Synergistic Perception-Reasoning Governance: Grounding Medical MLLMs with Verifiable Anatomical Evidence

Rui Hao[1], Qiankun Li[2,3†], Junyuan Mao[4], Linghao Meng[4], Dirui Xie[1], Dayu Tan[5], and Zhigang Zeng[1†]

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Imperial Global Singapore, Imperial College London, London, United Kingdom
[3] Nanyang Technological University, Singapore, Singapore
[4] National University of Singapore, Singapore, Singapore
[5] Anhui University, Hefei, China
ruihao@hust.edu.cn; q.li2@imperial.ac.uk; zgzeng@hust.edu.cn
[†]Corresponding author.

Abstract. Multimodal large language models (MLLMs) show strong promise for clinical VQA and radiology report generation, yet inference-time hallucinations still undermine trustworthy use: models can produce fluent conclusions that conflict with imaging evidence. Existing mitigation strategies typically rely on additional training, external retrieval/-knowledge bases, or multi-stage post-hoc verification, which increases cost and pipeline complexity and often generalizes poorly across models and tasks. To address this, we propose a holistic, training-free evidence-injection framework that systematically mitigates hallucinations through dual-side evidence injection. By leveraging ROI priors acquired using MedSAM in our implementation, we recalibrate the visual perception trajectory via ROI-guided activation modulation while anchoring the textual reasoning trajectory by mapping anatomical coordinates into discrete semantic tokens as verifiable external memory. Then we introduce a task-aware dynamic router to select modality-specific interventions based on task semantics, balancing perceptual grounding and linguistic fluency. We conduct systematic evaluations on 2 tasks and 5 datasets using `LLaVA-1.5-7B`, `LLaVA-Med-1.5-7B`, `Qwen3-VL-8B/32B`, and `InternVL-3.5-8B/38B`. Controlled ablations and visualizations further validate the framework, which consistently outperforms baselines across medical benchmarks, improving close-ended accuracy by up to $\sim$ **6**% $\uparrow$ and reducing open-ended hallucinations by $\sim$ **35**% $\downarrow$. The code has been made available on GitHub: https://github.com/Henry991115/SPRG.

Keywords: Multimodal large language models· Visual misinterpretation hallucination· Hallucination mitigation · Evidence injection.

1 Introduction

MLLMs combine visual perception and language generation with emerging clinical reasoning, enabling medical VQA and radiology report generation [1–5].

However, hallucinations remain a key barrier to trustworthy deployment [6], as models may generate fluent outputs that are unsupported by or even contradict imaging evidence [7,8]. Recent benchmarks categorize medical hallucinations into visual misinterpretation, knowledge deficiency, and context misalignment [9]; in high-stakes settings, such evidence-inconsistent outputs can lead to incorrect lesion judgments and unsafe clinical attributions [10, 11].

To mitigate these risks, many training-free, inference-time methods have emerged. VCD [12], DoLa [13], and OPERA [14]adjust decoding by contrasting perturbed inputs, contrasting layer logits, or penalizing over-trusted tokens. AVISC [15], DAMRO [16], and PAI [17]calibrate attention or visual tokens by correcting outliers, filtering vision-side noise, and reducing text-dominant bias. M3ID [18] strengthens image grounding through visually dominant constraints. However, most existing mitigations rely on internal statistical calibration without explicit anatomical grounding, which limits generalization across clinical tasks with different evidence requirements.

In this work, we propose a holistic, training-free inference-time evidence-injection framework to mitigate visual misinterpretation hallucinations in medical MLLMs. Using ROI priors as verifiable anatomical evidence, we improve visual grounding and clinical consistency through the perception-reasoning intervention mechanism, including text-side evidence injection and vision-side activation recalibration over visual tokens, with a joint strategy for complementary gains. To handle heterogeneous clinical tasks, we further introduce a lightweight task-aware dynamic router that dynamically selects modality-specific interventions based on task semantics. Extensive experiments across diverse medical benchmarks, supported by controlled ablations and visualizations, show superior robustness over existing baselines, with up to $\sim$ **6**%$\uparrow$ improvement in close-ended reasoning accuracy and up to $\sim$ **35**%$\downarrow$ reduction in open-ended hallucinations. Our key contributions can be summarized as follows:

