Counterfactual Anatomy-guided Spatial-Temporal Decoding for Annotation-Free Hallucination Mitigation in Medical VLMs

Yifan Lu, Adinath Dukre, Abhijit Das, Ziyun Zou, Haolin Yang, Yutong Xie, and Imran Razzak✉

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
MedOS, Abu Dhabi, UAE
Imran.Razzak@mbzuai.ac.ae

Abstract. Medical vision-language models (Med-VLMs) have demonstrated strong performance on medical visual question answering, yet they remain prone to hallucination, generating clinically unsupported statements that are insufficiently grounded in image evidence. Mitigation methods applied during decoding offer a practical solution, but they typically lack anatomical awareness or rely heavily on ground truth annotations, which limits their applicability. We propose Counterfactual Anatomy-guided Spatial-Temporal decoding (CAST), a framework that operates entirely during inference and requires no manual annotations for anatomically grounded hallucination mitigation. CAST automatically discovers anatomical regions relevant to the given query through broad medical segmentation. It then selects a compact, causally informative area using counterfactual intervention based on the drop in answer likelihood under occlusion. Guided by this chosen region, CAST performs a unified contrastive decoding process, combining classifier-free guidance to correct spatial attention with stepwise temporal contrast to regulate generation dynamics. Experiments on the SLAKE and MIMIC-CXR datasets across three Med-VLMs demonstrate that CAST consistently outperforms strong baselines and surpasses decoding strategies reliant on ground truth. Our results indicate that compact, automatically selected regions provide highly effective contrastive guidance without expert annotations, offering a practical and generalizable solution for improving spatial grounding and reducing hallucinations.

Keywords: Medical Vision-Language Models · Hallucination Mitigation · Contrastive Decoding · Causal Discovery

1 Introduction

Medical vision-language models (Med-VLMs) have recently demonstrated impressive performance on medical visual question answering (Med-VQA) by jointly reasoning over images and clinical language [5, 24, 11]. Despite their strong average accuracy, these models remain prone to hallucination, generating fluent

but clinically unsupported statements that lack sufficient grounding in the image [14, 20, 6]. In medical settings, such errors can have serious consequences, as a single hallucinated finding may mislead clinical interpretation or subsequent decision-making [19, 3, 26]. Improving faithfulness without expensive retraining has therefore become a central challenge for the reliable deployment of Med-VLMs [17, 27, 23].

Decoding-time methods are attractive because they are plug-and-play and do not require additional supervision [8, 7, 10]. Visual Contrastive Decoding (VCD) [15] mitigates hallucinations by contrasting predictions from the original image and a noise-perturbed version, encouraging tokens that remain stable under visual corruption. Decoding by Contrasting Layers (DoLA) [4] improves factuality by contrasting logits from early and late transformer layers, suppressing tokens driven by memorized knowledge rather than contextual evidence. OPERA [9] penalizes over-confident attention patterns and reallocates attention during generation to reduce over-trust in spurious visual cues. Although effective for general-purpose VLMs [13], these methods are largely anatomy-agnostic and apply global corrections that do not explicitly account for the localized nature of medical evidence. Anatomical Region-Guided Contrastive Decoding (ARCD) [18] addresses this limitation by introducing spatial grounding through classifier-free guidance (CFG) conditioned on an anatomical region-of-interest (ROI). At each decoding step, ARCD contrasts a full-image branch with an ROI-suppressed branch, steering generation toward tokens supported by the diagnostic region. However, ARCD depends on ground-truth (GT) ROI annotations, which are costly, scarce, and dataset-specific, limiting real-world applicability.

We investigate whether region-guided contrastive decoding can be made annotation-free without sacrificing effectiveness by automatically discovering question-relevant ROIs for guidance. Empirically, we find that such automatically selected regions not only remove the dependency on expert annotations but can also provide stronger guidance in practice. This is partly because the effectiveness of ROI-guided CFG depends on mask granularity: ground-truth bounding boxes in Med-VQA benchmarks are often coarse and high-coverage, and suppressing them removes most visual information from the unconditional branch, weakening the region-specific contrast. In contrast, compact and precise masks preserve contextual cues while selectively removing relevant evidence, yielding a cleaner contrastive signal during decoding.

