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

**Anonymous ACL submission**


## Abstract

Medical VLMs frequently hallucinate clinically unsupported findings. Yet current hallucination mitigation strategies face a fundamental tension: decoding-time methods apply weakly constrained perturbations that destroy subtle lesion and device cues, while training-time alignment requires costly annotation without any mechanism to verify whether a response is genuinely grounded in visual evidence. We identify a simple signal that resolves this tension: *visually grounded tokens maintain stable logits under semantics-preserving image perturbations, whereas hallucinated tokens exhibit sharp logit shifts* —a phenomenon we term as *stability gap*. We introduce **LENS** (**L**ogit-**EN**hanced **S**tability), a framework that operationalizes this gap at both training and inference. During training, *Semantics-Preserving Preference Optimization* (SPPO) constructs preference pairs scored jointly on clinical faithfulness and cross-view stability, teaching the model to favour grounded responses. At inference, *stability-aware contrastive decoding* adaptively penalizes tokens whose logits diverge between original and counterfactual views. Both stages share tightly constrained counterfactual machinery: attribution-guided masking and feature-space proximity guards ensure that lesion- and device-level evidence is preserved. LENS improves hallucination-sensitive by 1.8–2.9 points over the latest strongest baselines. Additionally, an expert evaluation by board-certified gastroenterologists confirms a reduction in hallucination rate 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).

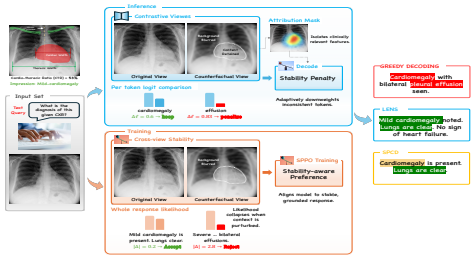


Figure 1: **The stability gap in medical VLM hallucinations.** *Top:* Per-token logit change $|\Delta\ell|$ under a semantics-preserving counterfactual view. *Bottom:* Response-level likelihood comparison.

*Decoding-time* methods contrast logits between original and perturbed images, but typically apply global perturbations that erase 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) reduces hallucinations through finetuning, but requires costly preference annotation and offers no mechanism to verify the response's truthfulness (Jiang et al., 2024; Xie et al., 2024; Fu et al., 2025). *Post-hoc detectors* flag hallucinated content after generation but cannot prevent it (Zhang et al., 2024; Park et al., 2025). A recent approach, Semantics-Preserving Contrastive Decoding (SPCD), partially addresses the perturbation quality problem by using proximity-constrained counterfactual views that preserve clinical evidence (Mahapatra et al., 2026). However, SPCD operates only at inference and cannot distinguish responses that are inherently stable from those whose apparent correctness is fragile.

These limitations share a common root: no existing method explicitly measures whether a model's output is *robust to clinically irrelevant visual vari-*

*ation*. We address this gap through a simple empirical observation. When a medical VLM generates a hallucinated token, its logit shifts sharply under perturbation of clinically irrelevant context—background texture, windowing, acquisition style. In contrast, tokens grounded in lesion- or device-level evidence remain 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 the *stability gap*, and it provides a direct, quantifiable signal for both training and decoding.

We propose **LENS** (**L**ogit-**EN**hanced **S**tability), which operationalizes the stability gap in two complementary stages. *Semantics-Preserving Preference Optimization* (SPPO) constructs preference pairs scored jointly on faithfulness and cross-view stability at training time. 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 share tightly constrained counterfactual machinery: attribution-guided masking preserves lesion and device regions, and feature-space proximity guards reject views that drift beyond safe bounds.

