Proximity-Constrained Counterfactual Decoding for Hallucination-Robust Medical VQA

Dwarikanath Mahapatra[1,6⋆], Abhijit Das[2⋆], Manish Pandey[3], Joy Dhar[4], Sudipta Roy[5,9], Zongyuan Ge[6], Behzad Bozorgtabar[7], and Imran Razzak[2,8]

[1]Khalifa University, UAE [2]MBZUAI, UAE [3]RoentGen Health, India [4]IIT Ropar, India [5]Jio Institute, India [6]Monash University, Australia [7]Aarhus University, Denmark [8]MedOS, UAE [9]UNSW Bengaluru, India

dwarikanath.mahapatra@ku.ac.ae, {abhijit.das, imran.razzak}@mbzuai.ac.ae

⋆Equal contribution.

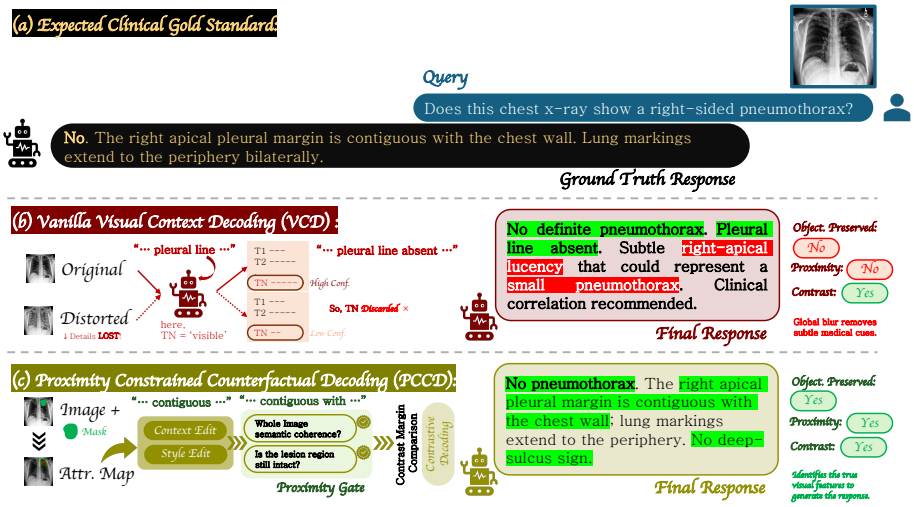

Fig. 1: **Proximity-constrained counterfactual decoding for medical VQA.** (a) Ground truth: no pneumothorax. (b) VCD's global blur erases subtle apical cues, producing a hallucinated response. (c) PCCD preserves the lesion via attribution-guided masking, gates edits through dual proximity checks, and selects the higher-margin branch yielding an evidence-aligned answer.

Abstract. Hallucinations in medical vision language models arise when language priors override subtle visual evidence during decoding. Existing training-free contrastive decoding methods suppress these priors by contrasting logits against a globally perturbed image, but global perturbations destroy the very diagnostic cues they are meant to protect —standard VCD blur flips nearly one in five CheXpert labels on held-out chest

radiographs. We propose **Proximity-Constrained Counterfactual Decoding (PCCD)**, which admits a counterfactual view only when it satisfies both a global and an object-masked feature-space similarity bound, constraining diagnostic drift while increasing the contrastive margin for visually grounded tokens to first order under explicitly stated assumptions. Two complementary branches address heterogeneous hallucination drivers: *Object-Aware VCD* preserves attribution-indicated lesion regions while attenuating contextual co-occurrence priors, and *Latent-Disentangled VCD* applies structure-preserving style edits to suppress attribute biases. When neither branch passes the proximity gate, PCCD falls back to greedy decoding, guaranteeing it never degrades below the unmitigated baseline. Across radiology VQA, chest X-ray report generation, and a new CXR existence probe (Med-POPE), PCCD improves GREEN AUC, QAAS, and RadGraph F1 by up to $\approx +2.5$ pp over SOTA with one auxiliary forward pass per token and no weight updates.

Keywords: Medical VLMs · Hallucination Robustness · Contrastive Decoding · Counterfactual Reasoning · Hallucination Mitigation

1 Introduction

Large vision–language models (VLMs) achieve strong performance on radiology reporting and medical VQA [17,5,7], yet they remain prone to *visual hallucinations*: spurious findings, incorrect attributes, or unsupported structures that directly affect diagnostic interpretation and patient safety [18,23,13]. Existing training-free approaches such as Visual Contrastive Decoding (VCD) [16] reduces hallucinations by contrasting token logits against a perturbed view, with variants such as HALC [4], VASparse [27], and Octopus [21] validating this paradigm for natural images.

Medical imaging differs fundamentally: diagnostically decisive cues appear as faint, localised structures (subtle apical lucencies, catheter tips, mucosal textures) that global perturbations inadvertently erase. We quantify this directly on 200 held-out CXRs and 200 GI frames: standard VCD blur (k=15) reduces macro CheXpert label agreement to 83.1%, with pneumothorax agreement falling to 0.70 — nearly one diagnostic label flip in five. Critically, silent drift is not unique to global noise: attention-reweighted views without proximity constraints alter at least one diagnostic label in 14.2% of cases, while PCCD-guarded views maintain >98% agreement across all findings. These measurements identify three necessary conditions for reliable mitigation: **(C1)** perturbations must preserve localised lesion evidence; **(C2)** existence and attribute errors stem from distinct drivers — co-occurrence priors vs. style biases — requiring separate interventions; **(C3)** counterfactual views must be explicitly constrained to prevent representational drift. No existing training-free decoder satisfies all three.

To bridge this gap, we propose **Proximity-Constrained Counterfactual Decoding (PCCD)**, a training-free framework built on one principle: a counterfactual view influences decoding only if it is *semantically admissible* — satisfying both a global similarity bound $s(v, v') \geq \tau$ and an object-masked bound

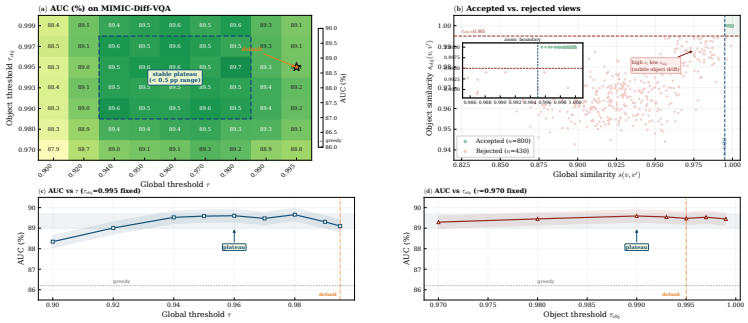

Fig. 2: **Proximity guard sensitivity.** AUC and CheXpert label agreement vs. τ and τ_{obj} on MIMIC-Diff-VQA. Performance sits on stable plateau $[0.990, 0.998]$; below this range semantic drift increases, above it greedy fallback dominates.