- ❶ We introduce a novel training-free framework that mitigates hallucinations by injecting dual-side verifiable anatomical evidence, bridging the gap between raw visual perception and clinical reasoning.
- ❷ We develop a perception-reasoning intervention mechanism that recalibrates the visual perception trajectory via ROI-guided visual-side activation modulation, while mapping anatomical coordinates into discrete semantic tokens as verifiable external memory to anchor the textual reasoning trajectory.
- ❸ We propose a task-aware dynamic router that adjusts modality-specific intervention strategies according to task semantics, thereby achieving a better balance between perceptual grounding and linguistic fluency.
- ❹ We establish a rigorous cross-model, cross-task evaluation framework involving 2 tasks, 5 datasets and 6 MLLMs, providing a highly comprehensive study to date on inference-time medical hallucination mitigation.

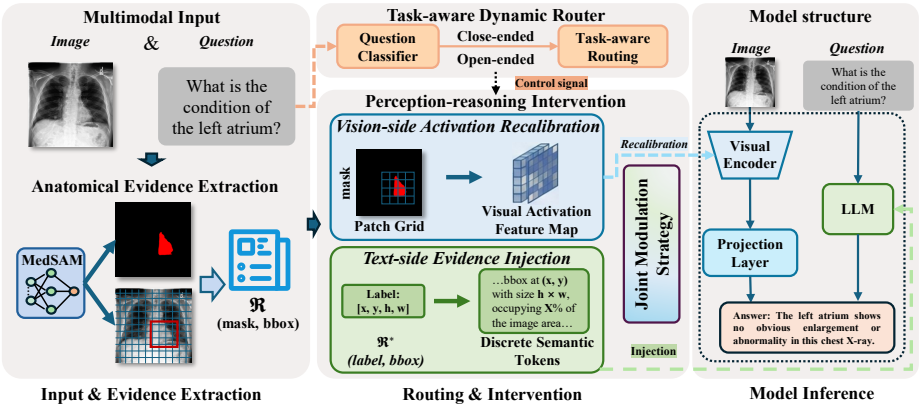

Fig. 1. The overall pipeline consists of all input, ROI prior extraction, intervention mechanism, task-aware routing, and model inference for the output generation.

2 Methods

Given an image I, a textual question q, and a pretrained MLLM parametrized by θ, the standard autoregressive generation is formulated as:

$$p_\theta(y \mid I, q) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, \mathrm{Enc}(I), q)\,, \tag{1}$$

where $\mathrm{Enc}(I) \in \mathcal{R}^{N\times d}$ denotes the sequence of N visual patch tokens. To mitigate visual misinterpretation, we introduce an intervention function that conditionally modulates the visual representation $\tilde{V}$ and textual prompt $\tilde{q}$ based on verifiable anatomical priors $\mathcal{R}$, yielding $p_{\theta,\phi}(y \mid I, q, R) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, \tilde{V}, \tilde{q})$.

As shown in Fig. 1, the pipeline comprises the anatomical evidence extraction, perception-reasoning intervention mechanism, and task-aware dynamic router.

2.1 Anatomical Evidence Extraction

We utilize MedSAM prompted by a candidate bounding box set $\mathcal{B} = \{b_k\}_{k=1}^{K}$ to extract verifiable anatomical evidence. For each valid box b_k, we obtain its corresponding segmentation mask m_k. After deduplication and invalid region filtering, we define the deterministic ROI prior set as $\mathcal{R} = \{(m_k, b_k, l_k)\}_{k=1}^{K'}$, where l_k is the anatomical semantic label derived from the clinical context.

2.2 Perception-reasoning Intervention Mechanism

Text-side Evidence Injection. We convert the ROI prior set $\mathcal{R}$ into a collection of verifiable region descriptors and inject them into the textual prompt in a

structured form. Given $\mathcal{R} = \{(m_k, \hat{b}_k)\}_{k=1}^{K'}$, we represent each region as a detection pair (l_k, b_k), where l_k is the anatomical label and $b_k = [x, y, w, h]$ specifies its location and scale. We then sequentially construct the structured region set, the serialized evidence block, and the injected prompt as

$$\tilde{q} = [\mathcal{S}(\mathcal{R}) \parallel q \parallel \eta_{\text{task}}], \quad \tilde{V} = V, \tag{2}$$

where $\mathcal{S}(\cdot)$ is a deterministic serialization operator. The operator $\parallel$ denotes token-level concatenation, and $\eta_{\text{task}} =$ ``Answer Yes/No'' is optionally appended for binary QA to reduce format drift. The final prediction is generated by conditioning on $(I, \tilde{q})$.

