

LENS: Exploiting the Cross-View Stability Gap for Hallucination Mitigation in Medical Vision–Language Models

Dwarikanath Mahapatra^{1*}, Abhijit Das^{2*}, Sudipta Roy^{3,4}
Behzad Bozorgtabar⁵, Imran Razzak²

¹Khalifa University, Abu Dhabi, UAE. ²MBZUAI, Abu Dhabi, UAE.

³Jio University, India. ⁴University of New South Wales, Bengaluru, India

⁵Aarhus University, Denmark.

Abstract

Medical VLMs frequently hallucinate clinically unsupported findings. We identify a *stability gap*: visually grounded tokens maintain relatively stable logits under semantics-preserving image perturbations, whereas visually misgrounded tokens exhibit substantially larger shifts. We introduce **LENS** (Logit-Enhanced Stability), which exploits this signal at both training and inference. During training, *Semantics-Preserving Preference Optimization* (SPPO) scores preference pairs using clinical faithfulness and cross-view stability; at inference, stability-aware contrastive decoding penalizes tokens whose logits diverge across original and counterfactual views. Constrained counterfactuals use attribution-guided masking and feature-space proximity guards to reduce disruption of lesion- and device-level evidence. Across radiology VQA, chest X-ray reporting, and gastrointestinal endoscopy, LENS consistently outperforms strong decoding-time and preference-optimization baselines. On MIMIC-CXR, entity-level RadGraph recall improves from 21.0 to 22.8 over PCCD, while expert evaluation shows a hallucination-rate reduction from 17.0% to 8.5%.

1 Introduction

Medical vision–language models (VLMs) now support radiology report generation, image-grounded question answering, and endoscopy description (Liu et al., 2023; Li et al., 2023; Chen et al., 2024b). However, they frequently *hallucinate*, generating clinically fluent text unsupported by visual evidence (Rohrbach et al., 2018; Bai et al., 2024). Because such outputs can propagate into health records and influence clinical decisions, even rare hallucinations carry disproportionate risk (Ostmeier et al., 2024; Liao et al., 2025).

Decoding-time methods contrast predictions between original and perturbed images, but often rely

*Equal contribution.

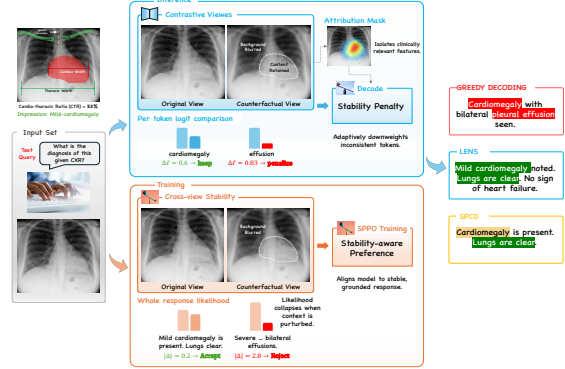


Figure 1: **The stability gap in medical VLM hallucinations.** *Top:* Per-token logit change $|\Delta\ell|$ under a semantics-preserving counterfactual view. *Bottom:* Response-level likelihood comparison.

on global perturbations that can disrupt subtle lesion and device cues critical to clinical interpretation (Leng et al., 2024; Chen et al., 2024a; Zhuang et al., 2025). *Training-time alignment* via preference optimization (DPO, PPO, GRPO) can reduce hallucinations, but typically relies on preference supervision without directly testing whether the preferred response is visually grounded (Jiang et al., 2024; Xie et al., 2024; Fu et al., 2025). *Post-hoc detectors* identify unsupported content only after generation (Zhang et al., 2025; Park et al., 2025). Our prior work, Proximity-Constrained Counterfactual Decoding (PCCD), introduced dual proximity-constrained counterfactual views for training-free medical VLM decoding (Mahapatra et al., 2026). LENS builds on this counterfactual machinery but treats cross-view stability as a supervision signal, using it for preference optimization during training and token-level correction at inference. Related token-perception work has also shown that token-level visual dependency signals can guide multimodal optimization (Huang et al., 2026).

Our starting point is a simple empirical observation. When a medical VLM generates a *visu-*

ally *misgrounded* token, its logit tends to shift sharply under perturbations of clinically irrelevant context—such as background texture, windowing, or acquisition style—whereas tokens supported by lesion- or device-level evidence are substantially more stable (Figure 1). Across 500 image–question pairs, hallucinated tokens exhibit mean logit change $|\Delta\ell| = 1.84$ versus 0.31 for grounded tokens ($p < 10^{-6}$). We term this difference the *stability gap*. Importantly, the signal targets hallucinations whose likelihood depends on visual context; purely language-driven errors may remain stable across visual perturbations and therefore require complementary mechanisms.

We propose **LENS** (Logit-**EN**hanced Stability), which operationalizes the stability gap in two complementary stages. During training, *Semantics-Preserving Preference Optimization* (SPPO) constructs preference pairs scored jointly on clinical faithfulness and cross-view stability. At inference, *stability-aware contrastive decoding* penalizes tokens whose logits diverge between original and counterfactual views, with penalty strength modulated by attribution-derived contextness. Both stages use same constrained counterfactual machinery: attribution-guided masking limits perturbations around salient regions, while global and object-level feature-space proximity guards reject views that drift beyond predefined similarity thresholds.

LENS requires access to token logits and an attribution signal over visual representations, and we evaluate it across CheXagent-3B, LLaVA-Med, and LLaVA-1.6-7B. For architectures without explicit spatial feature maps, such attribution can in principle be obtained from cross-attention or gradient-based saliency over visual tokens. Evaluation spans radiology VQA (MIMIC-Diff-VQA and VQA-RAD), chest X-ray report generation (MIMIC-CXR), and gastrointestinal endoscopy (Gut-VLM), with consistent improvements over decoding-time and preference-optimization baselines. Our **contributions**: (1) we empirically characterize a cross-view *stability gap* between visually grounded and misgrounded tokens under clinically constrained perturbations (§2); (2) we introduce SPPO and stability-aware contrastive decoding, which exploit this signal at response and token levels under a shared counterfactual mechanism; and (3) we demonstrate consistent gains across the evaluated clinical settings, including improved MIMIC-CXR entity-level recall and an expert-confirmed

reduction in hallucination rate from 17.0% to 8.5%.