Motivated by this observation, we propose **C**ounterfactual **A**natomy-guided **S**patial-**T**emporal (**CAST**) decoding, an annotation-free framework for hallucination mitigation in Med-VLMs. CAST consists of three stages. First, in the proposal stage, we generate diverse anatomical region candidates using a medical segmentation model through efficient multi-concept inference. Second, in the selection stage, we identify the most question-relevant and CFG-compatible region via a counterfactual intervention criterion that measures the likelihood drop under regional occlusion while discouraging overly large masks. Third, in the decoding stage, we introduce a spatial-temporal contrastive strategy that corrects where the model attends via anatomy-guided CFG and regulates what

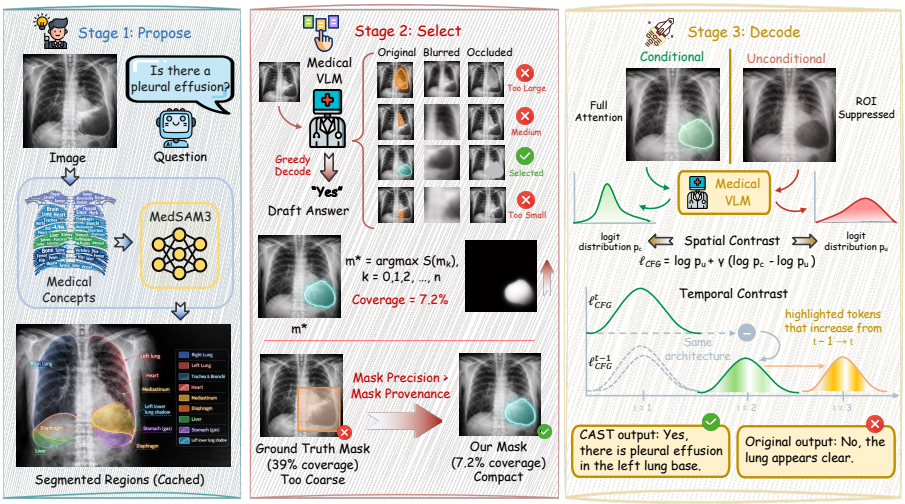

Fig. 1. Overview of the proposed CAST framework. It (1) proposes anatomical regions via multi-concept medical segmentation, (2) selects a compact question-relevant ROI using counterfactual likelihood drop under occlusion, and (3) applies spatial-temporal contrastive decoding: anatomy-guided CFG for grounding and step-wise temporal contrast to suppress hallucinations.

it predicts via step-wise temporal contrast. The entire pipeline operates at inference time and remains fully plug-and-play. Across three Med-VLMs on SLAKE and MIMIC-CXR, CAST outperforms strong decoding-time baselines and GT-dependent region-guided methods, achieving effective spatial grounding without expert annotations.

2 Method

CAST is an annotation-free, inference-time framework for mitigating hallucinations in Med-VLMs via automatic ROI discovery and spatial-temporal contrastive decoding. As illustrated in Fig. 1, we first generate candidate anatomical masks with multi-concept medical segmentation and filter them for CFG compatibility. Given a question, we select the ROI that most affects the model's likelihood under regional occlusion while favoring compactness. Finally, CAST performs anatomy-guided CFG for region grounding and a step-wise temporal contrast to damp language-driven momentum. The entire pipeline is fully plug-and-play, requiring no retraining or expert annotations.

