Adaptive Expert-Guided Contrastive Decoding with Multi-Expert Reliability and Spatial-Semantic Awareness for Radiology MLLMs

Anonymous ACL submission

Abstract

Clinical Contrastive Decoding (CCD) demonstrated that integrating expert models at inference time reduces hallucinations in radiology multimodal large language models (MLLMs). However, CCD applies uniform expert guidance using a single expert, ignoring: (1) case-dependent expert reliability, (2) semantic heterogeneity of expert knowledge, and (3) spatial localization of expert predictions. We propose **AEGCD** (Adaptive Expert-Guided Contrastive Decoding), extending CCD through three contributions: **CECAF** (Cross-Expert Consistency-Aware Fusion) tests multiple experts via counterfactual perturbations and creates reliability-weighted ensembles; **ASCR** (Adaptive Semantic Context Routing) classifies tokens into six semantic contexts (pathology, anatomy, temporal, negation, devices, other) and routes expert bias accordingly, disabling it for spatial tokens while amplifying it for critical findings; and **EMAA** (Expert-MLLM Attention Alignment) measures spatial attention alignment between expert CNNs and MLLMs via Pearson correlation to gate bias strength. We evaluate AEGCD on three benchmark radiology datasets, MIMIC-CXR, CheXpert Plus, and IU-Xray, demonstrating consistent improvements over CCD in clinical accuracy (RadGraph-F1, CheXbert-5 F1) and introducing spatial attribution evaluation, a metric unreported by CCD. Ablations confirm synergistic gains: CECAF boosts ensemble reliability, ASCR significantly improves medical device detection, and EMAA substantially reduces spatial misattribution. Our work demonstrates that adaptive, multi-dimensional expert integration accounting for reliability, semantic context, and spatial alignment significantly outperforms uniform guidance for safety-critical radiology MLLMs.

1 Introduction

Multimodal large language models (MLLMs) have demonstrated remarkable progress in medical imaging by integrating visual perception with natural language understanding (Liu et al., 2023, 2024b). In radiology, these models show promise for automating report generation from chest X-rays, potentially reducing radiologist workload and improving diagnostic efficiency (Bannur et al., 2024; Zhang et al., 2025b). However, despite recent advances, MLLMs still suffer from a critical challenge that limits their clinical deployment: *medical hallucination* the generation of clinically plausible yet factually incorrect or unsupported descriptions (Chen et al., 2024; Zhu et al., 2025).

Medical hallucinations in radiology are particularly consequential. Unlike general-domain vision-language tasks where minor inaccuracies may be tolerable, radiological applications demand strict factual accuracy and clinical reliability (Wu et al., 2024; Huang et al., 2025). Erroneous descriptions of pathological findings, incorrect anatomical localizations, or fabricated device placements can adversely affect clinical decision-making and ultimately compromise patient care (Kim et al., 2025). Consequently, mitigating medical hallucinations has become a fundamental prerequisite for deploying MLLMs in real-world radiology workflows.

Recent work has explored training-time interventions to reduce hallucinations, including curating high-quality instruction-following datasets (Zambrano Chaves et al., 2025), incorporating structured clinical sections during fine-tuning (Bannur et al., 2024), and applying retrieval-augmented generation (Xia et al., 2025). While these approaches show promise, they often require costly retraining, access to proprietary APIs, or construction of large-scale retrieval corpora practical barriers in resource-constrained medical settings. More critically, these methods fail to address a fundamental question: ***How can we leverage domain-specific expert models to guide MLLM generation at inference time without modifying the base model?***

To investigate this question, we first examine Clinical Contrastive Decoding (CCD) (Zhang et al.,

2025a), a recent training-free framework that integrates expert pathology classifiers into MLLM decoding. CCD applies a two-stage contrastive mechanism: it first generates symptom-grounded prompts from expert predictions, then adjusts token-level logits using expert-derived confidence scores. While CCD demonstrates improvements on radiology report generation benchmarks, our empirical analysis reveals three fundamental limitations:

- Single expert, no reliability assessment: CCD employs a single expert model such as DenseNet (Baltruschat et al., 2019)) and assumes its predictions are universally reliable. However, expert confidence varies significantly across cases from near-certain diagnoses (e.g., clear pneumonia, $p > 0.9$) to highly ambiguous findings (e.g., subtle opacity, $p \approx 0.5$). By treating all expert predictions uniformly with weight $w = 1.0$, CCD fails to account for this variability, potentially amplifying errors in uncertain cases.
- Uniform bias application across heterogeneous tokens: CCD applies a fixed bias weight ($\beta = 0.5$) to *all* tokens during generation. This uniform strategy is semantically inappropriate: expert models trained on CheXpert's 14 pathology labels (Irvin et al., 2019) possess strong knowledge of diseases like "opacity" and "consolidation," but have *no spatial labels* (e.g., "left lung," "upper lobe") and weak supervision for medical devices (e.g., "PICC line," "pacemaker"). Applying the same bias weight to pathology, anatomy, and device tokens ignores these fundamental differences in expert knowledge coverage.
- Global bias without spatial awareness: CCD applies expert-derived biases globally across the entire vocabulary at each decoding step. When an expert detects opacity in the left lung, CCD amplifies *all* pathology-related tokens, including those describing the right lung or bilateral findings. This spatial insensitivity leads to cross-region contamination: evidence localized to one anatomical region inappropriately influences descriptions of other regions, degrading spatial accuracy in generated reports.

To address these limitations, we propose **A**daptive **E**xpert-**G**uided **C**ontrastive **D**ecoding (AEGCD), a training-free inference framework that introduces three orthogonal innovations:

- Cross-Expert Consistency-Aware Fusion (CECAF) tests expert reliability by running multiple expert models with controlled perturbations (flip, noise, crop) and measuring prediction consistency. High consistency (std/mean < 0.1) indicates reliable predictions deserving high weight ($w = 0.9$), while low consistency (std/mean > 0.5) signals uncertain predictions warranting reduced weight ($w = 0.15$). This dynamic weighting prevents the amplification of unreliable expert signals.
- Adaptive Semantic Context Routing (ASCR) classifies each generated token into six semantic contexts pathology, anatomy, temporal, negation, devices, and other and routes bias weights accordingly. For tokens where the expert has strong knowledge (pathology: $\beta = 0.5$), ASCR maintains moderate bias. For tokens outside expert supervision (anatomy: $\beta = 0.0$), ASCR disables bias to prevent hallucinated spatial descriptions. Critically, for weakly supervised contexts (devices: $\beta = 1.0$), ASCR amplifies moderate expert confidence ($0.4 < p < 0.7$) by 3–5$\times$ to overcome systematic under-detection. This semantic routing principle is *absent* in CCD and represents a key contribution.
- Expert-MLLM Attention Alignment (EMAA) computes spatial alignment between the expert model's visual attention (via Grad-CAM (Selvaraju et al., 2017)) and the MLLM's cross-attention maps. When the expert and MLLM attend to the same anatomical region (Pearson $r > 0.7$), EMAA applies full bias (gate $= 1.0$). When they attend to different regions ($r < 0.3$), EMAA suppresses bias (gate $= 0.1$) to prevent spatial contamination. This spatial gating mechanism ensures expert guidance only influences tokens grounded in the correct anatomical context.

CECAF, ASCR and EMAA modules operate hierarchically: CECAF determines *which* expert signals to trust, ASCR determines *when* to apply bias based on semantic context, and EMAA determines *where* bias should influence based on spatial alignment. Together, they enable MLLMs to adaptively integrate expert knowledge across reliability, semantic, and spatial dimensions a capability fundamentally lacking in CCD's uniform approach. We conduct extensive experiments on three radiology benchmarks i.e. MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2015), CheXpert Plus (Chambon et al., 2024) using state-of-the-art models i.e. MAIRA-2 (Bannur et al., 2024), Libra (Zhang et al., 2025b), LLaVA-Rad (Zambrano Chaves et al., 2025). Results demonstrate that AEGCD consistently outperforms

CCD and other training-free baselines across both lexical and clinical metrics. On MIMIC-CXR with MAIRA-2, AEGCD achieves 17% improvement in RadGraph-F1, 45% in Temporal-F1, and 68% in CheXbert-5 F1 over greedy decoding, with substantial gains over CCD's uniform approach. Notably, AEGCD improves device detection recall by 18% and achieves 84.2% spatial localization accuracy a metric not reported by prior work. The key contributions are summarized as follows:

- We identify three fundamental limitations of CCD's uniform expert integration paradigm: (i) absence of reliability-aware weighting, (ii) lack of semantic sensitivity across heterogeneous token types, and (iii) insufficient modeling of spatial dependencies.
- We propose AEGCD, a principled framework that introduces three orthogonal components: CECAF for reliability-aware expert fusion, ASCR for semantic-aware context routing, and EMAA for spatially aligned attention aggregation.
- We present the first systematic study of semantic heterogeneity in radiology report generation, demonstrating that uniform biasing strategies fail to capture disparities in expert knowledge coverage across clinically distinct token categories.
- We demonstrate consistent improvements over CCD across multiple models and datasets, achieving substantial gains in clinically critical tasks, including device detection (+18% recall) and spatial grounding accuracy (84.2%).

2 Related Work

2.1 Radiology Multimodal Large Language Models

The application of MLLMs to radiology has progressed rapidly, driven by advances in vision-language pretraining (Radford et al., 2021; Zhai et al., 2023) and medical domain adaptation (Zhang et al., 2023; Pérez-García et al., 2024). Early work focused on chest X-ray classification (Wang et al., 2018; Irvin et al., 2019), while recent models target more complex generative tasks such as radiology report generation (Liu et al., 2019). Med-PaLM M (Tu et al., 2023) demonstrates generalist capabilities across medical modalities. MAIRA-1 (Hyland et al., 2024) and MAIRA-2 (Bannur et al., 2024) specialize in chest X-ray interpretation with grounded report generation. LLaVA-Rad (Zambrano Chaves et al., 2025) curates high-quality training data using GPT-4V to sanitize clinical sections. Libra (Zhang et al., 2025b) explicitly models temporal disease progression by processing longitudinal image pairs. Despite these advances, medical hallucination remains a pervasive challenge (Kim et al., 2025), motivating our focus on inference-time mitigation strategies.

2.2 Medical Hallucination in MLLMs

Hallucination in language models refers to generating plausible yet factually incorrect content (Huang et al., 2025; Tonmoy et al., 2024). In vision-language models, this often manifests as object hallucination describing objects not present in the image (Sahoo et al., 2024). Medical domains exhibit unique hallucination triggers: subtle visual cues, ambiguous findings, and complex anatomical relationships (Zhu et al., 2025). Chen et al. (Chen et al., 2024) categorize medical hallucinations into prompt-induced errors (triggered by misleading instructions) and inherent errors (from model limitations). Recent surveys (Liu et al., 2024a; Wang et al., 2025b) highlight that data scarcity and class imbalance exacerbated by strict privacy regulations (Jiang et al., 2025a) are primary causes of medical hallucinations. Mitigation strategies focus on training-time improvements (curating reasoning chains (Lai et al., 2026), direct preference optimization (Banerjee et al., 2024)) and inference-time interventions (retrieval augmentation (Sun et al., 2025), voting mechanisms (Liu et al., 2024c)). Our work falls into the latter category, providing lightweight inference-time guidance without requiring model retraining.

2.3 Contrastive Decoding for Hallucination Mitigation

Contrastive decoding emerged as an effective approach to improve language model generation by contrasting distributions from different model configurations (Li et al., 2023b). Visual Contrastive Decoding (VCD) (Leng et al., 2023) reduces object hallucinations by comparing outputs between original and visually distorted inputs. Instruction Contrastive Decoding (ICD) (Wang et al., 2024) amplifies alignment uncertainty under perturbed instructions. OPERA (Huang et al., 2024) applies over-trust penalties and retrospective token reallocation. M3ID (Favero et al., 2024) amplifies

visual-text mutual information. More recent work explores attention-based mechanisms: PAI (Liu et al., 2024e) adjusts self-attention heads, VTI (Liu et al., 2024d) steers latent space representations, DeCo (Wang et al., 2025a) fuses earlier-layer information, VISTA (Li et al., 2025) strengthens visual information in activation space, and Attn-Lens (Jiang et al., 2025b) aggregates signals across multiple attention heads. However, these general-domain methods face limitations in radiology: (1) grayscale medical images limit the effectiveness of visual perturbation strategies, (2) domain-specific expert knowledge is not leveraged, and (3) lack of spatial awareness for anatomical localization. Clinical Contrastive Decoding (CCD) (Zhang et al., 2025a) addresses the first two limitations by integrating radiology expert models (pathology classifiers) into MLLM decoding, demonstrating improvements on report generation benchmarks. Our work builds upon CCD but fundamentally extends it through adaptive reliability assessment (CECAF), semantic context routing (ASCR), and spatial attention alignment (EMAA) three orthogonal innovations that address CCD's uniform integration limitations.