$s_{\text{obj}}(v, v') \geq \tau_{\text{obj}}$ in the VLM representation space. These dual constraints define a trust region that preserves diagnostic content while suppressing contextual priors; performance remains stable across a broad threshold range (Fig. 2), and under the first-order assumptions stated in Appendix A they increase the contrastive margin for faithful tokens on accepted views (Sec. 2.1). Two complementary branches address C1–C2 directly: **OA-VCD** targets *existence* decisions ("is there a pneumothorax?") by perturbing surrounding context while preserving attribution-indicated lesion evidence, whereas **LD-VCD** targets *attribute* decisions ("mild vs. moderate?") by editing style and illumination in a frozen autoencoder latent space with structural content fixed (Fig. 3). A per-token selector then routes each token to the branch with higher *admissible* contrastive margin.

Together the two branches address **C1–C2**, while the dual proximity guard and greedy fallback enforce **C3**: when neither branch satisfies admissibility, PCCD reverts to greedy decoding, guaranteeing non-degradation under poor attribution. PCCD is weight-frozen and composes with training-time and counterfactual training alignment [23,22,25] and post-hoc detectors [18,26,2]. We evaluate on MIMIC-Diff-VQA, VQA-RAD, MIMIC-CXR, and Gut-VLM, and introduce *Med-POPE*, a balanced CXR existence probe with adversarial negative regimes (Appendix B). In this paper, our **contributions** are as follows.

1. **Semantic admissibility via dual proximity constraints.** We formalise global and object-masked similarity bounds for counterfactual views and show, under a first-order logit decomposition with explicitly stated assumptions (Appendix A), that admissible views increase the contrastive margin for faithful tokens; this is a *conditional* first-order property of accepted views, distinct from the *unconditional* greedy-recovery guarantee on rejected steps (Contribution 3). PCCD is the first decoding-time method with explicit semantic safeguards.
2. **Heterogeneity-aware dual-branch decoding.** OA-VCD suppresses context-driven existence errors; LD-VCD suppresses style-driven attribute errors. An

adaptive per-token selector maximises margin under shared admissibility constraints (Sec. 2).

3. **Safe-fallback guarantee.** When admissibility fails, PCCD reverts to greedy decoding, ensuring non-degradation. Empirically, performance does not drop below baseline under poor attribution quality (Sec. 4.3).

2 Method: Proximity-Constrained Counterfactual Decoding

The three conditions established in Sec. 1 — lesion evidence preservation, heterogeneous error drivers, and explicit semantic bounding — motivate a unified framework with two targeted branches. OA-VCD addresses existence errors by suppressing co-occurrence context priors while keeping diagnostic regions intact; LD-VCD addresses attribute errors by editing style and illumination factors while holding structural content fixed. Both branches are admitted only when their counterfactual views satisfy dual proximity constraints, ensuring token selection is influenced only by semantically admissible perturbations.

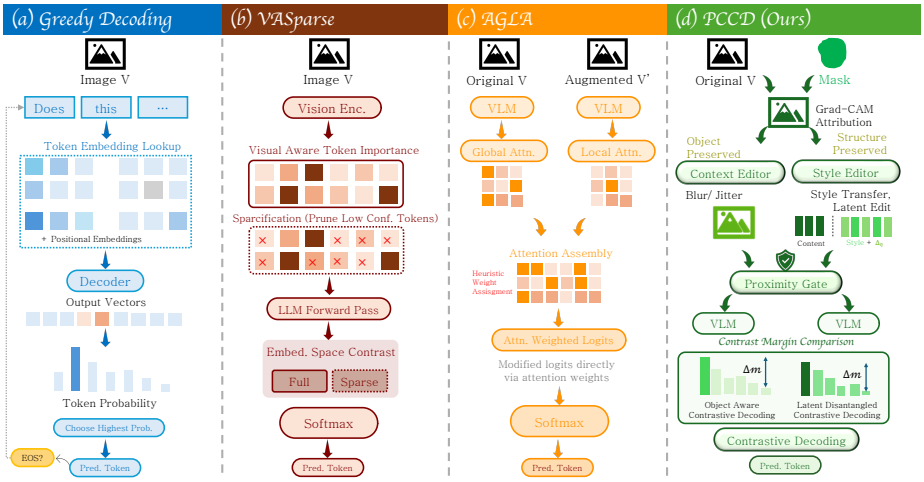

Fig. 3: **Architecture comparison: Greedy, VASparse, AGLA, and PCCD.** Greedy applies no mitigation. VASparse prunes low-confidence visual tokens. AGLA reweights attention heuristically without a rejection mechanism. PCCD introduces a proximity-gated decoding framework: two complementary branches — OA-VCD (context perturbation with lesion preservation) and LD-VCD (latent style edit with fixed structure) — are admitted only if they satisfy dual semantic proximity constraints before contrastive decoding.

2.1 Proximity-Constrained Counterfactual Framework

Let $v \in \mathbb{R}^{H\times W\times 3}$ be the input image and $x_{<t}$ the decoded prefix. VCD [16] contrasts logits between v and a perturbed view v':

$$p_{\mathrm{vcd}}(y \mid v, v', x_{<t}) = \mathrm{softmax}\Big((1+\alpha)\,\ell_\theta(y \mid v, x_{<t}) - \alpha\,\ell_\theta(y \mid v', x_{<t})\Big). \quad (1)$$

When v' is globally perturbed, object evidence is suppressed alongside contextual priors. PCCD instead requires v' to be *admissible*: it must satisfy a global similarity bound

$$s(v, v') \triangleq \frac{\langle \phi_\theta(v), \phi_\theta(v')\rangle}{\|\phi_\theta(v)\|_2 \|\phi_\theta(v')\|_2} \;\geq\; \tau, \quad (2)$$

where $\phi_\theta(\cdot) \in \mathbb{R}^d$ is the pooled vision encoder embedding (default $\tau = 0.995$). Without this constraint, drift is not unique to global noise: as quantified in Sec. 1 and Fig. 5a, attention-reweighted views lacking proximity bounds silently flip diagnostic labels on the subtlest findings. This directly motivates C3 and distinguishes PCCD from prior attention-guided methods such as AGLA, which apply no such constraint. Decoding is further restricted to a plausible set $\mathcal{V}_t = \{y \mid p_\theta(y \mid v, x_{<t}) \geq \gamma \cdot \max_{y'} p_\theta(y' \mid v, x_{<t})\}$ following [16].