2.1 Zero-Shot Anatomical Region Proposal

The first stage of CAST aims to extract a diverse set of anatomically coherent region candidates directly from the input image $\mathbf{x}$. To achieve this with-

out dataset-specific annotations, we leverage the robust medical priors embedded in a text-guided medical segmentation model, MedSAM3 [21]. Rather than sequentially querying individual anatomical structures, we employ an efficient multi-concept inference strategy. We construct a comprehensive vocabulary of anatomical concepts (spanning thoracic, abdominal, musculoskeletal, and vascular systems) and encode them into a unified prompt. The model's DETR-style [1] decoder processes this joint embedding to activate diverse parallel queries, generating segmentation masks for multiple structures in a single forward pass. To ensure the proposals are suitable for localized decoding, the raw masks undergo a compactness-aware filtering pipeline. Masks are filtered by area coverage to remove pixel-level noise and excessively global masks. We then apply spatial non-maximum suppression based on intersection over union to eliminate near-duplicates [25], followed by morphological filtering to discard degenerate shapes. This results in a spatially diverse and anatomically valid set of K candidate regions $[m_1, \ldots, m_K]$, extracted with minimal computational overhead.

2.2 Counterfactual Region Selection

Given the K candidates, CAST must autonomously identify the single region m^* that is most causally responsible for the model's clinical reasoning regarding the current query q. We frame this as a causal discovery problem, evaluating each region via counterfactual intervention. We first generate a draft sequence $\hat{y} = (y_1, \ldots, y_T)$ utilizing greedy decoding on the factual, unoccluded image $\mathbf{x}$. To isolate the causal contribution of a candidate region m_k, we define a counterfactual intervention do(occlude m_k) by selectively corrupting the pixel space strictly within the mask boundaries. To prevent the model from over-indexing on occlusion artifacts, we utilize two complementary operators, $o \in \{\text{blur}, \text{mean}\}$, which independently disrupt fine-grained texture and local contrast statistics. The causal effect (CE) of region m_k is quantified as the drop in the predictive log-likelihood of the draft answer when the region is removed:

$$\text{CE}_o(m_k) = \log p(\hat{y} \mid \mathbf{x}, q) - \log p(\hat{y} \mid \mathbf{x}_{\text{occ}}^{(o,k)}, q), \tag{1}$$

where $\mathbf{x}_{\text{occ}}^{(o,k)}$ represents the image under occlusion mode o. Effective contrastive guidance requires spatial precision; a region that spans the entire image provides no localized contrastive signal. Therefore, our selection objective explicitly balances causal relevance with mask compactness and cross-modal consistency:

$$S(m_k) = \frac{1}{2} \sum_o \text{CE}_o(m_k) - \lambda \cdot \text{area}(m_k) - \mu \cdot \text{Var}\big[\text{CE}_{\text{blur}}(m_k), \text{CE}_{\text{mean}}(m_k)\big]. \tag{2}$$

Here, the area penalty λ enforces mask compactness for localized contrast, while μ penalizes variance to filter out occlusion-specific artifacts. The highest-scoring mask $m^* = \arg\max_k S(m_k)$ is then selected as the optimal diagnostic anchor.

2.3 Unified Spatial-Temporal Contrastive Decoding

With the optimal region m^* identified, CAST introduces a lightweight, sequential decoding strategy to simultaneously correct two orthogonal hallucination axes: spatial grounding (*where* the model attends) and temporal generation (*what* it predicts). At each decoding step t, a dual-branch forward pass is constructed. The unconditional branch heavily suppresses attention to visual tokens within m^*, forcing reliance on surrounding context and language priors, while the conditional branch processes the full image natively. Spatially guided logits are computed as:

$$\boldsymbol{\ell}_{\text{CFG}}^{(t)} = \log p_u^{(t)} + \gamma \cdot \left(\log p_c^{(t)} - \log p_u^{(t)}\right), \tag{3}$$

where γ is the guidance scale, and p_c and p_u are conditional and unconditional token probabilities. This amplifies tokens dependent on the targeted anatomical region, reducing reliance on spurious background correlations. Following spatial grounding, CAST applies an orthogonal temporal correction across consecutive steps:

$$\boldsymbol{\ell}_{\text{final}}^{(t)} = w_m \cdot \boldsymbol{\ell}_{\text{CFG}}^{(t)} - w_p \cdot \boldsymbol{\ell}_{\text{CFG}}^{(t-1)}, \tag{4}$$