2 Hallucinations in Medical VLMs

Hypothesis. *Medical VLMs learn strong co-occurrence priors during pre-training: tokens such as “effusion” or “cardiomegaly” become statistically associated with common radiographic contexts regardless of whether the specific image supports them. We hypothesize that hallucinated tokens, being driven by these contextual priors rather than lesion-level evidence, should be disproportionately sensitive to perturbations of clinically irrelevant image regions. Grounded tokens, by contrast, should remain stable because their logit mass derives from visual features that the perturbation leaves intact.*

Experimental verification. To test this, we sample 500 image–question pairs from MIMIC-Diff-VQA and generate answers with greedy decoding using CheXagent. For each generated token y_t , we compute the logit shift $|\Delta\ell| = |\ell_\theta(y_t | v, x_{<t}) - \ell_\theta(y_t | v', x_{<t})|$ across five semantics-preserving counterfactual views v' , constructed via background blur and windowing shifts constrained to $s_{\text{global}} \geq 0.98$ (details in §3.2). We label tokens as hallucinated or grounded using GREEN (Ostmeier et al., 2024). Hallucinated tokens exhibit mean $|\Delta\ell| = 1.84$ versus 0.31 for grounded tokens ($p < 10^{-6}$, Welch’s t -test), confirming a clear *stability gap* (Figure 1, top). The effect is consistent across question types (yes/no, open-ended, difference) and persists on LLaVA-Med (mean $|\Delta\ell|$: 2.11 hallucinated vs. 0.38 grounded), showing that the stability-gap pattern persists across both evaluated backbones. This gap provides two actionable signals: at the *response level*, hallucinated responses show larger aggregate likelihood divergence across views, enabling preference-based training; at the *token level*, individual unstable tokens can be penalized during decoding. Prior work addresses complementary failure modes—reasoning-level hallucinations (Dong et al., 2025; Xu et al., 2025), representation amplification (Yin et al., 2025), and verifier-based resampling (Wu et al., 2025). These works address complementary signals. PCCD (Mahapatra et al., 2026) introduced proximity-constrained counterfactuals but operates only at inference, whereas LENS uses clinically constrained cross-view stability jointly for preference training and decoding. We adopt evaluation frameworks from VASE (Liao et al., 2025), HalLoc

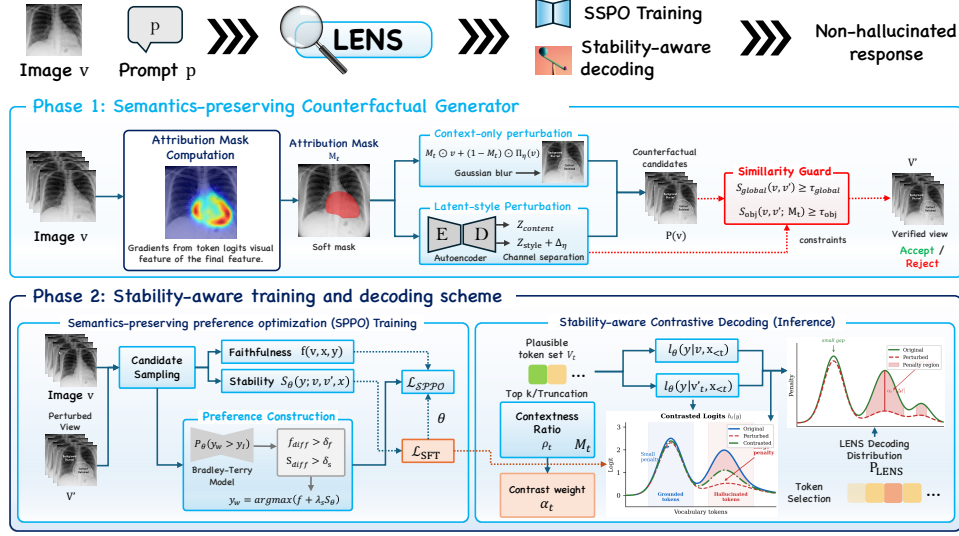


Figure 2: **LENS framework overview.** *Phase 1:* Semantics-preserving counterfactual generation via context-only (Eq. 2) and latent-style (Eq. 3) perturbations, filtered by proximity guards (Eq. 1). *Phase 2:* SPPO training (left) and stability-aware contrastive decoding (right).

(Park et al., 2025), and Gut-VLM (Khanal et al., 2025).

3 LENS: Method

The stability gap identified in §2 provides two complementary signals: at the response level, hallucinated outputs exhibit larger aggregate likelihood divergence across views, enabling preference-based training; at the token level, individual unstable tokens can be identified and penalized during decoding. LENS operationalizes both through three components: (i) a semantics-preserving counterfactual generator that produces constrained alternative views (§3.2), (ii) Semantics-Preserving Preference Optimization (SPPO) that uses cross-view stability as a training signal (§3.3), and (iii) stability-aware contrastive decoding that penalizes unstable tokens at inference (§3.4). Figure 2 provides an overview.

3.1 Problem Setup

An auto-regressive VLM $p_\theta(y | v, x)$ generates a response $y = (y_1, \dots, y_T)$ conditioned on image v and prompt x . A token y_t is *hallucinated* if it asserts a clinical finding unsupported by v . The goal of LENS is to modify both training and decoding so that the model favours responses whose token probabilities remain stable under clinically plausible perturbations of the input image—directly leveraging the stability gap as an operational criterion.