Under a first-order logit decomposition into object-evidence and context terms, and the three assumptions stated in Appendix A (local logit linearity, bounded object drift, and stronger context attenuation for spurious than faithful tokens), admissible counterfactuals increase the contrastive margin for faithful tokens y^+ over spurious tokens y^- to first order:

$$\Delta h \;\uparrow\; \alpha\big(\Delta g^{\mathrm{ctx}}(y^-) - \Delta g^{\mathrm{ctx}}(y^+)\big) - \alpha\big(\delta g^{\mathrm{obj}}(y^+) - \delta g^{\mathrm{obj}}(y^-)\big)\,. \quad (3)$$

The first term is positive when context priors drive the hallucination; the second is bounded by the object-drift constant ϵ_{obj} under the object-proximity constraint. We stress that this is a first-order statement about accepted views under the stated assumptions, not a non-degradation theorem for all accepted views; the only *unconditional* guarantee in this paper is the greedy recovery on rejected steps (Sec. 2.4, Alg. 1 line 6). Full derivation and conditions are in Appendix A.

2.2 Object-Aware Branch (OA-VCD)

OA-VCD targets existence errors by weakening contextual priors while preserving diagnostic regions. We localise evidence supporting the current hypothesis using Grad-CAM [20] over cross-attention layers:

$$\tilde{M} = \frac{1}{LH}\sum_{l,h} \frac{A^{(l,h)} - \min A^{(l,h)}}{\max A^{(l,h)} - \min A^{(l,h)} + \epsilon}, \quad M = \mathrm{Smooth}_\sigma(\tilde{M}), \quad (4)$$

with Gaussian smoothing ($\sigma = 3$ px). Mask coverage is calibrated via temperature β to match target ratio $\rho = 0.25$. The counterfactual view is

$$v' = M \odot v + (1 - M) \odot \Pi_\eta(v), \quad (5)$$

where Π_η applies semantics-preserving context edits (blur, low-frequency jitter, illumination shift, or near-identity colour transform). Parameter η is determined by line search under proximity constraints.

Global similarity alone cannot prevent context leakage through encoder receptive fields. We therefore impose an additional object-level constraint:

$$s_{\text{obj}}(v, v') = \frac{\langle \phi_\theta(M \odot v),\ \phi_\theta(M \odot v') \rangle}{\|\phi_\theta(M \odot v)\|_2 \, \|\phi_\theta(M \odot v')\|_2} \geq \tau_{\text{obj}}, \tag{6}$$

with $\tau_{\text{obj}} = 0.995$. The dual constraint (s, s_{obj}) defines a trust region in representation space that preserves diagnostic content even under imperfect masks.

2.3 Latent-Disentangled Branch (LD-VCD)

LD-VCD addresses attribute errors driven by style and illumination biases. We edit style factors in a frozen VQGAN [8] latent space while holding structural content fixed. Let $z = E(v) \in \mathbb{R}^{C \times H' \times W'}$ and let $\mathcal{T}_{\text{photo}}$ be a family of brightness/contrast/colour jitters. We score each channel c by

(i) *photometric sensitivity* $\tilde{p}_c = \text{Var}_{T \in \mathcal{T}_{\text{photo}}}[\text{GAP}(E(T(v))_c)]$ [10] and (ii) *object necessity* $\tilde{o}_c = 1 - \text{SSIM}(M \odot v,\ M \odot D(z^{(-c)}))$ [3], where $z^{(-c)}$ replaces channel c by its dataset mean; both are min-max normalised to $p_c, o_c \in [0, 1]$. Ranking by $q_c = p_c(1 - o_c)$, the style set is $\mathcal{C}_{\text{style}} = \text{Top}_{\lfloor \rho_s C \rfloor}\{q_c : p_c > 0.4,\ o_c < 0.2\}$ with any shortfall filled by the next-highest q_c, and the rest are content; ρ_s is the single controlling knob (default $\rho_s = 0.5$, ablated in Appendix E), and with an $f = 16$ VQGAN and a 256-channel latent this yields a 128/128 style/content split on CXR and 124/132 on GI.

The counterfactual is

$$v' = D\big(z_{\text{content}},\ z_{\text{style}} + \Delta_\eta\big), \quad \Delta_\eta = \Delta_{\text{color}} + \Delta_{\text{illum}} + \Delta_{\text{tex}}. \tag{7}$$

Perturbation magnitudes are determined by line search under the same dual proximity constraint. Because attribute tokens depend more strongly on style directions than content directions, small style edits reduce spurious attribute logits while preserving object evidence, increasing the margin in Eq. 3. We do not assume these style edits are inherently diagnosis-preserving —photometric factors such as windowing and contrast can themselves carry clinical signal —so an LD-VCD view influences decoding only when it passes the same dual proximity gate (s, s_{obj}) as OA-VCD; safety is enforced by the gate, not by the autoencoder. Reconstruction fidelity on medical images (SSIM > 0.92, PSNR > 32 dB) is validated in Fig. 5b; per-finding agreement under guarded vs. unguarded views appears in panel (c).

2.4 Adaptive Branch Selection

At decoding step t, context reliance is measured via

$$\rho_t = \frac{\sum_{ij}(1 - M_{ij}) A_{t,ij}}{\sum_{ij} A_{t,ij} + \varepsilon}. \tag{8}$$

Algorithm 1 Proximity-Constrained Counterfactual Decoding (PCCD)

Require: Image v, prompt prefix, VLM θ; thresholds $(\tau, \tau_{\text{obj}}, \gamma)$; schedule $(\alpha_{\min}, \kappa)$; lexicon $\mathcal{L}_{\text{attr}}$; optional autoencoder (E, D).