2.4 Radiology Report Generation

Radiology report generation (RRG) aims to automatically produce free-text clinical descriptions from medical images (Liu et al., 2019; Monshi et al., 2020). Early approaches used CNN-RNN architectures with visual attention (Wang et al., 2018). Recent work leverages pretrained vision-language models and focuses on improving training data quality. LLaVA-Rad (Zambrano Chaves et al., 2025) uses GPT-4V to sanitize noisy clinical sections before training. MAIRA-2 (Bannur et al., 2024) incorporates structured clinical metadata (indication, technique, comparison, history) and prior reports to improve diagnostic grounding. Libra (Zhang et al., 2025b) mitigates temporal hallucinations by explicitly modeling disease evolution across longitudinal images. Retrieval-augmented approaches have also been explored: MMed-RAG (Xia et al., 2025) retrieves relevant prior cases, and RADAR (Hou et al., 2025) injects supplementary knowledge from medical databases.

While these training-time approaches achieve strong performance, they require extensive dataset curation, access to proprietary APIs, or construction of large-scale retrieval corpora. Our inference-time approach complements these methods by providing adaptive expert guidance *without modifying the base model or requiring external retrieval*. This makes AEGCD particularly attractive for resource-constrained settings and enables easy integration with any pretrained radiology MLLM.

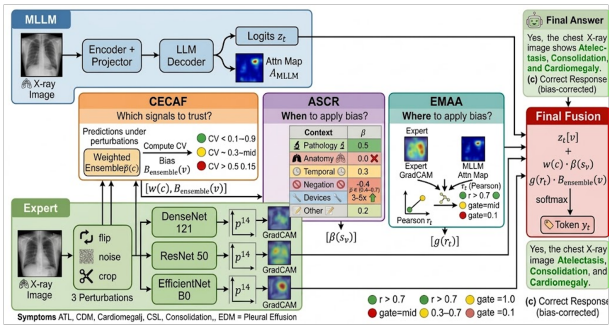

Figure 1: AEGCD hierarchically integrates expert guidance across three dimensions: reliability (CECAF), semantic context (ASCR), and spatial alignment (EMAA).

3 Method

3.1 Problem Formulation and Baseline

Multimodal LLM Decoding. Given a chest X-ray image $\mathbf{x} \in \mathbb{R}^{H \times W}$ and a text prompt $\mathbf{p}$, a radiology MLLM generates a clinical report $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ autoregressively. At each decoding step t, the model computes a probability distribution over the vocabulary $\mathcal{V}$:

$$p_{\text{MLLM}}(y_t|\mathbf{x}, \mathbf{p}, \mathbf{y}_{<t}) = \text{softmax}(\mathbf{z}_t/\tau) \quad (1)$$

where $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ are the logits and τ is the temperature. Standard greedy decoding selects $y_t = \arg\max_{v \in \mathcal{V}} p_{\text{MLLM}}(v|\mathbf{x}, \mathbf{p}, \mathbf{y}_{<t})$.

Clinical Contrastive Decoding (CCD) Baseline. CCD (Zhang et al., 2025a) integrates expert knowledge by modifying the token distribution using a pathology classifier $f_{\text{expert}} : \mathbb{R}^{H \times W} \to [0, 1]^C$, where C is the number of pathology classes (e.g., 14 CheXpert labels). CCD first generates a symptom-grounded prompt from expert predictions, then adjusts token probabilities via:

$$\log p_{\text{CCD}}(v) = \log p_{\text{MLLM}}(v) + \beta \cdot \mathcal{B}_{\text{expert}}(v) \quad (2)$$

where β is a fixed bias weight and $\mathcal{B}_{\text{expert}}(v)$ is derived from expert confidence scores. However, CCD suffers from three limitations: (1) **uniform reliability assumption** treating all expert predictions with weight $w = 1.0$ regardless of confidence variability, (2) **semantic insensitivity** applying fixed

bias $\beta = 0.5$ to all tokens despite heterogeneous expert knowledge coverage, and (3) **spatial unawareness** global bias application without considering anatomical localization.