3.2 Semantics-Preserving Counterfactuals

The effectiveness of both SPPO and stability-aware decoding depends on counterfactual views that perturb nuisance factors—background context, acquisition style, windowing—while preserving lesion- and device-level evidence. Poorly constrained perturbations would conflate grounded and hallucinated tokens, collapsing the stability gap. We therefore impose two layers of constraint: proximity guards that enforce feature-space similarity, and attribution-guided masking that restricts perturbations to clinically irrelevant regions.

Proximity guards. Let $\phi_\theta(v)$ denote the image encoder embedding. We require every candidate view v' to satisfy

$$s_{\text{global}}(v, v') \geq \tau_{\text{global}}, \quad s_{\text{obj}}(v, v'; M_t) \geq \tau_{\text{obj}}, \quad (1)$$

where s_{global} and s_{obj} are cosine similarities computed in the global and attribution-masked feature spaces, respectively. The global constraint limits overall representation drift, while the object-level constraint protects localized findings. We set $\tau_{\text{global}}=0.98$ and $\tau_{\text{obj}}=0.95$ on validation data; performance is robust across $[0.95, 0.99]$ for both thresholds (§4.3). Views violating either constraint are rejected (Figure 3).

Attribution mask M_t . We compute $M_t \in [0, 1]^{H \times W}$ via Grad-CAM (Selvaraju et al., 2017) on the final vision encoder block, using the gradient of $\ell_\theta(y_t | v, x_{<t})$ with respect to visual feature

maps. The resulting saliency map is min-max normalized to $[0, 1]$ and serves as a soft mask highlighting clinically salient regions. This mask plays a dual role: it guides where perturbations may be applied (Eq. 2) and determines the adaptive contrast weight during decoding (§3.4). In ablations (§4.3), attention-based attribution yields comparable results, and performance degrades gracefully under 20% random noise injection.

Context-only perturbation. Using M_t , we restrict edits to the complement of the salient region:

$$v'_{\text{ctx}} = M_t \odot v + (1 - M_t) \odot \Pi_\eta(v), \quad (2)$$

where Π_η is a low-amplitude operator—Gaussian blur ($\sigma=1.5$) or brightness shift ($\pm 10\%$)—chosen so that (1) holds. This limits edits in salient regions while perturbing their complement.

Latent-style perturbation. To capture acquisition-level variation that pixel-space operators cannot cleanly isolate, we use a convolutional autoencoder (E, D) pre-trained on domain-matched images. We encode $z = E(v)$ and partition channels into content and style based on activation variance under photometric shifts (channels above the 75th percentile form z^{style} , typically 20–30%). We perturb only style:

$$v'_{\text{style}} = D(z^{\text{content}}, z^{\text{style}} + \Delta_\eta), \quad (3)$$

where $\Delta_\eta \sim \mathcal{N}(0, \sigma_\eta^2 I)$ with $\sigma_\eta=0.1$, accepted only if (1) holds. The full set of accepted views is $\mathcal{P}(v)$.

3.3 Semantics-Preserving Preference Optimization

With a reliable set of counterfactual views $\mathcal{P}(v)$ in hand, we can now use the stability gap as a training signal. Given image v , an accepted view $v' \in \mathcal{P}(v)$, and prompt x , we define the *cross-view stability* of a response y as

$$S_\theta(y; v, v', x) = - \left| \log p_\theta(y | v, x) - \log p_\theta(y | v', x) \right|. \quad (4)$$

Values of S_θ near zero indicate that the response probability is robust to the perturbation, consistent with visual grounding. Large negative values signal reliance on contextual priors—precisely the hallucination mechanism identified in §2.

Preference construction. From a supervised dataset $\mathcal{D}_0 = \{(v, x, y^*)\}$ equipped with a hallucination-sensitive scorer $f(v, x, y)$ (GREEN for radiology, RadGraph F1 for CXR reporting, QAAS for endoscopy), we sample two candidate responses: $\tilde{y}^{\text{orig}} \sim p_\theta(\cdot | v, x)$ from the original view and $\tilde{y}^{\text{cf}} \sim p_\theta(\cdot | v', x)$ from the counterfactual. The winner is selected by a composite score:

$$y_w = \arg \max_{y \in \{\tilde{y}^{\text{orig}}, \tilde{y}^{\text{cf}}\}} \left(f(v, x, y) + \lambda_S S_\theta(y; v, v', x) \right). \quad (5)$$

with y_ℓ as the loser. This scoring favors responses that are both clinically faithful and cross-view stable. To mitigate self-reward bias when both candidates are poor, we discard pairs where the faithfulness gap $|f^{\text{orig}} - f^{\text{cf}}| < \delta_f = 0.05$ or the stability gap $|S^{\text{orig}} - S^{\text{cf}}| < \delta_s = 0.1$, filtering approximately 25% of candidates in practice.

SPPO objective. We define a stability-adjusted reward that explicitly penalizes cross-view instability:

$$r_\theta(y) = \log p_\theta(y | v, x) - \gamma \left| \log p_\theta(y | v, x) - \log p_\theta(y | v', x) \right|. \quad (6)$$

where $\gamma \geq 0$ controls the penalty strength ($\gamma=0.3$ by default; ablation in Table 6). The preference probability follows a Bradley–Terry model:

$$P_\theta(y_w \succ y_\ell) = \sigma(\beta(r_\theta(y_w) - r_\theta(y_\ell))). \quad (7)$$

The full SPPO loss combines this preference objective with supervised fine-tuning to prevent catastrophic forgetting:

$$\mathcal{L}(\theta) = \underbrace{-\mathbb{E}[\log P_\theta(y_w \succ y_\ell)]}_{\mathcal{L}_{\text{SPPO}}} + \lambda_{\text{SFT}} \mathcal{L}_{\text{SFT}}(\theta). \quad (8)$$

3.4 Stability-Aware Contrastive Decoding

SPPO aligns the model toward stable responses during training, but cannot guarantee stability for every token at inference on unseen inputs. We therefore complement SPPO with a decoding-time mechanism that directly penalizes tokens exhibiting the instability signature identified in §2.