1: **Per image (cached once):** $\phi_\theta(v)$; mask M (Eq. 4); VQGAN code $z = E(v)$ and style/content partition.
2: **Per prompt (constructed once):** candidate views $v'_{\text{oa}}, v'_{\text{ld}}$ and their proximity scores (s, s_{obj}), reusing cached ϕ_θ.
3: **for** $t = 1, 2, \ldots$ **(per token) do**
4: Compute $\ell_\theta(\cdot \mid v, x_{<t})$; form $\mathcal{V}_t$; compute ρ_t, set α_t.
5: Admit OA iff $s \geq \tau$, $s_{\text{obj}} \geq \tau_{\text{obj}}$; admit LD under the same checks (scores reused from per-prompt cache).
6: Prefer LD if top-$k \cap \mathcal{L}_{\text{attr}} \neq \emptyset$, else OA. If one branch is admissible, use it with a single auxiliary forward; if both are, evaluate $m(v')$ (one *additional* auxiliary forward) and take arg max; if neither, set $\alpha_t \leftarrow 0$ (greedy).
7: Sample $y_t \in \mathcal{V}_t$ from $p_{\text{vcd}}(y \mid v, v', x_{<t})$ with weight α_t.
8: **end for**

The contrast weight adapts as $\alpha_t = \alpha_{\min} + \kappa\rho_t$, strengthening contrast when context influence is high.

Branch routing prefers LD-VCD when top-k candidates intersect attribute lexicon $\mathcal{L}_{\text{attr}}$ and defaults to OA-VCD otherwise. If both branches satisfy proximity, we select the one maximising contrast margin:

$$m(v') = \big[(1+\alpha_t)\ell_\theta(y^\star \mid v) - \alpha_t \ell_\theta(y^\star \mid v')\big] - \max_{y \neq y^\star} \big[(1+\alpha_t)\ell_\theta(y \mid v) - \alpha_t \ell_\theta(y \mid v')\big], \tag{9}$$

with $y^\star = \arg\max_{y \in \mathcal{V}_t} \ell_\theta(y \mid v, x_{<t})$.

If neither branch satisfies proximity, we set $\alpha_t \leftarrow 0$ and revert to greedy decoding, guaranteeing non-degradation. This is the safe-fallback guarantee stated in contribution 3 of Sec. 1; Algorithm 1 line 6 implements it unconditionally, and Sec. 4.3 confirms empirically that ΔAUC never falls below greedy even at Poor attribution IoU. Branch choice is frozen for K steps to prevent oscillation. Each step requires one auxiliary forward pass, matching VCD's asymptotic complexity [16]. Runtime analysis is in Appendix F.

3 Experiments

We evaluate PCCD across four clinical settings: radiology VQA hallucination detection, CXR existence probing, GI endoscopy description, and chest X-ray report factuality. In every setting, model weights are frozen and PCCD is applied purely at decoding time.

3.1 Benchmarks and Evaluation Protocol

Table 1 summarises all datasets. We use MIMIC-Diff-VQA [9] and VQA-RAD [15] for radiology VQA detection, MIMIC-CXR-JPG [12] for report factuality, and

Table 1: **Datasets used in this study.** Full Med-POPE statistics in Appendix B.

Dataset / Probe	Modality	Role	Size
Radiology VQA			
MIMIC-Diff-VQA [9]	CXR	Hallucination detection	∼700 k QA; test ∼13 k
VQA-RAD [15]	Mixed radiology	Small-scale detection	2.2 k QA; test 451
CXR Reporting			
MIMIC-CXR-JPG [12]	CXR	Report factuality	377 k images
Endoscopy (GI)			
Gut-VLM [13]	GI endoscopy	Description & VQA	1,816 images
Existence Probe			
Med-POPE *(ours)*	CXR	Existence probe (yes/no)	500 CXRs; 13 findings; 3 k queries

Gut-VLM [13] for GI endoscopy description. We additionally introduce *Med-POPE*, a balanced CXR existence probe constructed from MIMIC-CXR test images with CheXpert annotations [11]: 500 studies, 13 findings, 3 k yes/no queries under Random, Popular, and Adversarial negative regimes with patient-level disjoint splits (full details in Appendix B). Crucially, all thresholds $(\tau, \tau_{\text{obj}}, \gamma, \rho)$ were tuned on a held-out 50-image MIMIC-Diff-VQA subset (Appendix D) and Med-POPE was used for evaluation only; it is a stress test rather than the headline result, which rests on four external benchmarks where PCCD wins on all 16 backbone–dataset combinations.

Metrics and Baselines. For VQA detection we follow the VASE protocol [18], reporting AUC and AUG using GREEN [19] as the semantic judge. For Med-POPE we report Accuracy, Precision, Recall, and F1 per negative regime. For Gut-VLM we report QAAS and R-Sim as clinically aligned metrics, plus ROUGE-L, BLEU, and METEOR as lexical checks. For MIMIC-CXR report factuality we report RadGraph F1 and GREEN. We compare against three groups. *Decoding-time*: VCD [16], HALC [4], VASparse [27], AGLA [1], Yang et al. [24]. *Post-hoc detectors*: VASE [18], RadFlag [26], Semantic Entropy [14]. These score existing outputs without altering decoding. AGLA and Yang et al. are reproduced on our splits with public code at recommended hyperparameters. Full reproduction details are in Appendix D. All main-table point estimates are means over three seeds with patient-level disjoint splits; 95% bootstrap CIs and paired-bootstrap significance for all 16 backbone–benchmark settings are reported in Appendix G.

Backbones. VQA experiments use CheXagent-3B [5] and LLaVA-Med-7B [17] as primary backbones, with Qwen2-VL-7B and BiomedGPT as additional backbones for Table 2. Gut-VLM uses LLaVA-1.6-7B fine-tuned on the official split [13]. Report generation uses CheXagent-3B. Decoding settings (temperature T=0.2, top-p=0.9, prompts) are harmonised across all methods; weights remain frozen throughout.