LENS is architecture-agnostic, requiring only access to vision-encoder features and token logits; we validate this across CheXagent-3B, LLaVA-Med, and LLaVA-1.6-7B. Evaluation on three clinical domains- radiology VQA (MIMIC-Diff-VQA, VQA-RAD), chest X-ray report generation (MIMIC-CXR), and gastrointestinal endoscopy (Gut-VLM)—shows consistent gains over both decoding-time and training-time baselines. In this paper, our **contributions** are: **(1)** We identify the *stability gap*: hallucinated tokens exhibit measurably larger logit shifts than grounded tokens under semantics-preserving perturbations, providing a novel training and decoding signal (§2); **(2)** we propose SPPO and a complementary stability-aware contrastive decoding rule; and **(3)** we demonstrate 1.8–2.9 point improvements across three clinical domains with expert-confirmed hallucination reduction 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 $|\Delta\ell| = |\ell_\theta(y_t \mid v, x_{<t}) - \ell_\theta(y_t \mid v', x_{<t})|$ 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), indicating that the gap is a property of the hallucination mechanism, not a backbone-specific artifact. 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)—but none exploits cross-view stability as a training signal. SPCD (Mahapatra et al., 2026) introduced proximity-constrained counterfactuals but operates only at inference. We adopt evaluation frameworks from VASE (Liao et al., 2025), HalLoc (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

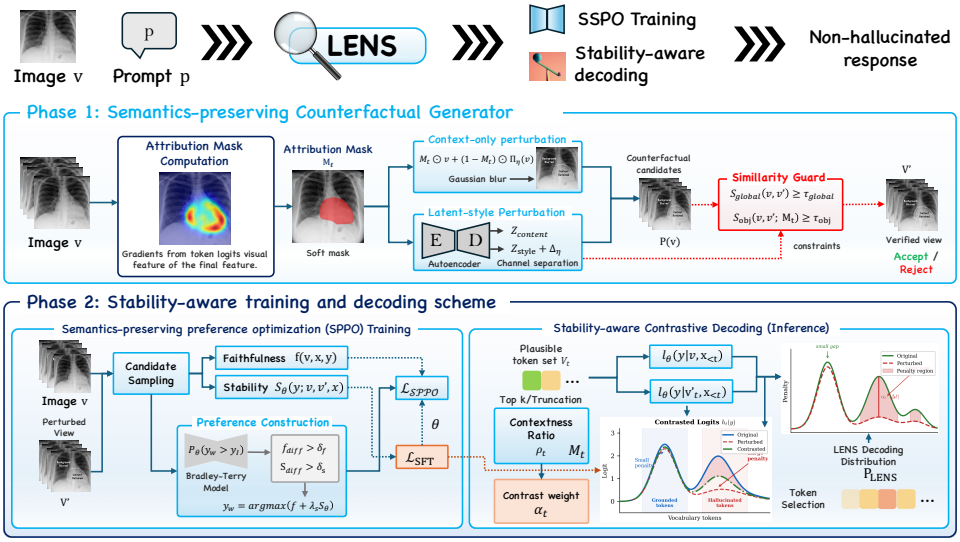


Figure 2: **LENS framework overview.** *Phase 1:* Semantics-preserving counterfactual generation via context-only (Eq. 2) and latent-style (Eq. 3) perturbations, filtered by proximity guards (Eq. 1). *Phase 2:* SPPO training (left) and stability-aware contrastive decoding (right).

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 clinically safe 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 $p_\theta(y \mid v, x)$ generates a response $y = (y_1, \ldots, y_T)$ 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

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

where $s_{\text{global}}$ and $s_{\text{obj}}$ are cosine similarities computed in the global and attribution-masked feature spaces, respectively. Global constraint ensures overall semantic preservation and object-level constraint protects findings that may occupy only a small image region (e.g., a pneumothorax apex). 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 $\ell_\theta(y_t \mid v, x_{<t})$ 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 $\Pi_\eta$ is a low-amplitude operator—Gaussian blur ($\sigma$=1.5) or brightness shift ($\pm$10%)—chosen so that (1) holds. This preserves all content within the attribution mask while altering only background context.