Vision-side Activation Recalibration. Let $\mathcal{P}_i \subset \mathbb{Z}^2$ denote the discrete spatial domain of the i-th visual patch token, where $i \in \{1, \ldots, N\}$. We compute the spatial coverage density γ_i by aggregating the global binary ROI mask $\mathbf{B}(u, v) = \bigvee_k m_k(u, v)$ over the patch, normalized by its cardinality $|\mathcal{P}_i|$:

$$\gamma_i = \frac{1}{|\mathcal{P}_i|} \sum_{(u,v)\in\mathcal{P}_i} \mathbf{B}(u, v). \tag{3}$$

To suppress out-of-distribution (OOD) structural collapse while extinguishing background noise, we formulate an activation soft-mask $\mathbf{m} \in \mathbb{R}^N$ governed by the threshold indicator $\mathbb{I}(\cdot)$:

$$m_i = \alpha \cdot \mathbb{I}(\gamma_i \geq \rho) + \beta \cdot (1 - \mathbb{I}(\gamma_i \geq \rho)), \tag{4}$$

where ρ is the density threshold, α serves as a gain factor about the ROI-aligned features, and $\beta \in (0, 1)$ acts as a severe L_2-norm attenuation factor. Implemented via a forward hook prior to the cross-modal projection layer, the recalibration is executed via broadcasting scalar multiplication:

$$\tilde{V} = \mathbf{Diag}(m)V, \quad \tilde{q} = q. \tag{5}$$

By shrinking the feature norm of non-ROI patches by a factor of β, their subsequent dot-product attention scores approach zero exponentially, effectively excising spurious visual pathways.

Joint Modulation Strategy. To achieve complementary regularization, the joint mapping simultaneously enforces L_2-norm visual filtering and discrete semantic anchoring:

$$\Phi_{\text{Joint}}: \quad \tilde{V} = \mathbf{Diag}(m)V, \quad \tilde{q} = [\mathcal{S}(\mathcal{R}) \parallel q \parallel \eta_{\text{task}}]. \tag{6}$$

2.3 Task-aware Dynamic Router

Since distinct clinical question categories exhibit varying sensitivities to visual vs. textual constraints, statically applying a single strategy yields sub-optimal

robustness. We therefore propose a zero-overhead dynamic router based on zero-shot LLM question classification at inference.

Given question q, a lightweight classifier yields its category $\hat{c} = h(q) \in \{\mathcal{C}_A, \mathcal{C}_M, \mathcal{C}_S, \mathcal{C}_R\}$, corresponding to Anatomy, Measurement, Symptom, and Radiology features, respectively. We define the routing strategy dynamically based on the query category: $s(\hat{c}) =$ **Vision-side** if $\hat{c} \in \{\mathcal{C}_A\}$, **Text-side** if $\hat{c} \in \{\mathcal{C}_M, \mathcal{C}_S\}$, and **Joint** if $\hat{c} \in \{\mathcal{C}_R\}$.

Rationale: Anatomy ($\mathcal{C}_A$) relies on spatial localization, naturally benefiting from vision-side L2-norm recalibration. Measurement ($\mathcal{C}_M$) and Symptoms ($\mathcal{C}_S$) demand precise semantic anchoring, making structured textual prompts superior. Radiology features ($\mathcal{C}_R$) typically require complex cross-modal reasoning, necessitating joint constraints on both visual activations and textual generation.

3 Experiments and Results

3.1 Datasets and Evaluation Metric

We evaluate hallucinations following the benchmark by Chang et al. [9]. For close-ended tasks, we report overall and per-type accuracy (anatomy, measurement, symptom, radiology) on MM-VisHal (comprising SLAKE [19] and VQA-RAD [20]) and CXR-VisHal (featuring IU-Xray [21] and MIMIC-CXR [22]). For open-ended evaluation, we assess hallucinations and text quality via CHAIR, CheXbert, RadGraph, RaTEScore, and Recall using 490 image-report pairs sampled from the MIMIC-CXR test set.

3.2 Results

On both close-ended and open-ended evaluation tasks, we systematically evaluate LLaVA-1.5-7B [23], LLaVA-Med-1.5-7B [24], Qwen3-VL-8B/32B [25], and InternVL-3.5-8B/38B [26]; the following sections present these results in turn.