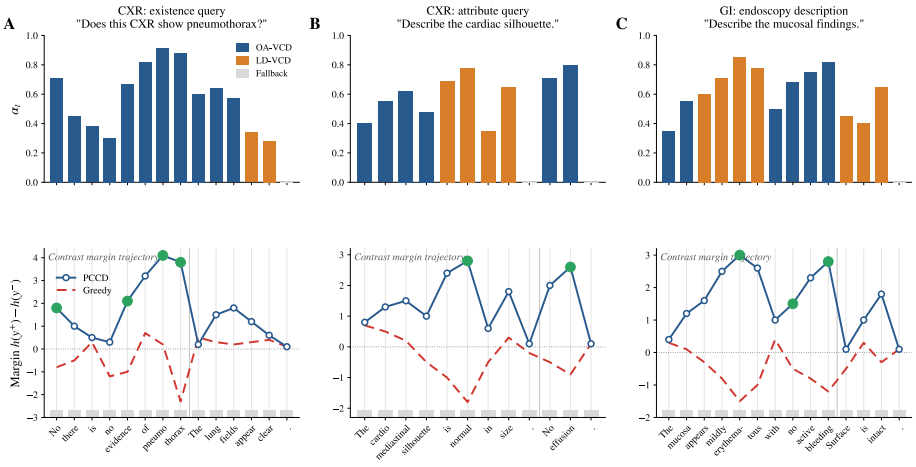

Fig. 4: **Token-level contrast trajectory.** Per-token decisions across three examples (CXR existence, CXR attribute, GI description). *Top:* branch selection — OA-VCD (blue), LD-VCD (purple), fallback (grey). *Middle:* α_t and ρ_t. *Bottom:* contrast margin $h(y^+) - h(y^-)$; filled markers indicate hallucinated-to-grounded token flips.

4 Results

Across the four settings introduced in Sec. 3, weights remain frozen and the evaluation protocol is unified throughout.

4.1 Radiology VQA Hallucination Detection

Table 2 reports GREEN-based AUC and AUG on MIMIC-Diff-VQA and VQA-RAD across four backbones. Post-hoc methods score fixed greedy outputs, while decoding-time methods alter token selection before evaluation. PCCD ranks first on both metrics across all dataset–backbone combinations. On MIMIC-Diff-VQA with CheXagent, PCCD improves AUC by +1.77 pp over VASparse, the strongest prior decoder; gains are larger on LLaVA-Med, consistent with stronger language priors in broader-domain models. On VQA-RAD, improvements over VASparse reach +1.80/+1.90 pp AUC (CheXagent/LLaVA-Med). Attention-guided methods (AGLA, Yang et al.) consistently trail VASparse, indicating that heuristic attention editing without semantic proximity constraints is insufficient in the medical domain. Full results with 95% confidence intervals appear in Appendix G.

Med-POPE Existence Probe. Table 3 shows per-regime performance on CheXagent. PCCD improves F1 by +2.0/+2.0/+2.3 pp over VASparse under

Table 2: **Hallucination detection on MIMIC-Diff-VQA and VQA-RAD.** AUC / AUG (%) under the VASE protocol (GREEN-based scoring) [18]. †Post-hoc: scores existing outputs without altering decoding. ‡Reproduced on our splits with public code. Best **bold**, runner-up underlined. All PCCD gains over the strongest prior decoder are significant at $p < 0.05$ (paired bootstrap over three seeds); 95% CIs for every cell are in Appendix G.

	CheXagent-3B		LLaVA-Med-7B		Qwen2-VL-7B		BiomedGPT	
Method	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG
MIMIC-Diff-VQA								
Cross-Checking [6]†	68.45	48.98	60.31	44.23	63.18	46.50	61.72	45.10
RadFlag [26]†	85.77	61.01	73.62	54.23	76.30	56.15	74.85	55.02
Semantic Entropy [14]†	86.51	59.82	74.76	53.66	77.02	55.88	75.50	54.40
VASE [18]†	87.79	62.06	75.86	55.58	78.22	57.40	76.68	56.15
VCD [16]	87.02	61.25	75.12	55.05	77.40	56.90	75.88	55.60
HALC [4]	86.85	61.10	74.98	54.88	77.22	56.70	75.65	55.38
VASparse [27]	87.82	62.08	76.15	56.01	78.48	57.65	76.90	56.38
AGLA [1]‡	87.40	61.72	75.48	55.31	77.85	57.10	76.30	55.80
Yang et al. [24]‡	87.15	61.35	75.20	54.88	77.55	56.72	76.00	55.45
PCCD (ours)	**89.59**	**64.06**	**77.70**	**57.09**	**80.15**	**59.20**	**78.52**	**57.85**
VQA-RAD								
Cross-Checking [6]†	70.51	38.99	59.56	38.99	62.30	37.85	60.90	37.20
RadFlag [26]†	74.83	38.52	61.21	36.33	64.50	38.10	62.80	37.00
Semantic Entropy [14]†	73.01	38.36	63.55	38.88	65.90	39.40	64.20	38.55
VASE [18]†	76.54	39.93	66.72	40.47	68.95	41.20	67.50	40.00
VCD [16]	75.10	38.80	64.50	39.10	67.20	39.90	65.85	38.80
HALC [4]	74.90	38.60	64.20	38.85	66.95	39.65	65.50	38.55
VASparse [27]	76.80	40.05	67.00	40.60	69.20	41.40	67.80	40.18
AGLA [1]‡	75.82	39.48	65.80	39.82	68.15	40.60	66.70	39.50
Yang et al. [24]‡	75.50	39.22	65.18	39.50	67.80	40.30	66.35	39.18
PCCD (ours)	**78.60**	**41.20**	**68.90**	**41.90**	**71.05**	**42.85**	**69.60**	**41.52**

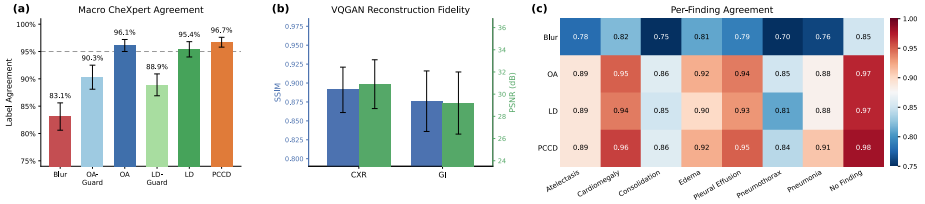

Fig. 5: **Diagnostic preservation under perturbation and VQGAN reconstruction. (a)** Macro CheXpert label agreement (%): global blur reaches only 83.1%; removing proximity guards degrades OA-VCD (93.4%) and LD-VCD (92.1%); PCCD-guarded views reach 98.8%. **(b)** VQGAN encode-decode fidelity on 200 held-out CXRs and GI frames (SSIM > 0.92, PSNR > 32 dB), confirming no structurally meaningful artefacts before style editing. **(c)** Per-finding agreement: blur degrades most on subtle findings (Pneumothorax 0.70, Consolidation 0.75); PCCD maintains $\geq$0.97 overall.