where w_m and w_p are weighting hyper-parameters. The theoretical foundation for this temporal contrast is that the prior step's distribution, $\boldsymbol{\ell}_{\text{CFG}}^{(t-1)}$, heavily encodes the default trajectory driven by the model's internal language momentum. By subtracting this prior distribution from the current active prediction $\boldsymbol{\ell}_{\text{CFG}}^{(t)}$, CAST selectively amplifies tokens that are newly activated by the current visual evidence, while suppressing tokens that are merely passively inherited from previous steps. Because the spatial contrast modulates the source of visual attention and the temporal contrast regulates habitual token transitions, the two mechanisms operate harmoniously without interference.

3 Experiments

3.1 Experimental Setup

Datasets and Baselines. We evaluate on SLAKE [22] and MIMIC-CXR [12]. SLAKE contains 1,061 multi-modal (CT/MRI/X-ray) QA pairs with open- and closed-ended questions and bounding boxes. MIMIC-CXR includes 2,134 chest X-ray QA pairs, primarily closed-ended with pathology annotations. Following prior work [18], we report Accuracy (closed), Recall (open), and weighted Overall performance. CAST is compared against the standard decoding baseline and four decoding-time methods: VCD [15], DoLA [4], OPERA [9], and ARCD [18].

Implementation Details. To demonstrate the plug-and-play versatility of CAST, we apply it to three Med-VLMs: Phi-3.5V-Med [18], HuatuoGPT-Vision-7B [2], and LLaVA-Med-7B [16]. MedSAM3 processes images at a resolution of 1008×1008. We apply NMS with an IoU threshold of 0.9 to ensure spatial diversity, retaining up to K=16 candidate regions per question for counterfactual scoring. For the selection objective, both the area penalty factor λ and the

Table 1. Comprehensive comparison on SLAKE and MIMIC benchmarks. Best in **bold**, second underlined within each model block.

	SLAKE			MIMIC		
Method	Open	Closed	Overall	Open	Closed	Overall
Phi-3.5V-Med	38.58	61.42	50.00	14.29	58.12	47.24
+ VCD	40.16	61.42	50.79	15.87	59.16	48.43
+ DoLA	**47.24**	59.84	53.54	19.05	56.54	47.24
+ OPERA	36.22	65.35	50.79	14.29	61.78	50.00
+ ARCD	44.54	65.87	52.90	28.85	60.44	52.61
+ **CAST (Ours)**	44.90	**81.49**	**59.25**	**33.03**	**67.85**	**59.22**
HuatuoGPT-V-7B	47.94	75.48	58.74	29.84	62.68	54.54
+ VCD	**49.73**	75.48	59.83	31.05	63.61	55.54
+ DoLA	43.97	78.61	57.55	29.64	65.30	56.46
+ OPERA	46.32	74.04	57.19	**34.74**	63.99	56.74
+ ARCD	48.33	75.00	58.79	30.97	63.12	55.15
+ **CAST (Ours)**	47.56	**81.49**	**60.86**	31.31	**67.29**	**58.37**
LLaVA-Med-7B	39.62	62.50	48.59	22.52	53.71	45.98
+ VCD	40.00	62.98	49.01	22.77	53.71	46.04
+ DoLA	32.75	66.35	45.92	21.34	56.14	47.51
+ OPERA	**40.59**	65.38	50.31	26.17	54.52	47.49
+ ARCD	37.95	62.02	47.39	22.88	53.77	46.11
+ **CAST (Ours)**	36.49	**73.32**	**50.93**	**27.16**	**59.69**	**51.63**

consistency coefficient μ are set to 0.1. In the decoding stage, the spatial CFG employs a guidance scale of γ=1.5. The attention modulation weights for regions inside and outside the ROI (α and β, respectively) are empirically determined via grid search over the set $\{0.01, 0.1, 0.5, 1.0, 2.0, 3.0\}$. For the step-wise temporal contrast, we set the active momentum weight w_m=1.0 and the prior-step penalty w_p=0.5. All experiments are executed on a single NVIDIA A100 40GB GPU.