**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 $z^{\text{style}}$, typically 20–30%). We perturb only style:

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

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

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

Values of $S_\theta$ 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 $\mathcal{D}_0$=$\{(v, x, y^\star)\}$ 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: $\tilde{y}^{\text{orig}} \sim p_\theta(\cdot \mid v, x)$ from the original view and $\tilde{y}^{\text{cf}} \sim p_\theta(\cdot \mid v', x)$ from the counterfactual. The winner is selected by a composite score:

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

with $y_\ell$ as the loser. This scoring ensures that the preferred response is not only clinically faithful but also visually grounded, addressing the key limitation of standard preference optimization methods that optimize faithfulness without verifying grounding. 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:

$$r_\theta(y) = \log p_\theta(y \mid v, x) - \gamma\big|\log p_\theta(y \mid v, x) - \log p_\theta(y \mid v', x)\big|, \quad (6)$$

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:

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

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

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

### 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: $\mathcal{V}_t = \{y : p_\theta(y \mid v, x_{<t}) \geq \gamma_{\text{top}} \max_{y'} p_\theta(y' \mid v, x_{<t})\}$, 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 $\rho_t$ 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:

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

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

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 $\alpha_t$ prevents over-suppression when the token is primarily evidence-driven ($\rho_t$ low). If no valid counterfactual view exists for a given step, we set $\alpha_t$=0 and fall back to standard decoding, ensuring LENS never degrades below the baseline.

# 4 Experiments

We evaluate LENS across three clinical domains 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) (∼700k QA pairs; test ∼13k), 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. 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$_{\text{ftH}}$ for GI endoscopy. Baselines span post-hoc detectors (Cross-Checking, RadFlag (Zhang et al., 2024), 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), SPCD (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.

Table 1: **Main results.** Best **bold**, second underlined. '–' = not applicable. BS$^{\dagger}$ = BERTScore F1 (not used in training).

| | | Radiology VQA | | | | Indep. |
|---|---|---|---|---|---|---|
| | | MIMIC-Diff-VQA | | VQA-RAD | | |
| **Cat.** | **Method** | AUC | AUG | AUC | AUG | BS$^{\dagger}$ |
| 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 | – | – | – |
| | SPCD | 89.91 | 64.46 | 78.60 | 41.20 | .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 SPCD | +2.3 | +2.4 | +2.0 | +2.1 | +.002 |
| | Δ best PO | +1.8 | +1.8 | +1.5 | +1.7 | +.004 |

| | | 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$_{\text{ftH}}$ | – | – | 90.89 | 3.96 | .89 |
| Dec.-Time | VCD | 21.3 | 29.1 | 91.20 | 4.07 | .90 |
| | SPCD | 23.1 | 30.8 | 92.70 | 4.19 | .91 |
| 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 SPCD | +2.2 | +2.4 | +2.1 | +.13 | +.01 |
| | Δ best PO | +1.7 | +1.9 | +2.4 | +.17 | +.01 |

**Implementation.** SPPO freezes the visual encoder and finetunes language and projection layers with AdamW (lr $2\times10^{-5}$, weight decay $10^{-2}$, $\beta$=0.1, $\lambda_{\text{SFT}}$=0.1) for 3 epochs on 8×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, $T_{\max}$=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.

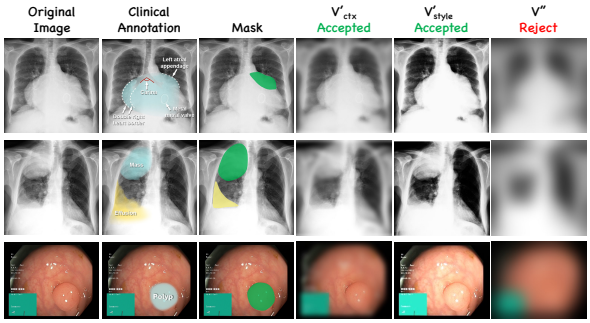


Figure 3: **Counterfactual view gallery.** Rows span chest radiography (1–2) and GI endoscopy (3).