Plausible set and contextness. At decoding step t , we first restrict to a plausible token set via adaptive truncation: $\mathcal{V}_t = \{y : p_\theta(y | v, x_{<t}) \geq$

$\gamma_{\text{top}} \max_{y'} p_{\theta}(y' \mid v, x_{<t})$, with $\gamma_{\text{top}}=0.1$. We then compute how much of the current token’s evidence comes from context versus salient regions. Using the attribution mask M_t , we define a *contextness ratio* $\rho_t = 1 - \frac{1}{HW} \sum_{i,j} M_t(i, j)$: high ρ_t indicates the token depends primarily on background context and should receive stronger contrastive correction. The contrast weight is set as $\alpha_t = \text{clip}(\alpha_{\min} + \kappa\rho_t, \alpha_{\min}, \alpha_{\max})$, with $\alpha_{\min}=0.1$, $\alpha_{\max}=1.0$, and $\kappa=0.8$.

Contrasted logits. We select a counterfactual view $v'_t \in \mathcal{P}(v)$ satisfying (1) and compute the stability-penalized logit:

$$h_t(y) = \ell_{\theta}(y \mid v, x_{<t}) - \alpha_t \left| \ell_{\theta}(y \mid v, x_{<t}) - \ell_{\theta}(y \mid v'_t, x_{<t}) \right|. \quad (9)$$

$$p_{\text{LENS}}(y \mid v, x_{<t}) \propto \exp(h_t(y)), \quad y \in \mathcal{V}_t. \quad (10)$$

The intuition is direct: tokens with large logit divergence—the hallucination signature from §2—receive a proportionally larger penalty, while stable, grounded tokens pass through largely unchanged. The adaptive α_t prevents over-suppression when the token is primarily evidence-driven (ρ_t low). If no valid counterfactual view exists, we set $\alpha_t = 0$ and fall back to standard decoding at that step.

4 Experiments

We evaluate LENS across three clinical settings to answer four questions: (1) Does LENS improve over both decoding-time and training-time baselines? (2) Are gains consistent across backbones with different pre-training strengths? (3) Do improvements on hallucination-sensitive metrics come at the cost of language quality? (4) Which components and design choices drive the gains?

4.1 Experimental Setup

Benchmarks and metrics. We evaluate on standard train/test splits of MIMIC-Diff-VQA (Hu et al., 2023) ($\sim 700\text{k}$ QA pairs; test $\sim 13\text{k}$), VQA-RAD (Lau et al., 2018) (2.2k QA; test 451), MIMIC-CXR-JPG (Johnson et al., 2019) (377k images), and Gut-VLM (Khanal et al., 2025) (1,816 images; 1,450/366 train/test) for radiology VQA (AUC, AUG under GREEN/VASE (Ostmeier et al., 2024; Liao et al., 2025)), CXR reporting (RadGraph F1 (Jain et al., 2021), GREEN), and GI endoscopy (QAAS, R-Sim, ROUGE-L) respectively.

Table 1: **Main results.** Best **bold**, second underlined. ‘–’ = not applicable. BS[†] = BERTScore F1 (not used in training).

		Radiology VQA				Indep.
		MIMIC-Diff-VQA		VQA-RAD		
Cat.	Method	AUC	AUG	AUC	AUG	
Post-hoc	Cross-Check.	68.45	48.98	70.51	38.99	–
	RadFlag	85.77	61.01	74.83	38.52	–
	Sem. Entr.	86.51	59.82	73.01	38.36	–
	VASE	87.79	62.06	76.54	39.93	–
Greedy	CheXagent-3B	86.20	60.10	74.80	39.10	.862
Dec.-Time	VCD	87.02	61.25	75.90	39.85	.869
	HALC	87.31	61.44	–	–	–
	VASparse	87.82	62.08	–	–	–
	PCCD	<u>89.59</u>	<u>64.06</u>	<u>78.60</u>	<u>41.20</u>	.873
Train. PO	DPO	90.12	64.81	78.80	41.40	.870
	PPO	89.85	64.52	78.20	41.05	.868
	GRPO	90.38	65.04	79.10	41.60	.871
Ours	LENS	92.21	66.86	80.60	43.30	.875
	Δ PCCD	+2.6	+2.8	+2.0	+2.1	+0.02
	Δ best PO	+1.8	+1.8	+1.5	+1.7	+0.04

		CXR Report		GI Endoscopy		
		MIMIC-CXR		Gut-VLM		
Cat.	Method	RG-F1	GREEN	QAAS	R-Sim	R-L
Rep. Gen.	RaDialog-7B	16.3	24.4	–	–	–
	RadVLM-7B	17.7	24.9	–	–	–
Greedy	CheXagent-3B	20.9	28.5	–	–	–
	LLaVA-1.6 _{7B}	–	–	90.89	3.96	.89
Dec.-Time	VCD	21.3	29.1	91.20	4.07	.90
	PCCD	<u>23.1</u>	<u>30.8</u>	<u>92.70</u>	<u>4.19</u>	<u>.91</u>
Train. PO	DPO	23.4	31.0	92.10	4.11	.90
	PPO	23.0	30.6	91.80	4.08	.90
	GRPO	23.6	31.3	92.45	4.15	.91
Ours	LENS	25.3	33.2	94.80	4.32	.92
	Δ PCCD	+2.2	+2.4	+2.1	+0.13	+0.01
	Δ best PO	+1.7	+1.9	+2.4	+0.17	+0.01

We additionally report BERTScore F1 as an independent metric never used in training.