3.2 Experimental Results and Analysis

Comparison with Decoding-Time Methods. Table 1 reports the comprehensive comparison across three Med-VLMs on SLAKE and MIMIC-CXR. CAST consistently improves overall performance over strong decoding-time baselines (VCD, DoLA, OPERA) and GT-dependent region-guided decoding across all models. Most notably, CAST demonstrates substantial and consistent improvements in Closed-ended question accuracy, which directly reflects a model's ability to remain factually grounded and resist hallucination. For instance, when applied to Phi-3.5V-Med on SLAKE, CAST elevates closed-ended accuracy from 61.42 to a striking 81.49. On MIMIC, CAST improves both open and closed metrics simultaneously, indicating that spatial-temporal contrast benefits both factual grounding and short-form diagnostic responses. While anatomy-agnostic baselines such as DoLA, OPERA, and VCD occasionally yield isolated gains on Open-ended queries for specific models, they frequently exhibit volatile performance, trading closed-ended accuracy for open-ended recall. In contrast, CAST maintains highly competitive open-ended performance while decisively dominating closed-ended factual reasoning. Notably, CAST consistently outperforms

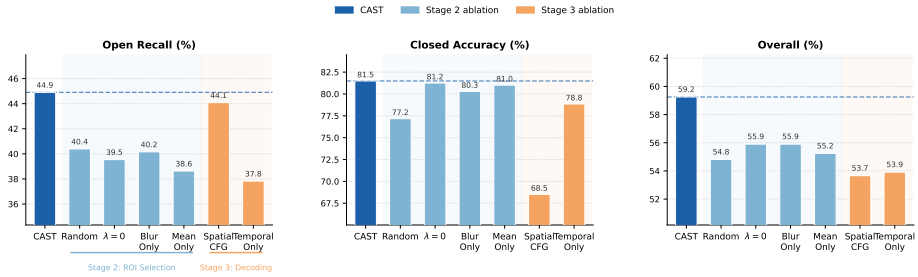

Fig. 2. Ablation study on SLAKE (Phi-3.5V-Med).

ARCD, this empirical superiority strongly validates the core theoretical hypothesis introduced in our methodology: automatically discovered, dynamically selected, and highly compact region masks provide a cleaner, more targeted contrastive signal than coarse GT annotations. By explicitly optimizing for causal relevance and spatial compactness, CAST achieves superior spatial grounding and hallucination mitigation entirely free of expert annotation dependency.

Ablation Study. Fig. 2 ablates the core components of CAST on SLAKE, demonstrating that both our unified decoding mechanisms and counterfactual selection criteria are essential. Isolating spatial CFG or temporal contrast drops overall accuracy from 59.25% to 53.66% and 53.91%, respectively, confirming their orthogonal roles: spatial CFG anchors where the model attends to rescue closed-question accuracy (from 68.51% to 81.49%), while temporal contrast regulates what it predicts to sustain open-question recall (44.90% vs. 37.83%). Beyond decoding, counterfactual region selection provides a critical signal; replacing it with random mask selection degrades overall accuracy by 4.4%. Within the scoring function itself, removing the area penalty (λ=0) causes a severe 3.4% drop, cementing that mask compactness is strictly required for effective CFG. Finally, relying on a single occlusion mode (blur or mean) lowers accuracy by 3.3% and 4.0%, highlighting the necessity of combining dual occlusions to average out modality-specific artifacts.