Close-ended Evaluation. Table 1 shows that our training-free evidence injection consistently improves accuracy across all MLLMs. Notably, text-side injection dominates, achieving the highest overall Acc in 9 of 12 settings (with 2 ties) and an average gain of +1.9%, peaking at +3.3% for Qwen3-VL-8B on MM-VisHal. Conversely, the vision-side constraint exhibits a strong bias toward anatomy-related questions, delivering the best Acc-A in 10 of 12 settings. Finally, the joint strategy excels in radiology-feature questions, attaining the best Acc-R in 9 of 12 settings. The domain-specific benefits of interventions make dynamic routing essential for superior robustness.

As shown in Fig. 2, our task-aware Router consistently outperforms less robust baselines. Each module exhibits selective advantages for different questions, By dynamically exploiting these complementary strengths, routing achieves the highest overall accuracy improvements (+3.2% on MM-VisHal; +2.4% on CXR-VisHal), demonstrating superior robustness over any fixed intervention.

Table 1. Close-ended evaluation results on MM-VisHal and CXR-VisHal datasets. A, M, S, and R denote Acc-A, Acc-M, Acc-S, and Acc-R, respectively. (**Note:** Due to space constraints, MM-VisHal includes SLAKE and VQA-RAD, while CXR-VisHal comprises IU-Xray and MIMIC-CXR.)

MLLM	Method	MM-VisHal					CXR-VisHal				
		A	M	S	R	Acc	A	M	S	R	Acc
LLaVA-1.5-7B	Original	58.5	42.3	55.7	50.8	52.3	73.2	43.0	52.7	58.2	55.7
	Text-side	65.6	**44.1**	**57.3**	51.3	**55.0**	73.6	**45.2**	**55.9**	56.9	**57.9**
	Vision-side	**69.8**	42.9	54.7	52.7	**55.0**	**74.7**	38.6	54.3	58.3	56.5
	Joint	61.7	42.8	56.2	**52.8**	53.6	72.4	41.7	54.9	**59.1**	56.9
LLaVA-Med-1.5-7B	Original	65.9	38.2	51.1	53.2	51.7	85.3	46.3	70.7	85.3	71.9
	Text-side	65.9	**40.1**	**57.3**	52.8	**54.4**	86.2	48.9	**73.3**	86.6	**74.1**
	Vision-side	**66.9**	39.1	53.6	51.8	52.8	**86.5**	48.1	71.3	85.6	72.7
	Joint	65.3	37.9	55.0	**54.6**	53.2	85.2	**49.2**	71.5	**87.6**	72.9
Qwen3-VL-8B	Original	73.1	61.7	66.0	72.0	67.6	77.0	46.8	78.5	80.7	75.0
	Text-side	76.9	**65.1**	70.2	**72.7**	**70.9**	77.8	**50.4**	**79.7**	81.7	**76.3**
	Vision-side	**77.9**	63.1	65.7	72.2	69.0	**78.3**	48.0	78.3	82.3	75.4
	Joint	72.9	63.0	**70.5**	**72.7**	69.6	78.1	50.0	78.0	**83.6**	75.5
InternVL3.5-8B	Original	76.6	65.0	70.4	84.4	72.9	88.1	73.5	85.1	93.7	85.1
	Text-side	**80.2**	**66.7**	**73.6**	84.4	**75.3**	**88.5**	74.5	85.4	92.9	85.4
	Vision-side	78.1	65.2	70.5	**84.8**	73.4	88.4	73.5	**85.7**	94.1	**85.6**
	Joint	78.6	66.0	73.0	84.1	74.5	87.8	**74.8**	85.3	**94.5**	85.4
Qwen3-VL-32B	Original	78.6	68.3	66.6	74.7	71.1	79.0	60.7	81.9	79.2	78.8
	Text-side	79.5	**69.9**	**70.6**	**76.7**	**73.5**	78.8	**62.5**	**83.1**	76.7	**79.5**
	Vision-side	**80.3**	68.0	69.5	74.8	72.5	**79.7**	62.2	81.8	79.8	79.1
	Joint	77.9	67.4	70.5	75.9	72.4	79.3	61.9	82.3	**81.5**	**79.5**
InternVL3.5-38B	Original	79.1	69.4	70.8	81.2	74.1	80.4	71.4	86.3	91.7	84.2
	Text-side	79.8	70.2	**74.8**	81.0	**75.9**	80.9	**75.6**	**87.0**	91.6	**85.1**
	Vision-side	**81.0**	**70.5**	71.9	81.4	75.3	**81.2**	72.8	86.4	91.7	84.5
	Joint	80.7	67.3	**74.8**	**82.6**	75.6	81.1	72.5	86.5	**93.1**	84.7

Obs 1: Dynamic dual-side routing transcends the fragility of post-hoc corrections, consistently mitigating multi-source hallucinations.