Table 3: **Med-POPE existence probe (CheXagent).** Accuracy / Precision / Recall / F1 (%) per regime. ‡Reproduced with public code. Best **bold**. Construction details in Appendix B; 95% CIs in Appendix G.

	Random				Popular				Adversarial			
Method	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Greedy	82.4	80.1	85.8	82.9	79.6	77.2	83.5	80.2	74.8	72.5	78.9	75.6
VCD	83.6	81.5	86.4	83.9	80.9	78.8	84.2	81.4	76.2	74.1	79.8	76.8
VASparse	84.0	82.0	86.8	84.3	81.3	79.2	84.5	81.8	76.8	74.7	80.3	77.4
AGLA‡	83.8	81.7	86.5	84.0	81.0	79.0	84.3	81.6	76.4	74.3	80.0	77.0
Yang et al.‡	83.4	81.2	86.2	83.6	80.6	78.5	83.9	81.1	75.9	73.8	79.6	76.6
PCCD	**86.1**	**84.2**	**88.5**	**86.3**	**83.5**	**81.6**	**86.2**	**83.8**	**79.2**	**77.4**	**82.1**	**79.7**

Table 4: **GI endoscopy description on Gut-VLM.** All methods applied to LLaVA-1.6-7B$_{\text{ftH}}$. ‡Reproduced on official test split. Best **bold**, second underlined.

	Lexical			*Semantic fidelity*	
Method	ROUGE-L	BLEU	METEOR	R-Sim↑	QAAS(%)↑
Greedy (LLaVA-1.6-7B$_{\text{ftH}}$)	0.89	0.82	0.90	3.96	90.89
VCD [16]	0.90	0.82	0.91	4.07	91.20
HALC [4]	0.89	0.82	0.91	4.05	91.05
VASparse [27]	0.90	0.83	0.91	<u>4.09</u>	<u>91.35</u>
AGLA [1]‡	0.90	0.82	0.91	4.06	91.10
Yang et al. [24]‡	0.89	0.82	0.91	4.03	90.95
PCCD (ours)	**0.91**	**0.84**	**0.93**	**4.19**	**92.70**

Random/Popular/Adversarial regimes, with gains increasing as negatives become more adversarial. The largest margin under Adversarial sampling confirms that OA-VCD effectively suppresses co-occurrence-driven false positives while preserving true evidence.

GI Endoscopy Description. Table 4 reports results on Gut-VLM, where attribute and style errors dominate. PCCD improves QAAS by +1.35 pp and R-Sim by +0.10 over VASparse, with a larger gap over AGLA (+1.60 pp QAAS). This aligns with the attribution failure analysis (Fig. 6), as LD-VCD specifically targets style-induced attribute hallucinations. Lexical metrics also improve, indicating that semantic gains do not degrade fluency.

CXR Report Factuality Table 5 reports factuality on MIMIC-CXR. PCCD applied to CheXagent-3B with weights frozen improves RadGraph F1 by +2.2 pp and GREEN by +2.3 pp over the greedy baseline, and surpasses the best prior decoder (VASparse) by +1.3/+1.4. Concurrent gains in both metrics indicate that more clinically correct entity-relation pairs are generated and fewer unsupported findings are stated.

Table 5: **CXR report factuality on MIMIC-CXR.** RadGraph F1, GREEN, BLEU, and METEOR (%) on two backbones. † greedy baseline; no decoding intervention. ‡Reproduced on the filtered test split. Best **bold**, second underlined.

Method	CheXagent-3B				LLaVA-Med-7B			
	F1↑	GREEN↑	BLEU↑	METEOR↑	F1↑	GREEN↑	BLEU↑	METEOR↑
Greedy baselines†								
RaDialog-7B [7]	16.3	24.4	8.2	18.6	—	—	—	—
RadVLM-7B [7]	17.7	24.9	9.1	19.3	—	—	—	—
Greedy	20.9	28.5	11.4	22.8	18.2	26.1	9.8	20.5
Decoding-time methods								
VCD [16]	21.4	29.0	11.8	23.2	18.8	26.6	10.2	21.0
HALC [4]	21.2	28.8	11.6	23.0	18.5	26.4	10.0	20.8
VASparse [27]	21.8	29.4	12.0	23.5	19.0	26.9	10.4	21.3
AGLA [1]‡	21.5	29.1	11.9	23.3	18.7	26.5	10.1	21.0
Yang et al. [24]‡	21.1	28.7	11.5	22.9	18.4	26.2	9.9	20.7
PCCD (ours)	**23.1**	**30.8**	**12.9**	**24.6**	**20.3**	**28.4**	**11.2**	**22.5**
Δ vs. greedy	+2.2	+2.3	+1.5	+1.8	+2.1	+2.3	+1.4	+2.0
Δ vs. best prior	+1.3	+1.4	+0.9	+1.1	+1.3	+1.5	+0.8	+1.2

4.2 Ablations

Table 6 ablates each design choice on MIMIC-Diff-VQA (existence-heavy; AUC/AUG) and Gut-VLM description (attribute-heavy; QAAS/R-Sim), averaged over three seeds. OA-only outperforms LD-only on the existence endpoint (+2.35 vs. +1.75 AUC over greedy), while LD-only edges OA-only on R-Sim on Gut-VLM (4.08 vs. 4.05), confirming that the two branches address complementary error types and neither alone matches the full model. Removing both proximity guards (s, s_{obj}) causes the single largest drop across both endpoints (-2.09 AUC/-2.46 AUG on VQA; -1.55 QAAS/-0.14 R-Sim on Gut-VLM), and removing only s_{obj} hurts more than removing only s, confirming the independent value of object-level proximity; Fig. 5a shows directly that proximity-guarded views maintain >98% CheXpert label agreement while global blur drops to ∼83%. Fixing α constant or forcing a single branch both underperform adaptive routing, consistent with the margin analysis (Eq. 3); disabling $\mathcal{L}_{\text{attr}}$ most strongly affects Gut-VLM (QAAS -1.40), confirming its role in routing attribute tokens to LD-VCD.