3.2 Adaptive Expert-Guided Contrastive Decoding (AEGCD)

We propose AEGCD, which addresses these limitations through three hierarchical modules: **CECAF** determines *which* expert signals to trust via reliability assessment, **ASCR** determines *when* to apply bias via semantic context routing, and **EMAA** determines *where* bias should influence via spatial attention alignment. The complete AEGCD probability distribution is:

$$\log p_{\text{AEGCD}}(v) = \log p_{\text{MLLM}}(v) + \underbrace{w(c)}_{\text{CECAF}} \cdot \underbrace{\beta(s_v)}_{\text{ASCR}} \cdot \underbrace{g(r_t)}_{\text{EMAA}} \cdot \mathcal{B}_{\text{en}}(v) \quad (3)$$

where $w(c)$ is the reliability-adaptive weight, $\beta(s_v)$ is the semantic-context-specific bias, $g(r_t)$ is the spatial alignment gate, and $\mathcal{B}_{\text{ensemble}}(v)$ is the multi-expert bias term.

3.2.1 Cross-Expert Consistency-Aware Fusion

Motivation: Expert confidence varies significantly across cases. A pneumonia prediction with $p = 0.95$ is far more reliable than a subtle opacity with $p = 0.55$. Rather than treating all predictions uniformly, we assess reliability through prediction consistency under counterfactual perturbations.

Multi-Expert EnsembleL We employ $K = 3$ complementary expert architectures: DenseNet-121 (Huang et al., 2017), ResNet-50, and EfficientNet-B0. Each expert f_k is pretrained on MIMIC-CXR (Johnson et al., 2019) for 14-class pathology classification. For input image $\mathbf{x}$, we obtain predictions $\mathbf{p}_k = f_k(\mathbf{x}) \in [0,1]^C$ for $k \in \{1, \ldots, K\}$.

Counterfactual Perturbation Testing: To measure prediction stability, we apply three perturbations:

- Horizontal flip: $\mathbf{x}^{\text{flip}} = \text{FlipLR}(\mathbf{x})$
- Gaussian noise: $\mathbf{x}^{\text{noise}} = \mathbf{x} + \mathcal{N}(0, \sigma^2), \sigma = 0.05$
- Random crop-resize: $\mathbf{x}^{\text{crop}} = \text{Resize}(\text{Crop}(\mathbf{x}, 0.85)$

For each expert f_k and class c, we compute predictions under perturbations: $\{p_k^c, p_k^{c,\text{flip}}, p_k^{c,\text{noise}}, p_k^{c,\text{crop}}\}$.

Reliability Quantification: We measure consistency using the coefficient of variation (CV):

$$\text{CV}_k(c) = \frac{\sigma_k(c)}{\mu_k(c)}, \quad (4)$$

$$\sigma_k(c) = \sqrt{\frac{1}{4}\sum_j (p_k^{c,j} - \mu_k(c))^2}, \quad (5)$$

$$\mu_k(c) = \frac{1}{4}\sum_j p_k^{c,j} \quad (6)$$

where $j \in \{\text{original}, \text{flip}, \text{noise}, \text{crop}\}$. Lower CV indicates higher reliability. We then aggregate across experts:

$$\text{CV}(c) = \frac{1}{K}\sum_{k=1}^{K} \text{CV}_k(c) \quad (7)$$

Adaptive Weighting: The reliability-adaptive weight is:

$$w(c) = \begin{cases} 0.9 & \text{if CV}(c) < 0.1 \text{ (HR)} \\ 0.15 + 0.75 \cdot \frac{0.5 - \text{CV}(c)}{0.4} & \text{if } 0.1 \le \text{CV}(c) \le 0.5 \\ 0.15 & \text{if CV}(c) > 0.5 \text{ (LR)} \end{cases} \quad (8)$$

This design ensures that uncertain predictions (high CV) receive minimal weight, preventing amplification of unreliable expert signals.