Open-ended Evaluation. Fig. 3 shows that ROI-based training-free interventions reduce hallucinations with metric-dependent trade-offs: vision-side constraints most consistently lower CHAIR and improve clinical consistency, text-side injection better preserves coverage and fluency (higher RaTEScore/Recall), and the joint strategy often provides the best balance, yielding strong Recall with competitive CHAIR (notably on InternVL3.5).

Fig. 4 highlights the trade-offs of our ROI-based interventions against baselines. While OPERA yields the highest Recall (19.13%) and PAI excels in clinical metrics (CheXbert 23.13%, RadGraph 9.09%), both severely increase hallucinations. Our vision-side intervention achieves the lowest CHAIR (12.76%) and improves clinical consistency by curbing spurious visual findings, albeit at a slight cost to Recall. Meanwhile, our text-side intervention maintains Recall and boosts

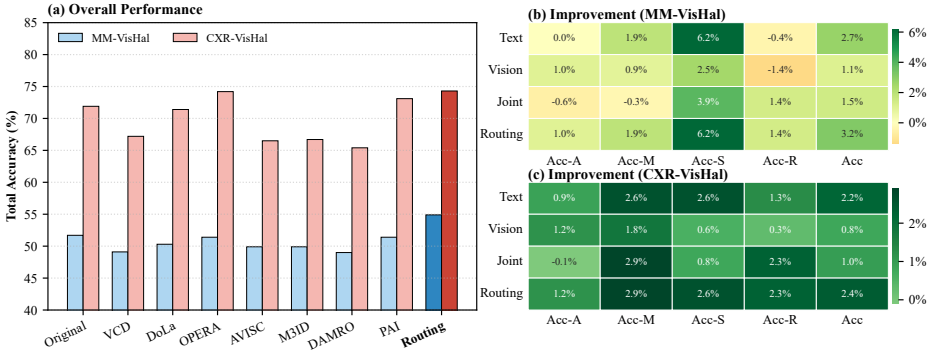

Fig. 2. Evaluation of mitigation strategies. (a) Overall accuracy comparison on MM-VisHal and CXR-VisHal datasets. (b)-(c) Absolute metric improvements of individual components and our Routing over the LLaVA-Med-1.5-7B baseline.

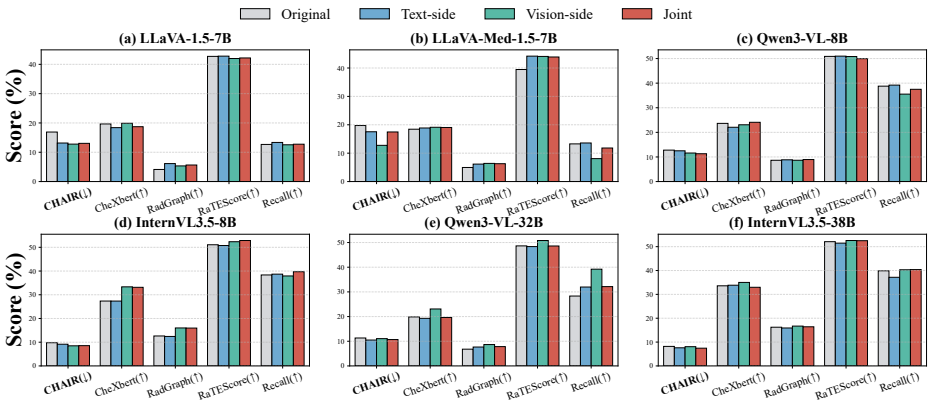

Fig. 3. Open-ended evaluation results on the MIMIC-CXR for report generation. Subfigures (a) to (f) illustrate the evaluation results across six different MLLMs.

RaTEScore to 44.18%, proving structured textual injection enhances narrative quality and coherence while effectively mitigating hallucinations.