Backbones and baselines. We use CheXagent-3B (Chen et al., 2024b) for radiology/CXR and LLaVA-1.6-7B_{7B} for GI endoscopy. Baselines span post-hoc detectors (Cross-Checking, RadFlag (Zhang et al., 2025), Semantic Entropy, VASE (Liao et al., 2025)), decoding-time methods (VCD (Leng et al., 2024), HALC (Chen et al., 2024a), VASparse (Zhuang et al., 2025), PCCD (Mahapatra et al., 2026)), training-time PO (DPO (Fu et al., 2025), PPO (Xie et al., 2024), GRPO (Liu et al., 2025)), and report generators (RaDialog-7B, RadVLM-7B). DPO, PPO, and GRPO use the same preference data as SPPO without stability scoring.

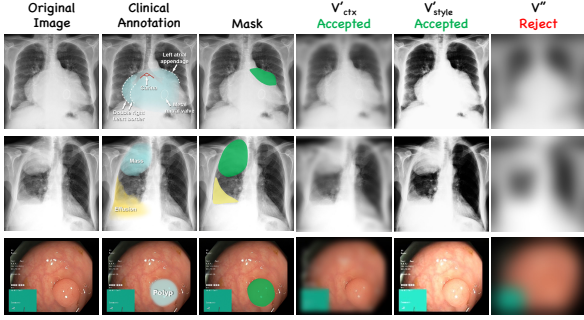


Figure 3: **Counterfactual view gallery.** Rows span chest radiography (1–2) and GI endoscopy (3).

Implementation. SPPO freezes the visual encoder and finetunes language and projection layers with AdamW ($\text{lr } 2 \times 10^{-5}$, weight decay 10^{-2} , $\beta=0.1$, $\lambda_{\text{SFT}}=0.1$) for 3 epochs on $8 \times \text{A100 GPUs}$ (12–18h wall-clock time). We sample $K_{\text{cand}}=4$ candidate counterfactual views per image with $\lambda_S=0.3$. At inference, we use $K_{\text{dec}}=16$ candidate tokens per step, $T_{\text{max}}=256$ for report generation and 64 for VQA. All hyperparameters are fixed across benchmarks.

4.2 Main Results

Table 1 presents results across all benchmarks. We organize discussion around the four questions posed at the start of section 4.

(Q1 and Q2) Improvement over baselines and backbone generalization. LENS achieves the best performance on every metric (Table 1), improving over PCCD by 2.0–2.6 AUC on radiology VQA, 2.2 RadGraph F1 on CXR reporting, and 2.1 QAAS on GI endoscopy, while outperforming the best PO baseline by 1.5–2.4 points on the corresponding primary metrics. Standard PO methods yield only modest gains over PCCD, whereas LENS provides a larger margin. On LLaVA-Med (Table 2), gains over PCCD are comparable to or larger than on CheXagent: +2.3 AUC on MIMIC-Diff-VQA and +2.2 on VQA-RAD, versus +2.3 and +2.0, respectively. This shows that the benefit of cross-view stability transfers across the evaluated backbones.

(Q3) Independent quality and entity recovery. Table 3 reports independent semantic and lexical quality metrics not used in preference construction. LENS maintains or improves these metrics across domains. On MIMIC-CXR, compared with PCCD, LENS improves BERTScore from 0.864 to 0.866, CheXpert F1 from 0.445 to 0.463, BLEU

Table 2: **LLaVA-Med backbone** on Rad-VQA. Gains over PCCD are comparable to or larger than on CheXagent (Table 1).

Method	MIMIC-Diff-VQA		VQA-RAD	
	AUC	AUG	AUC	AUG
Cross-Checking	60.31	44.23	59.56	38.99
RadFlag	73.62	54.23	61.21	36.33
Sem. Entropy	74.76	53.66	63.55	38.88
VASE	75.86	55.58	66.72	40.47
PCCD	77.70	57.09	68.90	41.90
LENS	80.00	59.29	71.10	43.90

Table 3: Independent quality metrics and entity-level recovery. Panel (a) reports semantic and lexical metrics not used in preference construction. Panel (b) decomposes RadGraph F1 into entity-level precision and recall on MIMIC-CXR. [†]Independent metrics. Best results are shown in bold.

(a) Independent quality metrics							
Method	Primary metric		Independent [†]				
	Metric	Value	BERTScore	CheXpert F1	BLEU	R-L	MET
<i>MIMIC-Diff-VQA (CheXagent)</i>							
Greedy	AUC	86.20	0.871	–	–	–	–
PCCD	AUC	89.59	0.873	–	–	–	–
LENS	AUC	92.21	0.875	–	–	–	–
<i>MIMIC-CXR (CheXagent)</i>							
Greedy	RG-F1	20.9	0.862	0.421	0.21	0.32	0.25
PCCD	RG-F1	23.1	0.864	0.445	0.23	0.35	0.27
LENS	RG-F1	25.3	0.866	0.463	0.26	0.37	0.30
<i>Gut-VLM (LLaVA-1.6-7B_{phi})</i>							
Greedy	QAAS	90.89	0.884	–	0.82	0.89	0.90
PCCD	QAAS	92.70	0.886	–	0.84	0.91	0.93
LENS	QAAS	94.80	0.889	–	0.85	0.92	0.94
(b) MIMIC-CXR entity-level RadGraph							
Method	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$				
Greedy	22.8	19.3	20.9				
PCCD	25.6	21.0	23.1				
LENS	28.4	22.8	25.3				

from 0.23 to 0.26, ROUGE-L from 0.35 to 0.37, and METEOR from 0.27 to 0.30. We additionally decompose RadGraph F1 into entity-level precision and recall (Table 3(b)). LENS improves precision from 25.6 to 28.4 and recall from 21.0 to 22.8 over PCCD. The simultaneous improvement in lexical quality, clinical entity precision, and recall indicates that the hallucination reduction is not explained solely by more conservative reporting or uniform suppression of predicted findings.