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

| | MIMIC-Diff-VQA | | VQA-RAD | |
|---|---|---|---|---|
| Method | 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 |
| SPCD | 77.70 | 57.09 | 68.90 | 41.90 |
| **LENS** | **80.00** | **59.29** | **71.10** | **43.90** |

**(Q1 and Q2) Improvement over baselines and backbone generalization.** LENS achieves the best performance on every metric (Table 1), improving over SPCD by 2.0–2.4 AUC on radiology VQA, 2.2 RadGraph F1 on CXR reporting, and 2.1 QAAS on GI endoscopy, while outperforming the best PO baseline (GRPO) by 1.7–2.4 points. Standard PO methods yield only modest gains over SPCD (0.2–0.5 AUC), confirming that the stability signal in SPPO—not faithfulness scoring alone—provides the larger margin. On LLaVA-Med (Table 2), a backbone with weaker medical pre-training, gains are even larger (+2.3 AUC on MIMIC-Diff-VQA, +2.2 on VQA-RAD vs. +2.3 and +2.0 on CheXagent), consistent with the hypothesis that weakly pre-trained models exhibit a larger stability gap for LENS to exploit.

**(Q3) Independent metric quality.** Table 3 reports BERTScore F1 and CheXpert F1, neither used in training. LENS maintains or improves both across domains (e.g., CheXpert F1: 0.421→0.463 on MIMIC-CXR), ruling out the possibility that hallucination gains trade off language quality.

**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 SPCD 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 3: **Independent metrics**† not used in training or preference construction. LENS maintains or improves BERTScore F1 and CheXpert F1 across all domains.

| | Training metric | | Independent† | |
|---|---|---|---|---|
| Method | Metric | Value | BERTScore | CheXpert F1 |
| *MIMIC-Diff-VQA (CheXagent)* | | | | |
| Greedy | AUC | 86.20 | 0.871 | – |
| SPCD | AUC | 89.91 | 0.873 | – |
| **LENS** | AUC | **92.21** | **0.875** | – |
| *MIMIC-CXR (CheXagent)* | | | | |
| Greedy | RG-F1 | 20.9 | 0.862 | 0.421 |
| SPCD | RG-F1 | 23.1 | 0.864 | 0.445 |
| **LENS** | RG-F1 | **25.3** | **0.866** | **0.463** |
| *Gut-VLM (LLaVA-1.6-7B$_{ftH}$)* | | | | |
| Greedy | QAAS | 90.89 | 0.884 | – |
| SPCD | QAAS | 92.70 | 0.886 | – |
| **LENS** | QAAS | **94.80** | **0.889** | – |

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

| | Description ($n$=100) | | VQA ($n$=80) | |
|---|---|---|---|---|
| Method | Halluc. (%) ↓ | Adequacy ↑ | Corr. (%) ↑ | Pref. vs. VCD ↑ |
| Greedy | 17.0 | 3.80 | 84.0 | – |
| VCD | 13.0 | 3.96 | 86.3 | 50.0 |
| SPCD | 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 SPCD by 0.6 AUC; LENS decoding without SPPO improves by 1.4. Their combination yields +2.3, substantially more than either alone, confirming complementarity: SPPO reduces the *frequency* of unstable tokens while stability-aware decoding catches residual instabilities at inference.

**(b) Design choice.** Replacing adaptive $\alpha_t$ 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 ($s_{\text{obj}}$) costs 1.7 AUC, with largest degradation on MIMIC-CXR where find-

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

| Variant | MIMIC-Diff-VQA AUC | MIMIC-Diff-VQA AUG | MIMIC-CXR GREEN | MIMIC-CXR RG-F1 |
|---|---|---|---|---|
| *(a) Component decomposition* | | | | |
| Greedy (no mitigation) | 86.20 | 60.10 | 28.5 | 20.9 |
| + SPCD decoding only | 89.91 | 64.46 | 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** |
| *(b) Stability weighting and constraints* | | | | |
| Fixed $\alpha_t = 0.5$ | 91.05 | 65.62 | 31.8 | 24.1 |
| No adaptive $\alpha_t$ ($\kappa$=0) | 90.82 | 65.38 | 31.5 | 23.9 |
| Global sim. only (no $s_{\text{obj}}$) | 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** |
| *(c) Attribution robustness* | | | | |
| 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 |

*(d) Computational overhead (A100, $T_{\max}$=64)*

| | Latency ms | Latency Rel. | Peak mem GB |
|---|---|---|---|
| Greedy | 42 | 1.00× | 8.2 |
| SPCD | 68 | 1.62× | 13.6 |
| **LENS** | **74** | **1.76×** | **14.1** |