Obs 2: The central challenge in open-ended evaluation remains the difficulty of balancing report completeness and hallucination suppression.

3.3 Ablation Study and Qualitative Analysis

Ablation Study. Fig. 5 illustrates the ablation studies for both interventions. **(1) Text-side:** Random bbox degrade performance across all benchmarks, proving that text-side improvements rely on the injected spatial grounding of ROI coordinates. **(2) Vision-side:** In vision-side weighting, $(\alpha = 1.0, \beta = 0.1)$ is the optimal setting. Over-amplifying ROI features (α) provides no gain, while increasing non-ROI weights (β) consistently introduces distracting signals.

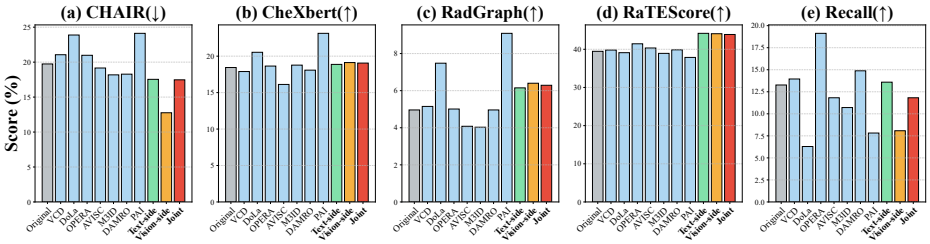

Fig. 4. Open-ended mitigation comparison on MIMIC-CXR using LLaVA-Med-1.5-7B.

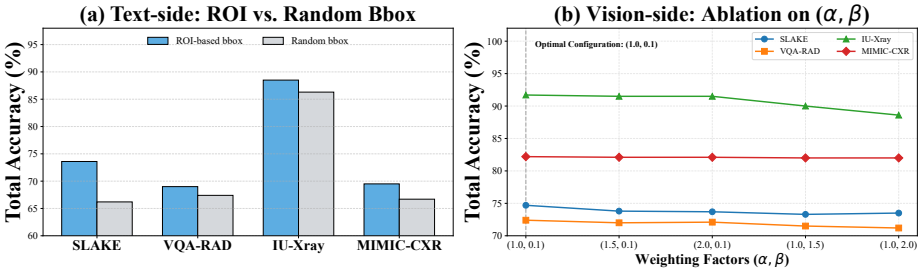

Fig. 5. Controlled comparisons for text-side and vision-side methods. (a) Performance comparison between ROI-based and random bounding boxes for Qwen3-VL-8B. (b) Ablation study of weighting factors (α, β) for InternVL3.5-8B.

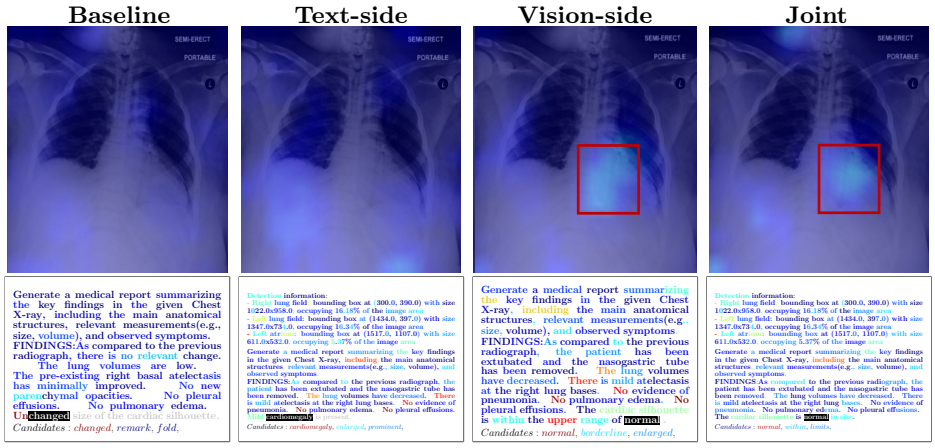

Fig. 6. Token Activation Map (TAM) for visualization presents high-quality localization results on MIMIC-CXR datasets under the InternVL3.5-8B.

Qualitative Analysis. Fig. 6 provide TAM visualizations to elucidate how our interventions steer the grounding mechanism. The baseline's diffuse attention during the generation of *"changed"* is sharpened by vision-side injection, which

refocuses the model on the left atrial region—a key semantic anchor for the ground-truth *"normal"* heart size.