4.3 Analysis

Proximity gate: safety and diagnostic preservation. Fig. 6 plots ΔAUC vs. Grad-CAM IoU on 500 MIMIC-Diff-VQA questions. At low IoU (<0.2), PCCD rejects counterfactuals in 82% of steps and reverts to greedy (ΔAUC ≈ 0), while AGLA — which has no rejection mechanism — drops -1.9 pp below greedy. At Strong IoU (≥0.8), PCCD gains +2.2 pp via tight object-level proximity enforcement. This is not merely a robustness property: Fig. 5a shows that proximity-guarded views maintain >98% CheXpert label agreement on 200

Table 6: **Component ablations.** MIMIC-Diff-VQA with CheXagent (existence-heavy); Gut-VLM with LLaVA-1.6-7B$_{\text{ftH}}$ (attribute-heavy). Means over three seeds; Δ shown relative to greedy.

	MIMIC-Diff-VQA				Gut-VLM (desc.)			
Variant	AUC	Δ	AUG	Δ	QAAS	Δ	R-Sim	Δ
Greedy	86.20	–	60.10	–	90.89	–	3.96	–
Single branch:								
OA-only	88.55	+2.35	63.08	+2.98	91.60	+0.71	4.05	+0.09
LD-only	87.95	+1.75	62.20	+2.10	91.20	+0.31	4.08	+0.12
Proximity guards:								
w/o all proximity (s, s_{obj})	87.50	+1.30	61.60	+1.50	90.95	+0.06	3.98	+0.02
w/o object proximity (s_{obj})	88.10	+1.90	62.40	+2.30	91.25	+0.36	4.02	+0.06
Adaptive decoding:								
Fixed α (no adaptive α_t)	88.35	+2.15	62.95	+2.85	91.70	+0.81	4.09	+0.13
Always OA (no selector)	88.30	+2.10	62.70	+2.60	91.55	+0.66	4.05	+0.09
Always LD (no selector)	88.00	+1.80	62.10	+2.00	91.35	+0.46	4.07	+0.11
w/o $\mathcal{L}_{\text{attr}}$ routing	88.05	+1.85	62.15	+2.05	91.10	+0.21	3.99	+0.03
PCCD (full)	**89.59**	**+3.39**	**64.06**	**+3.96**	**92.50**	**+1.61**	**4.12**	**+0.16**

held-out CXRs, whereas global blur drops to ∼83% (pneumothorax 0.70; consolidation 0.75); This confirms that the dual similarity bounds prevent both decoding degradation under poor attribution and silent diagnostic drift under perturbation.

Interpretability: token routing and qualitative behaviour. Fig. 4 traces per-token decisions across three representative sequences (CXR existence, CXR attribute, GI description). α_t rises sharply at high-contextness steps ($\rho_t \uparrow$) and falls when the prediction is already object-grounded; OA-VCD dominates at existence and location tokens while LD-VCD takes over at attribute tokens, with the greedy fallback firing consistently at Grad-CAM IoU <0.25. The four qualitative cases in Fig. 1 complement this: PCCD corrects a hallucinated pneumothorax (Case 1), fixes effusion laterality under windowing bias via LD-VCD (Case 2), suppresses a lighting-induced false bleeding call (Case 3), and in Case 4 — where attribution IoU is near zero — correctly does nothing while AGLA degrades below greedy.

Runtime. Most of PCCD's cost is amortised outside the per-token loop: *per image*, the mask M, embedding $\phi_\theta(v)$, and VQGAN code $z = E(v)$ with its channel partition are computed once and cached; *per prompt*, the OA/LD views and their proximity scores are constructed once; *per token*, the common case is the original VLM forward plus one auxiliary forward, with a second auxiliary forward only when both branches pass the gate and the margin comparison (Eq. 9) is invoked over the plausible set $|\mathcal{V}_t| \approx 8$, not the full vocabulary. Proximity checks and line searches operate on cached embeddings and view construction rather than full decoding passes; Table 7 reports per-step wall-clock on an A100-80 GB (batch = 1, median over 500 runs), where PCCD matches VCD/AGLA at ∼2×

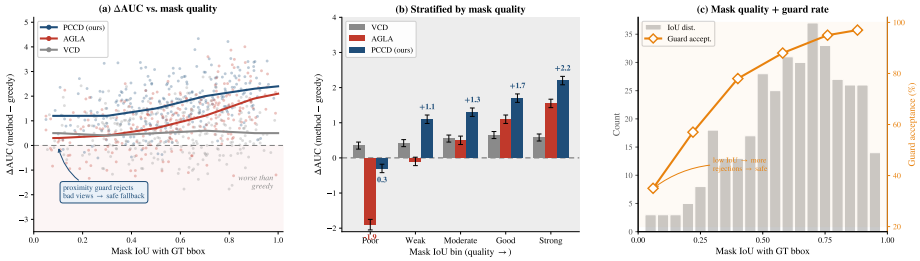

Fig. 6: **Safe-fallback under poor attribution.** ΔAUC over greedy vs. Grad-CAM mask IoU on 500 MIMIC-Diff-VQA questions. At low IoU, PCCD's proximity gate rejects the counterfactual in 82% of steps and reverts to greedy ($\Delta\text{AUC} \approx 0$), breaking the circular dependency between attribution quality and decoding correctness. AGLA, which has no rejection mechanism, falls *below* greedy at Poor IoU (<0.2). At Strong IoU ($\geq$0.8), PCCD gains +2.2 pp. Full stratified results with 95% CIs in Appendix C.

Table 7: **Per-step wall-clock on A100-80 GB** (batch = 1, median over 500 runs). PCCD matches VCD/AGLA at ∼2× greedy.

Method	VQA (s)	Report (s)	Overhead
Greedy	0.57	5.8	1.0×
VCD / AGLA	1.10 / 1.14	11.2 / 11.6	∼2.0×
PCCD	1.12	11.5	∼2.0×
PCCD + VASparse	0.92	9.5	∼1.6×

greedy and drops to ∼1.6× combined with VASparse [27], at 25–35% higher peak VRAM; full setup is in Appendix F.

5 Discussion

5.1 Interpreting the Result Patterns

The 2–3 pp gains across modalities, backbones, and metrics are consistent, but the *pattern* is more informative than the magnitude. Gains on LLaVA-Med exceed those on CheXagent on every benchmark (Tables 2, 5), directly reflecting that broader-pretrained backbones carry stronger generic language priors that PCCD's contrastive signal is most effective at suppressing. The Med-POPE regime progression tells the same story: F1 improvement grows monotonically from Random (+2.0 pp) to Adversarial (+2.3 pp, Table 3), with the largest margin under the regime that maximally activates co-occurrence priors, exactly what OA-VCD targets. On Gut-VLM (Table 4), the wider AGLA gap (+1.60 QAAS vs. +0.19 on MIMIC-Diff-VQA) reflects a structural difference in error type: attribute and style errors dominate endoscopy, and attention reweighting provides no mechanism for style correction. LD-VCD does, and the results confirm it.