Table 6: **PO baselines and $\gamma$ 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 AUC | MIMIC-Diff-VQA AUG | MIMIC-CXR GREEN | MIMIC-CXR RG-F1 | Gut-VLM QAAS | Gut-VLM R-Sim |
|---|---|---|---|---|---|---|
| *No training (decoding only)* | | | | | | |
| Greedy | 86.20 | 60.10 | 28.5 | 20.9 | 90.89 | 3.96 |
| + SPCD decoding | 89.91 | 64.46 | 30.8 | 23.1 | 92.70 | 4.19 |
| *Standard PO + LENS decoding* | | | | | | |
| 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 |
| *SPPO $\gamma$ sweep + LENS decoding* | | | | | | |
| $\gamma$=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) | 91.40 | 66.12 | 32.5 | 24.8 | 93.90 | 4.26 |
| $\gamma$=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** |

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

| $\tau_{\text{obj}}$ \ $\tau_{\text{global}}$ | 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 |

ings 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 SPCD at inference (74 vs. 68 ms on A100, $T_{\max}$=64), a 1.09× overhead. Both methods require one additional forward pass per decoding 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×A100 GPUs, across all subsequent inference.

**Stability penalty strength ($\gamma$ ablation).** Table 6 examines how the stability penalty weight $\gamma$ 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 $\tau_{\text{global}}$ and $\tau_{\text{obj}}$, 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.

**Full Gut-VLM metrics.** Table 8 reports the complete set of Gut-VLM description metrics including lexical overlap. LENS improves across all met-

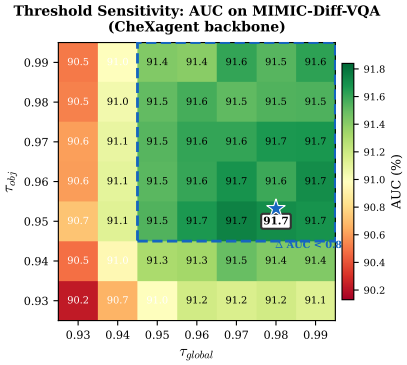


Figure 4: **Threshold sensitivity.** AUC on MIMIC-Diff-VQA (CheXagent) as $\tau_{\text{global}}$ and $\tau_{\text{obj}}$ vary. Performance is stable across $[0.95, 0.99]$ (<0.4 pt variation).

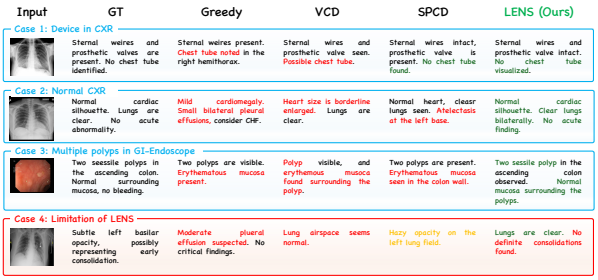


Figure 5: **Qualitative comparison across methods.** Green text: clinically correct. *Red italic*: hallucinated content. Orange text (case 4): over-conservative miss. Cases 1–3 show LENS eliminating hallucinations that persist under VCD and SPCD. Case 4 illustrates the specificity–sensitivity trade-off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

rics: BLEU (+0.01 over SPCD), 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.

### 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 prosthetic valve. In Case 2, greedy fabricates cardiomegaly with effusions—a high-risk false positive; LENS correctly reports normal findings. In Case 3, greedy hallucinates erythematous mucosa; LENS accurately describes sessile polyps with normal surrounding mucosa.