Qualitative Analysis. Fig. 3 shows two failure cases where CAST overcomes limitations of coarse GT masks. In Case 1 (MIMIC X-ray), GT-ARCD fails to detect a subtle Swan-Ganz catheter because its bounding box spans the entire right lung, diluting the contrastive signal. CAST instead isolates a compact 10.6% region around the catheter tip and correctly confirms its presence. In Case 2 (SLAKE MRI), the model must identify the anatomical region (head) rather than the highly salient organ (brain). Both the Baseline and GT-ARCD incorrectly output "brain" because the broad GT segmentation reinforces organ identity over spatial location, whereas CAST selects a highly compact mask highlighting specific craniofacial landmarks to correctly guide the response. Together, these cases concretely demonstrate that dynamically selected, compact masks provide superior contrastive guidance compared to static GT annotations, which can actively mislead or dilute critical visual evidence.

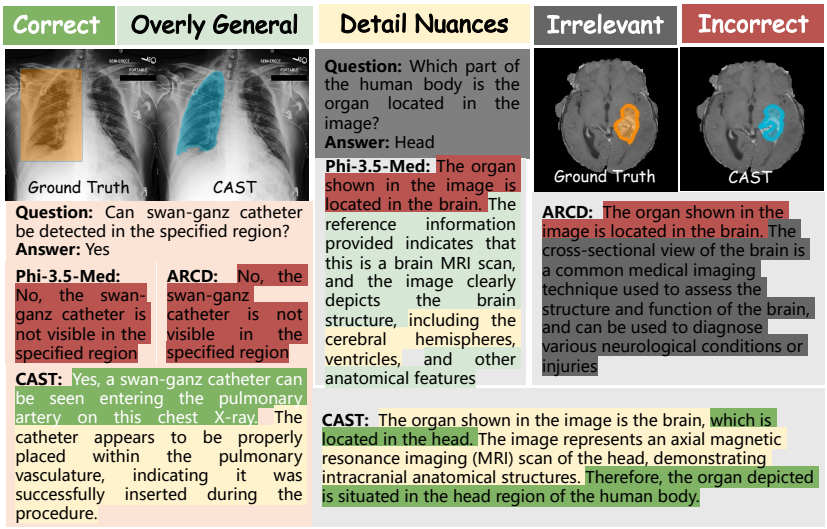

Fig. 3. Qualitative comparison of model-generated responses.

Mask Compactness Analysis. To understand CAST's advantage over GT-based methods, we analyze mask coverage and performance on HuatuoGPT-V-7B (SLAKE). GT bounding boxes are substantially larger than CAST-selected masks (39.41% vs. 7.05% mean coverage), with some spanning the entire image. When GT coverage exceeds 40%, closed-question accuracy drops noticeably. Per-sample analysis shows that CAST mainly improves cases with large GT masks, while GT performs competitively only when annotations are compact. Notably, GT slightly reduces closed accuracy compared to NULL decoding, whereas CAST significantly improves it. These results indicate that oversized masks dilute the localized contrast required by classifier-free guidance, while compact, causally selected regions provide sharper and more effective signals.

Computational Efficiency. As a plug-and-play inference-time framework, CAST introduces minimal overhead compared to standard decoding method. In Stage 1, MedSAM3 proposals are precomputed and cached, effectively eliminating per-question segmentation costs; even in an online setting, this overhead remains marginal relative to VLM decoding. In Stage 2, counterfactual ROI scoring operates via a lightweight teacher-forced evaluation over a short sequence ($T_{score} = 32$ tokens), which is roughly an order of magnitude cheaper than a full autoregressive generation pass. Although Stage 3 employs dual-branch spatial-temporal contrast, the overall wall-clock latency of CAST remains nearly identical to that of DoLA, placing it squarely within the practical operating range of accepted inference-time methods. Furthermore, the K candidate regions are independent and can be seamlessly parallelized at the engineering level for further efficiency gains.