Human evaluation. To validate automatic metrics against clinical judgment, two board-certified gastroenterologists and two senior trainees independently rated 100 image descriptions and 80 VQA responses from Gut-VLM (Table 4). LENS reduces the clinician-assessed hallucination rate to 8.5% (vs. 10.9% for PCCD and 17.0% for greedy decoding), raises clinical adequacy to 4.25/5, and achieves 92.3% VQA correctness. In pairwise preference, LENS is preferred over VCD 75% of times.

Table 4: **Human evaluation on Gut-VLM** (LLaVA-1.6-7B_{ftH}), rated by 2 board-certified gastroenterologists and 2 senior trainees.

Method	Description ($n=100$)		VQA ($n=80$)	
	Halluc. (%) $\downarrow$	Adequacy $\uparrow$	Corr. (%) $\uparrow$	Pref. vs. VCD $\uparrow$
Greedy	17.0	3.80	84.0	—
VCD	13.0	3.96	86.3	50.0
PCCD	10.9	4.12	90.0	66.2
LENS (ours)	8.5	4.25	92.3	75.0

4.3 Ablation and Analysis

We analyze LENS across four analytical dimensions (Table 5), with all experiments conducted using the CheXagent backbone.

(a) Component complementarity. SPPO training alone improves over PCCD by 0.9 AUC, while LENS decoding without SPPO improves by 1.7. Their combination yields +2.6, outperforming either component alone and confirming complementarity: SPPO reduces the *frequency* of unstable tokens while stability-aware decoding catches residual instabilities at inference.

(b) Design choice. Replacing adaptive α_t with fixed constant ($\alpha=0.5$) costs 1.2 AUC, confirming contextness-aware weighting prevents over-suppression of grounded tokens that happen to depend on small salient regions. Dropping object-level similarity guard (s_{obj}) costs 1.7 AUC, with largest degradation on MIMIC-CXR where findings such as small nodules and tube tips occupy minimal image area- precisely setting where global similarity alone cannot detect evidence destruction. Removing preference filtering (§3.3) degrades by 0.7, supporting its role in mitigating self-reward bias when both candidate responses are poor.

(c) Attribution robustness. A practical concern is whether LENS depends critically on the quality of the attribution mask M_t . Replacing Grad-CAM with attention-based attribution reduces AUC by only 0.3 points. Injecting 10% and 20% uniform random noise into M_t degrades AUC by 0.4 and 0.9 respectively—a graceful degradation. This robustness arises because the proximity guards (§3.2) act as a safety net: even with a noisy mask, views that destroy clinical evidence are rejected by the feature-space constraints.

(d) Computational overhead. LENS adds 6 ms/sample over PCCD at inference (74 vs. 68 ms on A100, $T_{\text{max}}=64$), a $1.09\times$ overhead. Both methods require one additional forward pass per de-

Table 5: **Ablation studies and computational overhead** on CheXagent. Each component and design choice contributes measurably. LENS adds only $1.09\times$ latency over PCCD at inference.

Variant	MIMIC-Diff-VQA		MIMIC-CXR	
	AUC	AUG	GREEN	RG-F1
<i>(a) Component decomposition</i>				
Greedy (no mitigation)	86.20	60.10	28.5	20.9
+ PCCD decoding only	89.59	64.06	30.8	23.1
+ SPPO training only	90.50	65.10	31.4	23.8
+ LENS decoding w/o SPPO	91.30	65.90	32.0	24.4
Full LENS (SPPO + dec.)	92.21	66.86	33.2	25.3
<i>(b) Stability weighting and constraints</i>				
Fixed $\alpha_t = 0.5$	91.05	65.62	31.8	24.1
No adaptive α_t ($\kappa=0$)	90.82	65.38	31.5	23.9
Global sim. only (no s_{obj})	90.48	65.01	31.0	23.5
No preference filtering	91.54	66.10	32.4	24.6
Full LENS	92.21	66.86	33.2	25.3
<i>(c) Attribution robustness</i>				
Grad-CAM (default)	92.21	66.86	33.2	25.3
Attention-based M_t	91.88	66.52	32.9	25.0
M_t + 10% noise	91.79	66.41	32.8	24.9
M_t + 20% noise	91.31	65.98	32.5	24.6
<i>(d) Computational overhead ($A100$, $T_{\text{max}}=64$)</i>				
	Latency		Peak mem	
	ms	Rel.	GB	
Greedy	42	$1.00\times$	8.2	
PCCD	68	$1.62\times$	13.6	
LENS	74	$1.76\times$	14.1	

coding step for the counterfactual view. Peak memory increases marginally from 13.6 to 14.1 GB. The SPPO training phase is a one-time cost of 12–18h on $8\times A100$ GPUs, across all subsequent inference.

Stability penalty strength (γ ablation). Table 6 examines how the stability penalty weight γ interacts with different training objectives. Three findings emerge. First, SPPO at $\gamma=0$ (faithfulness scoring only, no stability) already outperforms standard DPO/PPO/GRPO by 0.1–0.7 AUC, attributable to the cross-view preference construction generating higher-quality pairs. Second, moderate stability penalty ($\gamma \in [0.1, 0.3]$) yields the best results across all metrics, with $\gamma=0.3$ optimal. Third, excessive penalty ($\gamma=1.0$) degrades performance below $\gamma=0$, consistent with the over-conservatism failure mode discussed in §4.4: too-strong penalization suppresses borderline findings alongside genuine hallucinations.

Threshold sensitivity. Figure 4 and Table 7 show AUC varies by <0.4 points across $[0.95, 0.99]$ for both τ_{global} and τ_{obj} , with degradation only below 0.95. The default ($\tau_{\text{global}}=0.98$, $\tau_{\text{obj}}=0.95$) is used without modification across all experiments.

Table 6: **PO baselines and γ ablation** (CheXagent, with LENS decoding). Moderate stability penalty ($\gamma \in [0.1, 0.3]$) is optimal; excessive penalty ($\gamma=1.0$) induces over-conservatism.

