Effect of Sliding Window Width. We also analyze the impact of sliding window width $w = \{\frac{L}{4}, \frac{L}{2}, \frac{3L}{4}, L\}$ on intervention efficacy. As illustrated in Fig. 4, performance peaks at $w = \frac{L}{2}$. Narrower windows $\frac{L}{4}$ restrict contextual propagation of perturbations, limiting their discriminative power, while wider windows $\frac{3L}{4}$ introduce excessive noise by intervening semantically heterogeneous layers. The $\frac{L}{2}$ configuration strikes an optimal balance, enabling localized yet contextually coherent perturbation propagation. We adopt $w = \frac{L}{2}$ for all experiments.

Effect of Masking Ratio. We investigate the impact of masking ratio $\rho \in \{5\%, 10\%, 25\%, 50\%\}$ on VQA-RAD. As shown in Fig. 5, insufficient masking ($\rho = 5\%$) yields weak perturbations that fail to calibrate semantic distribu-

tion, while aggressive masking ($\rho > 25\%$) corrupts visual semantics and degrades calibration fidelity. In contrast, the ratio of 10% achieves the optimal detection performance by introducing discriminative yet non-destructive interventions. Therefore, we fix $\rho = 10\%$ throughout experiments.

Effect of Sampling Strategy. To assess robustness under diverse decoding strategies, we evaluate all methods using Top-K sampling (K=50) [4] on VQA-RAD. As shown in Table 3, VIHD maintains superior performance across both AUC and AUG metrics, surpassing the strongest baseline (VASE) by 6.67% in AUC and 4.77% in AUG overall. These results validate the effectiveness of visual intervention for hallucination detection across multiple sampling paradigms.

4 Conclusion

We have proposed **V**isual **I**ntervention-based **H**allucination **D**etection (VIHD), a training-free method that exploits fine-grained visual intervention to detect hallucinations. VIHD identifies visually dominant decoder layers, performs visual intervention decoding to derive calibrated semantic entropy as a hallucination score. Extensive experiments on three medical VQA datasets and two MLLMs have demonstrated the superiority of VIHD in hallucination detection.

Acknowledgments. This work was supported by the Medical Research Future Fund (MRFF) under Grant NCRI000074.

Disclosure of Interests. The authors declare no relevant competing interests.

References

1. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF. vol. 2380 (2019)
2. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
3. Cohen, R., Hamri, M., Geva, M., Globerson, A.: LM vs LM: Detecting factual errors via cross examination. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12621–12640 (2023)
4. Fan, A., Lewis, M., Dauphin, Y.: Hierarchical neural story generation. In: ACL. pp. 889–898 (2018)
5. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
6. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: AAAI. vol. 38, pp. 18135–18143 (2024)
7. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced bert with disentangled attention. In: ICLR (2021)
8. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: ICLR (2020)

9. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR. pp. 13418–13427 (2024)
10. Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., Zhao, P.: Self-introspective decoding: Alleviating hallucinations for large vision-language models. In: ICLR (2024)
11. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
12. Jin, M., Liao, Z., Xia, Y.: V-Loop: Visual logical loop verification for hallucination detection in medical visual question answering. arXiv preprint arXiv:2601.18240 (2026)
13. Kuhn, L., Gal, Y., Farquhar, S.: Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In: ICLR (2023)
14. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(1), 1–10 (2018)
15. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR. pp. 13872–13882 (2024)
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS **36**, 28541–28564 (2023)
17. Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., Karray, F.: Reference-free hallucination detection for large vision-language models. In: EMNLP. pp. 4542–4551 (2024)
18. Liao, Z., Hu, S., Zou, K., Fu, H., Zhen, L., Xia, Y.: Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In: MICCAI. pp. 669–679. Springer (2025)
19. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: ISBI. pp. 1650–1654. IEEE (2021)
20. Ma, H., Chen, J., Zhou, J.T., Wang, G., Zhang, C.: Estimating LLM uncertainty with evidence. arXiv preprint arXiv:2502.00290 (2025)
21. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: GREEN: Generative radiology report evaluation and error notation. In: EMNLP. pp. 374–390 (2024)
22. Prabhakaran, V., Aggarwal, P., Verma, V.K., Swamy, G., Saladi, A.: Vade: Visual attention guided hallucination detection and elimination. In: ACL. pp. 14949–14965 (2025)
23. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: MedGemma technical report. arXiv preprint arXiv:2507.05201 (2025)
24. Sun, Z., Zang, X., Zheng, K., Xu, J., Zhang, X., Yu, W., Song, Y., Li, H.: ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In: ICLR (2025)
25. Wu, J., Liu, Q., Wang, D., Zhang, J., Wu, S., Wang, L., Tan, T.: Logical closed loop: Uncovering object hallucinations in large vision-language models. In: ACL. pp. 6944–6962 (2024)

26. Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., Zhu, L.: Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In: AAAI. vol. 39, pp. 25543–25551 (2025)
27. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
28. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: MM-Vet: Evaluating large multimodal models for integrated capabilities. In: ICML. pp. 57730–57754. PMLR (2024)
29. Zhang, J., Li, Z., Das, K., Malin, B., Kumar, S.: SAC3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In: EMNLP. pp. 15445–15458 (2023)
30. Zhang, R., Zhang, H., Zheng, Z.: VL-Uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919 (2024)
31. Zhang, S., Sambara, S., Banerjee, O., Acosta, J., Fahrner, L.J., Rajpurkar, P.: RadFlag: A black-box hallucination detection method for medical vision language models. arXiv preprint arXiv:2411.00299 (2024)