4 Conclusion

We address visual misinterpretation hallucinations in medical MLLMs with a training-free, evidence-injection framework. Using MedSAM-extracted ROI priors as verifiable anatomical evidence, our method improves visual grounding and clinical consistency through text-side evidence injection, vision-side activation recalibration, and their joint strategy. We further introduce a lightweight decoupled router that dynamically selects modality-specific interventions based on task semantics. Extensive evaluations on multiple MLLMs across close-ended and open-ended tasks, supported by controlled ablations and visualizations, show robust hallucination mitigation over existing baselines while improving reasoning accuracy and factual reliability. Overall, this work provides a practical path toward visually grounded and controllable generation in clinical workflows.

Acknowledgments. The work was supported by the National Key R&D Program of China under Grant 2025YFE0216500, the Major Program of National Natural Science Foundation of China under Grant 62595802, the Foundation for Outstanding Research Groups of Hubei Province of China under Grant 2025AFA012, the 111 Project on Computational Intelligence and Intelligent Control under Grant B18024, and the National Research Foundation, Prime Minister's Office, Singapore under its IN-CYPHER Campus for Research Excellence and Technological Enterprise (CREATE) Programme.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., et al.: CheXagent: Towards a foundation model for chest X-ray interpretation. In: AAAI 2024 Spring Symposium Series (Clinical Foundation Models). OpenReview.net (2024). https://openreview.net/forum?id=P3LOmrZWGR
2. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-Flamingo: A multimodal medical few-shot learner. In: Proceedings of the 3rd Machine Learning for Health Symposium (ML4H 2023). Proceedings of Machine Learning Research, vol. 225, pp. 353–367. PMLR (2023)
3. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.: XrayGPT: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing (BioNLP 2024), pp. 440–448. Association for Computational Linguistics (2024)
4. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications **16**, 7866 (2025)

5. Lee, S., Kim, W.J., Chang, J., Ye, J.C.: LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation. In: International Conference on Learning Representations (ICLR) (Poster) (2024). OpenReview.net. https://openreview.net/forum?id=BqHaLnans2
6. Jing, L., Zhang, Y., Du, X.: Tutorial proposal: Hallucinations in large language models and large vision-language models. In: Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR), pp. 2138–2139 (2025)
7. Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., Zhang, H.: Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. Artificial Intelligence Review **57**(9) (2024)
8. Tu, C., Ye, P., Zhou, D., Bai, L., Yu, G., Chen, T., Ouyang, W.: Attention reallocation: Towards zero-cost and controllable hallucination mitigation of MLLMs. International Journal of Computer Vision **134**(1) (2025)
9. Chang, A., Huang, L., Bhatia, P., Kass-Hout, T., Ma, F., Xiao, C.: MedHEval: Benchmarking hallucinations and mitigation strategies in medical large vision-language models. arXiv:2503.02157 (2025)
10. Zhu, Z., Zhang, Y., Zhuang, X., Zhang, F., Wan, Z., Chen, Y., Long, Q., Zheng, Y., Wu, X.: Can we trust AI doctors? A survey of medical hallucination in large language and large vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 6748–6769. Association for Computational Linguistics (2025)
11. Asgari, E., Montaña-Brown, N., Dubois, M., et al.: A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine **8**, 274 (2025)
12. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13872–13882 (2024)
13. Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., He, P.: DoLa: Decoding by contrasting layers improves factuality in large language models. arXiv:2309.03883 (2023)
14. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13418–13427 (2024)
15. Woo, S., Kim, D., Jang, J., Choi, Y., Kim, C.: Don't miss the forest for the trees: Attentional vision calibration for large vision language models. arXiv:2405.17820 (2024)
16. Gong, X., Ming, T., Wang, X., Wei, Z.: DAMRO: Dive into the attention mechanism of LVLM to reduce object hallucination. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7696–7712 (2024)
17. Liu, S., Zheng, K., Chen, W.: Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In: European Conference on Computer Vision (ECCV), pp. 125–140. Springer (2025)
18. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14303–14312 (2024)

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: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1650–1654 (2021)
20. 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)
21. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
22. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-Y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 26296–26306 (2024)
24. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J.-W., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems, vol. 36, pp. 28541–28564. Curran Associates, Inc. (2023)
25. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
26. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)