Method	MIMIC-Diff-VQA		MIMIC-CXR		Gut-VLM	
	AUC	AUG	GREEN	RG-F1	QAAS	R-Sim
<i>No training (decoding only)</i>						
Greedy	86.20	60.10	28.5	20.9	90.89	3.96
+ PCCD decoding	89.59	64.06	30.8	23.1	92.70	4.19
<i>Standard PO + LENS decoding</i>						
DPO	90.12	64.81	31.0	23.4	92.10	4.11
PPO	89.85	64.52	30.6	23.0	91.80	4.08
GRPO	90.38	65.04	31.3	23.6	92.45	4.15
<i>SPPO γ sweep + LENS decoding</i>						
$\gamma=0$ (no stability)	90.50	65.10	31.4	23.8	92.60	4.17
$\gamma=0.1$	91.02	65.68	32.0	24.3	93.40	4.22
$\gamma=0.3$ (default)	<u>91.40</u>	<u>66.12</u>	<u>32.5</u>	<u>24.8</u>	<u>93.90</u>	<u>4.26</u>
$\gamma=0.5$	91.28	65.95	32.3	24.5	93.60	4.24
$\gamma=1.0$	90.65	65.22	31.6	23.9	92.80	4.18
Full LENS	92.21	66.86	33.2	25.3	94.80	4.32

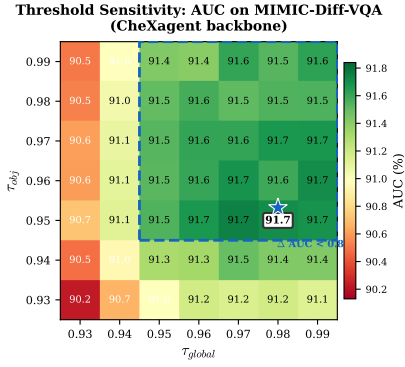


Figure 4: **Threshold sensitivity.** AUC on MIMIC-Diff-VQA (CheXagent backbone) as τ_{global} and τ_{obj} vary. Performance is stable across $[0.95, 0.99]$ (<0.4 pt variation).

Full Gut-VLM metrics. Table 8 reports the complete set of Gut-VLM description metrics including lexical overlap. LENS improves across all metrics: BLEU (+0.01 over PCCD), METEOR (+0.01), and R-Sim (+0.13), alongside the headline QAAS gain (+2.1). The consistent improvement across both semantic (QAAS, R-Sim) and lexical (BLEU, ROUGE-L) metrics indicates that LENS does not merely suppress hallucinated tokens but produces descriptions that are closer to expert reference text at all granularities.

False-negative analysis. To assess whether hallucination reduction is obtained at the expense of missed findings, we further analyze entity-level recall on MIMIC-CXR. Overall, LENS improves both RadGraph precision and recall over PCCD, from 25.6/21.0 to 28.4/22.8, respectively, indicating that its gains are not explained by uniform under-reporting. We additionally stratify recall

Input	GT	Greedy	VCD	SPCD	LENS (Ours)
Case 1: Device in CXR		Sternal wires and prosthetic valve present. No chest tube identified.	Sternal wires present. Chest tube noted in the right hemithorax.	Sternal wires intact, prosthetic valve seen. Possible chest tube.	Sternal wires intact, prosthetic valve present. No chest tube visualized.
Case 2: Normal CXR		Normal cardiac silhouette. Lungs are clear. No acute abnormality.	Mild cardiomegaly. Small bilateral pleural effusions, consider CHF.	Heart size is borderline enlarged. Lungs are clear.	Normal heart, clear lungs seen. Atelectasis at the left base.
Case 3: Multiple polyps in GI-Endoscopy		Two sessile polyps in the ascending colon. Normal surrounding mucosa, no bleeding.	Two polyps are visible. Erythematous mucosa present.	Polyp visible, erythematous mucosa found surrounding the polyp.	Two sessile polyps are present. Erythematous mucosa seen in the colon wall.
Case 4: Limitation of LENS		Subtle left basilar opacity, representing consolidation.	Moderate pleural effusion suspected. No critical findings.	Lung airspace seems normal.	Hazy opacity on the left lung field.

Figure 5: **Qualitative comparison across methods.** Green text: clinically correct. Red italic: hallucinated content. Orange text (case 4): over-conservative miss. Cases 1–3 show LENS eliminating hallucinations that persist under VCD and PCCD. Case 4 illustrates the specificity–sensitivity trade-off for borderline findings.

Table 7: **Threshold grid** (companion to Figure 4). AUC on MIMIC-Diff-VQA as proximity guard thresholds vary. Bold marks the default configuration.

τ_{obj}	τ_{global}				
	0.95	0.96	0.97	0.98	0.99
0.99	91.42	91.43	91.55	91.52	91.55
0.98	91.48	91.56	91.54	91.52	91.50
0.97	91.50	91.58	91.62	91.68	91.65
0.96	91.51	91.60	91.65	91.62	91.66
0.95	91.52	91.67	91.68	91.70	91.66

across clinically challenging subsets involving subtle opacity, small pleural effusion, support-device positioning, and low-confidence abnormalities. As shown in Table 9, LENS improves recall over PCCD across all four subsets. Nevertheless, individual borderline findings can still be missed, as illustrated by the subtle-opacity failure case in Figure 5, indicating that aggregate recall improvements do not eliminate condition-specific false-negative risk.

4.4 Qualitative Analysis

Figure 5 presents four representative cases spanning all evaluation domains. In Case 1, greedy decoding invents a nonexistent chest tube; VCD hedges; LENS correctly excludes it while fully describing sternal wires and the prosthetic valve. In Case 2, greedy fabricates cardiomegaly with effusions—a high-risk false positive—whereas LENS correctly reports normal findings. In Case 3, greedy hallucinates erythematous mucosa; LENS accurately describes sessile polyps with normal surrounding mucosa. Case 4 illustrates an important failure mode: for a subtle left basilar opacity, LENS produces an over-conservative response and misses the borderline finding. This example shows that individual subtle abnormalities can still

Table 8: **Full Gut-VLM description metrics.** LENS improves across both semantic (QAAS, R-Sim) and lexical (BLEU, METEOR, ROUGE-L) measures.