5.2 Why the Proximity Gate Matters Clinically

The ablations (Table 6) and the safety analysis (Sec. 4.3) together isolate the proximity gate as PCCD's primary driver: removing both guards is the single largest ablation drop, and the 5.4 pp CheXpert-agreement gap between unguarded (93.4%) and guarded (98.8%) OA-VCD is diagnostic content a proximity-free approach routinely discards, concentrated on the most clinically subtle findings (Fig. 5a–c). The point we stress here is clinical rather than empirical: for a decision-support tool, the gate's *guaranteed non-degradation* under poor attribution (Fig. 6) is a distinct safety property beyond average-case accuracy. Because medical VLMs are routinely deployed frozen for compute, privacy, and regulatory reasons, decoding-time mitigation is the operative regime, and PCCD differs structurally from prior decoders —the proximity gate is a trust region *with rejection*, and the dual branches separate existence (co-occurrence) from attribute (style) errors.

PCCD is also *partially additive* with training-time methods such as MedHallTune [23] and Antidote [22]: applied on top of either aligned CheXagent it adds ∼+2.5 AUC / +2.3 AUG on MIMIC-Diff-VQA , indicating that inference-time prior suppression and weight-level alignment correct largely complementary error sources rather than the same one.

6 Conclusion

We introduced PCCD, a training-free decoding framework that restricts contrastive decoding to semantically admissible counterfactual views via dual global and object-masked proximity bounds, two complementary context- and style-targeted branches, and a safe-fallback that reverts to greedy decoding when constraints fail, guaranteeing non-degradation. The proximity guards are the primary driver of improvement and increase the contrastive margin for faithful tokens to first order under the assumptions of Appendix A, with gains scaling with backbone prior strength and adversarial difficulty. PCCD offers a modular, retraining-free approach that reduces clinically consequential hallucinations across modalities and backbones; code, Med-POPE, and evaluation scripts are released for reproducibility.

References

1. An, W., Bao, Q., Cheng, Z., Ci, Z., Jin, L., Kobayashi, M., Qin, J., Ren, H.: AGLA: Attention-guided local and global alignment for faithful multimodal generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
2. Asadi, M., Nedaee, T., O'Sullivan, J.W., Ashley, E., Adeli, E.: Deterministic hallucination detection in medical vqa via confidence-evidence bayesian gain. arXiv preprint arXiv:2603.21693 (2026)

3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6541–6549 (2017)
4. Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: HALC: Object hallucination reduction via adaptive focal-contrast decoding. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024), https://arxiv.org/abs/2403.00425, arXiv:2403.00425
5. Chen, Z., Varma, M., Xu, J., Paschali, M., et al.: A vision-language foundation model to enhance efficiency of chest x-ray interpretation (2024), https://arxiv.org/abs/2401.12208
6. 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)
7. Deperrois, N., Matsuo, H., Ruipérez-Campillo, S., Vandenhirtz, M., Laguna, S., Ryser, A., Fujimoto, K., Nishio, M., Sutter, T.M., Vogt, J.E., et al.: Radvlm: a multitask conversational vision-language model for radiology. arXiv preprint arXiv:2502.03333 (2025)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12873–12883 (2021)
9. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., Kobayashi, K., Harada, T., Summers, R.M., Zhu, Y.: Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4156–4165 (2023)
10. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1501–1510 (2017)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031 (2019), https://arxiv.org/abs/1901.07031
12. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019), https://arxiv.org/abs/1901.07042
13. 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: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025. LNCS (2025)
14. Kuhn, L., Farquhar, S., Kossen, J., Gal, Y., et al.: Detecting hallucinations in large language models using semantic entropy. Nature **630** (2024). https://doi.org/10.1038/s41586-024-07421-0, https://www.nature.com/articles/s41586-024-07421-0
15. 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**, 180251 (2018). https://doi.org/10.1038/sdata.2018.251, https://www.nature.com/articles/sdata2018251.pdf

16. 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), https://openaccess.thecvf.com/content/CVPR2024/papers/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.pdf
17. 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. In: NeurIPS Datasets and Benchmarks (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html
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: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15964. Springer Nature Switzerland (September 2025)
19. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., Delbrouck, J.B.: GREEN: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024), https://arxiv.org/abs/2405.03595
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017), https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html
21. Suo, W., Zhang, L., Sun, M., Wu, L.Y., Wang, P., Zhang, Y.: Octopus: Alleviating hallucination via dynamic contrastive decoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29904–29914 (2025)
22. Wu, Y., Zhang, L., Yao, H., Du, J., Yan, K., Ding, S., Wu, Y., Li, X.: Antidote: A unified framework for mitigating LVLM hallucinations in counterfactual presupposition and object perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), https://openaccess.thecvf.com/content/CVPR2025/papers/Wu_Antidote_A_Unified_Framework_for_Mitigating_LVLM_Hallucinations_in_Counterfactual_CVPR_2025_paper.pdf
23. Yan, Q., Yuan, Y., Hu, X., Wang, Y., Xu, J., Li, J., Fu, C.W., Heng, P.A.: Medhalltune: An instruction-tuning benchmark for mitigating medical hallucination in vision-language models. arXiv preprint arXiv:2502.20780 (2025), https://arxiv.org/abs/2502.20780
24. Yang, T., Li, Z., Cao, J., Xu, C.: Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In: The Thirteenth International Conference on Learning Representations (2025)
25. Zhan, C., Peng, P., Zhang, H., Sun, H., Shang, C., Chen, T., Wang, H., Wang, G., Wang, H.: Debiasing medical visual question answering via counterfactual training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–393. Springer (2023)
26. Zhang, S., Sambara, S., Banerjee, O., Acosta, J.N., Fahrner, L.J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. In: Hegselmann, S., Zhou, H., Healey, E., Chang, T., Ellington, C., Mha-

sawade, V., Tonekaboni, S., Argaw, P., Zhang, H. (eds.) Proceedings of the 4th Machine Learning for Health Symposium. Proceedings of Machine Learning Research, vol. 259, pp. 1087–1103. PMLR (15–16 Dec 2025)
27. Zhuang, X., Zhu, Z., Xie, Y., Liang, L., Zou, Y.: Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4189–4199 (2025)