4 Conclusion

In this paper, we presented CAST, a plug-and-play decoding framework for mitigating hallucinations in Med-VLMs. By formulating spatial grounding as a causal discovery problem, CAST automatically identifies compact and question-relevant anatomical regions through counterfactual intervention. We show that these dynamically selected regions address the compactness limitations of coarse ground-truth annotations and provide a sharper and more informative contrastive signal. Combined with a unified spatial-temporal decoding strategy that jointly refines visual attention and step-wise generation dynamics, CAST consistently achieves state-of-the-art performance across multiple Med-VLMs on SLAKE and MIMIC-CXR. Overall, CAST offers a practical inference-time approach to improving factual reliability without requiring expert annotations or model retraining.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7346–7370 (2024)
3. Chen, Y., Yang, Z., Huang, Y., Yang, X., Yeo, S., Zhang, Y.: Trustworthy disentangled framework for multi-label medical image classification with multimodal refinement. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 839–846. IEEE (2025)
4. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. In: International Conference on Learning Representations. vol. 2024, pp. 54158–54183 (2024)
5. Deria, A., Kumar, K., Dukre, A.M., Segal, E., Khan, S., Razzak, I.: Medmo: Grounding and understanding multimodal large language model for medical images. arXiv preprint arXiv:2602.06965 (2026)
6. Dutta, N., Bose, K., Syailendra, E., Chu, L., Gupta, P.: Vision-language models in diagnostic imaging: review of technical advances, clinical validation, and practical deployment. International Journal of Medical Informatics p. 106227 (2025)
7. 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. pp. 14303–14312 (2024)
8. Ghosh, S., Evuru, C.K., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., Manocha, D.: Visual description grounding reduces hallucinations and boosts reasoning in lvlms. In: International Conference on Learning Representations. vol. 2025, pp. 66510–66547 (2025)

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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
10. Ji, Y., Zhang, J., Xia, H., Chen, J., Shou, L., Chen, G., Li, H.: Specvlm: Enhancing speculative decoding of video llms via verifier-guided token pruning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 7216–7230 (2025)
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. Johnson, A.E., 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)
13. Kang, J., Shu, H., Li, W., Zhai, Y., Chen, X.: Vispec: Accelerating vision-language models with vision-aware speculative decoding. Advances in Neural Information Processing Systems **38**, 115511–115532 (2026)
14. Khanal, B., Pokhrel, S., Bhandari, S., Rana, R., Shrestha, N., Gurung, R.B., Linte, C., Watson, A., Shrestha, Y.R., Bhattarai, B.: Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 235–245. Springer (2025)
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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 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. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
17. Liang, X., Hu, J., Wang, D., Ma, Z., Zhao, L., Li, R., Wan, B., Wang, Q.: Chexpo: Preference optimization for chest x-ray vlms with counterfactual rationale. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2606–2615 (2025)
18. Liang, X., Liu, C., Ma, Z., Wang, D., Jing, B., Wang, Q., Shi, Y.: Anatomical region-guided contrastive decoding: A plug-and-play strategy for mitigating hallucinations in medical vlms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6871–6879 (2026)
19. Liang, X., Wang, D., Jiao, Z., Li, R., Yang, P., Wang, Q., Chua, T.S.: Uncertainty-driven expert control: Enhancing the reliability of medical vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21144–21154 (2025)
20. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 669–679. Springer (2025)
21. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)

22. 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). pp. 1650–1654. IEEE (2021)
23. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14788–14798 (2025)
24. Shu, Y., Liu, C., Chen, R., Li, D., Dai, B.: Fleming-vl: Towards universal medical visual reasoning with multimodal llms. arXiv preprint arXiv:2511.00916 (2025)
25. Si, K.S., Sun, L., Zhang, W., Gong, T., Wang, J., Liu, J., Sun, H.: Accelerating non-maximum suppression: a graph theory perspective. Advances in Neural Information Processing Systems **37**, 121992–122028 (2024)
26. Silva-Rodríguez, J., Ben Ayed, I., Dolz, J.: Trustworthy few-shot transfer of medical vlms through split conformal prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 658–668. Springer (2025)
27. Zhao, Y., Zhong, E., Yuan, C., Li, Y., Zhao, M., Li, C., Hu, J., Liu, W., Liu, C.: Med-vlm: Enhancing medical image segmentation accuracy through vision-language model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7283–7293 (2025)