Model	R-L	BLEU	MET	R-Sim	QAAS
Qwen-7B _{HH}	.88	.82	.90	4.04	90.53
LLaVA-1.6-7B _{HH}	.89	.82	.90	3.96	90.89
VCD	.90	.82	.91	4.07	91.20
PCCD	.91	.84	.93	4.19	92.70
LENS	.92	.85	.94	4.32	94.80

Subset	Greedy	PCCD	LENS
Overall entity recall	19.3	21.0	22.8
Subtle opacity	18.1	20.8	21.9
Small pleural effusion	19.1	21.4	23.1
Tube/line positioning	22.8	24.9	27.2
Low-confidence abnormality	17.2	19.5	21.3

Table 9: Entity-level recall (%) on MIMIC-CXR, including clinically challenging borderline-finding subsets.

be suppressed even though the condition-stratified analysis in Table 9 shows improved recall over PCCD across all four evaluated subsets. The result therefore reflects a residual specificity–sensitivity trade-off rather than a systematic reduction in recall.

5 Discussion and Conclusion

LENS demonstrates that cross-view stability is an effective signal for mitigating visually misgrounded hallucinations. Adaptive contrast weighting and object-level proximity guards are important for avoiding unnecessary suppression of grounded tokens and protecting small findings (Table 5). The method is complementary to preference alignment and can also be applied as an inference-time correction.

Borderline findings remain challenging: although LENS improves aggregate and subset-level recall, individual subtle abnormalities can still be missed (Figure 5). The parameters γ and $\alpha_{\max}$ can shift this specificity–sensitivity operating point but do not eliminate case-level false-negative risk.

Conclusion. LENS uses cross-view stability jointly for preference optimization and inference-time decoding. Across radiology VQA, CXR reporting, and GI endoscopy, it consistently outperforms strong decoding-time and preference-optimization baselines. Expert evaluation reduces hallucination rate from 17.0% to 8.5%, while MIMIC-CXR entity recall improves from 21.0 to 22.8 over PCCD.

Limitations

LENS primarily targets visually misgrounded errors whose predictions change under visual counterfactuals; purely language-driven hallucinations may remain stable and require complementary mechanisms. Borderline findings can also yield ambiguous stability signals: despite improved aggregate recall, individual low-contrast abnormalities may still be missed. Finally, evaluation is limited to 2D chest radiography and GI endoscopy in English; generalization to other modalities, institutions, scanners, and multilingual settings remains to be established.

Acknowledgments: This research was funded by Khalifa University of Science and Technology through the Faculty Startup grant under Project ID: KU-INT-FSU-2026-8471000024.

References

- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. [Hallucination of multimodal large language models: A survey](#). *arXiv preprint arXiv:2404.18930*.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024a. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Blüthgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jena Jitsev, and 4 others. 2024b. [A vision-language foundation model to enhance efficiency of chest x-ray interpretation](#). *arXiv preprint arXiv:2401.12208*.
- Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. 2025. [MI-RAGE: Assessing hallucination in multimodal reasoning chains of MLLM](#). In *Advances in Neural Information Processing Systems* 38.
- Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. 2025. [Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.
- Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada,

- Ronald M. Summers, and Yingying Zhu. 2023. [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*.
- Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2026. [Spotlight on token perception for multimodal reinforcement learning](#). In *The Fourteenth International Conference on Learning Representations*.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. HACL: Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. [MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs](#). *arXiv preprint arXiv:1901.07042*.
- Bidur Khanal, Sandesh Pokhrel, Sanjay Bhandari, Ramesh Rana, Nikesh Shrestha, Ram B. Gurung, Cristian Linte, Angus Watson, Yash R. Shrestha, and Binod Bhattarai. 2025. [Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 235–245. Springer.
- Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. [A dataset of clinically generated visual questions and answers about radiology images](#). *Scientific Data*, 5:180251.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*.
- Chunyan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*.
- Zehui Liao, Shishuai Hu, Ke Zou, Huazhu Fu, Liangli Zhen, and Yong Xia. 2025. [Vision-amplified semantic entropy for hallucination detection in medical visual question answering](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*.
- Wenqi Liu, Xuemeng Song, Jiayi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. 2025. [Mitigating hallucination through theory-consistent symmetric multimodal preference optimization](#). In *Advances in Neural Information Processing Systems 38*.
- Dwarikanath Mahapatra, Abhijit Das, Manish Kumar Pandey, Joy Dhar, Sudipta Roy, Zongyuan Ge, Behzad Bozorgtabar, and Imran Razzak. 2026. [Proximity-constrained counterfactual decoding for hallucination-robust medical vqa](#). In *Computer Vision – ECCV 2026*.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Blüthgen, Arne Edward Michalson, Michael Moseley, Curtis P. Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.
- Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. [Halloc: Token-level localization of hallucinations for vision language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29893–29903.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.
- Tsung-Han Wu, Heekyoung Lee, Jiayin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. 2025. [Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling](#). In *Advances in Neural Information Processing Systems 38*.
- Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

- Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Y. Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. [More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models](#). In *Advances in Neural Information Processing Systems* 38.
- Hao Yin, Guangzong Si, and Zilei Wang. 2025. [Clear-sight: Visual signal enhancement for object hallucination mitigation in multimodal large language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14625–14634.
- Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian N. Acosta, L. John Fahrner, and Pranav Rajpurkar. 2025. [Radflag: A black-box hallucination detection method for medical vision language models](#). In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 1087–1103. PMLR.
- Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. 2025. [VASparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4189–4199.