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$_{\text{ftH}}$ | .88 | .82 | .90 | 4.04 | 90.53 |
| LLaVA-1.6-7B$_{\text{ftH}}$ | .89 | .82 | .90 | 3.96 | 90.89 |
| VCD | .90 | .82 | .91 | 4.07 | 91.20 |
| SPCD | .91 | .84 | .93 | 4.19 | 92.70 |
| **LENS** | **.92** | **.85** | **.94** | **4.32** | **94.80** |

## 5 Discussion and Conclusion

LENS demonstrates that the stability gap—the measurable logit divergence between grounded and hallucinated tokens under semantics-preserving perturbation—is a practical and effective signal for hallucination mitigation in medical VLMs. Two design choices proved critical: the adaptive contrast weight $\alpha_t$, which prevents over-suppression of grounded tokens by modulating penalty strength based on contextness (Table 5b), and object-level proximity guards, without which global similarity alone fails to protect small findings such as nodules and tube tips. LENS is complementary to existing alignment: because stability-aware decoding is independent of the training objective, it can be layered on top of DPO- or GRPO-aligned models as an inference-time safeguard (Table 6). The specificity–sensitivity trade-off observed for borderline findings (§4.4) is controllable via $\gamma$ and $\alpha_{\max}$, and motivates future work on adaptive conservatism scheduling. Further extensions include learning counterfactual generators end-to-end, integrating chain-of-thought supervision for parametric-knowledge hallucinations (Dong et al., 2025), and evaluation on volumetric modalities (CT, MRI), pathology, and multilingual settings.

**Conclusion.** LENS introduces cross-view stability as a first-class signal for hallucination mitigation, operationalizing it through SPPO at training and stability-aware contrastive decoding at inference. Across radiology VQA, CXR reporting, and GI endoscopy, LENS achieves 1.8–2.9 point improvements over the SOTA, with clinical evaluation confirming a hallucination rate reduction from 17.0% to 8.5% establishing cross-view stability as a principled, architecture-agnostic, and practically deployable criterion for safer medical VLMs.

## 6 Limitations

LENS's stability penalty operates on a principled assumption: tokens whose logits shift under

semantics-preserving perturbation are likely hallucinated. This assumption holds strongly for clear-cut findings but becomes ambiguous for borderline cases—such as the subtle basilar opacity in Figure 5, Case 4—where even grounded tokens produce equivocal logit patterns because the underlying visual evidence is itself uncertain. In such cases, LENS conservatively withholds the assertion, favouring specificity over sensitivity. Importantly, this behaviour is not a fixed limitation but a tunable operating point: reducing $\gamma$ and $\alpha_{\max}$ shifts the model toward higher sensitivity, enabling deployment-specific calibration along the specificity–sensitivity curve. We plan to address this further through adaptive $\gamma$-scheduling conditioned on per-token attribution confidence, allowing the model to relax conservatism precisely when the evidence is spatially localized but low-contrast. Additionally, our current evaluation spans chest radiography and GI endoscopy in English; validating LENS on volumetric modalities (CT, MRI), pathology, and multilingual clinical settings remains an important direction.

## 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. *arXiv preprint arXiv:2403.00425*.

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, Jenia 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. Mirage: Assessing hallucination in multimodal reasoning chains of multimodal large language models. *arXiv preprint arXiv:2505.24238*.

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, Mengliang Zhang, Tatsuya Harada, Ronald Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical vqa. In *KDD 2023*. Introduces MIMIC-Diff-VQA / Medical-Diff-VQA.

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, Jiaxing 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 EW 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 Bahadur Gurung, Cristian Linte, Angus Watson, Yash Raj 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*. 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*.

Chunyuan 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 vi-

sual 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, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. 2025. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *arXiv preprint arXiv:2506.11712*.

Dwarikanath Mahapatra, Behzad Bozorgtabar, and Imran Razzak. 2026. Semantics preserving contrastive decoding for hallucination robust medical vision–language models. *ResearchGate*. https://www.researchgate.net/publication/399488093_Semantics_Preserving_Contrastive_Decoding_for_Hallucination_Robust_Medical_Vision-Language_Models.

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*.

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 (ICCV)*, pages 618–626.

Tsung-Han Wu, Heekyung Lee, Jiaxin 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*.

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 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*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: 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*.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian Acosta, John Fahrner, and Pranav Rajpurkar. 2024. Radflag: A black-box hallucination detection method for medical vision language models. *arXiv preprint arXiv:2411.00299*.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. 2025. VASparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. *arXiv preprint arXiv:2501.06553*.