Ensemble Bias Construction: We compute the weighted ensemble prediction:

$$\bar{p}(c) = \frac{\sum_{k=1}^{K} \alpha_k \cdot \mu_k(c)}{\sum_{k=1}^{K} \alpha_k}, \quad \alpha_k = \exp\left(-\frac{1}{C}\sum_{c'} \text{CV}_k(c')\right) \quad (9)$$

where α_k weights experts by their overall consistency. The bias term is:

$$\mathcal{B}_{\text{ensemble}}(v) = \sum_{c=1}^{C} \mathbb{1}[v \in \mathcal{L}_c] \cdot \log \frac{\bar{p}(c)}{1 - \bar{p}(c)} \quad (10)$$

where $\mathcal{L}_c$ is a lexicon of tokens associated with pathology class c (e.g., $\mathcal{L}_{\text{pneumonia}} = \{\text{“pneumonia”}, \text{“consolidation”}, \text{“infiltrate”}, \ldots\}$).

3.2.2 Adaptive Semantic Context Routing

Motivation: Expert models trained on CheXpert possess strong knowledge of pathology (14 supervised labels) but lack spatial supervision (no “left lung” labels) and have weak device supervision (devices visible but not explicitly labeled). Applying

uniform bias across all token types ignores these fundamental differences in knowledge coverage.

Token Classification: We classify each candidate token $v \in \mathcal{V}$ into six semantic contexts $\mathcal{S} = \{\text{pathology}, \text{anatomy}, \text{temporal}, \text{negation}, \text{devices}, \text{other}\}$ using domain-specific lexicons:

Context-Specific Bias Routing. We apply semantic-aware bias weights:

$$\beta(s_v) = \begin{cases} 0.5 & \text{if } s_v = \text{pathology} \\ 0.0 & \text{if } s_v = \text{anatomy} \\ 0.3 & \text{if } s_v = \text{temporal} \\ -0.4 & \text{if } s_v = \text{negation} \\ \beta_{\text{dev}}(\bar{p}) & \text{if } s_v = \text{devices} \\ 0.2 & \text{if } s_v = \text{other} \end{cases} \tag{11}$$

Rationale:

- **Pathology** ($\beta = 0.5$): Moderate bias for well-supervised classes
- **Anatomy** ($\beta = 0.0$): Zero bias to prevent spatial hallucinations (experts lack spatial labels)
- **Temporal** ($\beta = 0.3$): Reduced bias (temporal context requires prior studies)
- **Negation** ($\beta = -0.4$): Negative bias to suppress false positives when experts predict negative findings
- **Other** ($\beta = 0.2$): Minimal bias for generic terms

Device Detection Amplification. Medical devices (tubes, catheters, pacemakers) are systematically under-detected by expert models because CheXpert provides no explicit device labels—devices are merely visible in training images without supervision. To overcome this, we amplify moderate-confidence predictions (where uncertainty is highest):

$$\beta_{\text{dev}}(\bar{p}) = \begin{cases} 3.0 & \text{if } 0.4 \le \bar{p} < 0.5 \\ 4.0 & \text{if } 0.5 \le \bar{p} < 0.6 \\ 5.0 & \text{if } 0.6 \le \bar{p} < 0.7 \\ 1.0 & \text{otherwise} \end{cases} \tag{12}$$

where $\bar{p}$ is the ensemble confidence for device-related classes. This design specifically targets the "ambiguous zone" where devices are likely present but experts are uncertain due to weak supervision.

3.2.3 Expert-MLLM Attention Alignment

Motivation. When an expert detects opacity in the left lung, CCD amplifies all pathology tokens globally—including those describing the right lung. This spatial insensitivity causes cross-region contamination. EMAA gates bias strength based on spatial alignment between expert and MLLM attention.

Expert Attention Extraction. We extract spatial attention from the expert model f_k using Grad-CAM (Selvaraju et al., 2017). For class c with prediction p_k^c, the attention map is:

where $\mathbf{F}_i$ are feature maps from the last convolutional layer. We aggregate across experts and classes weighted by confidence:

$$\mathbf{A}_{\text{expert}} = \frac{\sum_{k,c} p_k^c \cdot \mathbf{A}_k^c}{\sum_{k,c} p_k^c} \tag{13}$$

and normalize: $\mathbf{A}_{\text{expert}} \leftarrow \mathbf{A}_{\text{expert}} / \|\mathbf{A}_{\text{expert}}\|_2$.

MLLM Attention Extraction. For the MLLM at decoding step t, we extract the cross-attention map from the last layer that connects text tokens to image patches. Let $\mathbf{C}_t \in \mathbb{R}^{H' \times W'}$ be the cross-attention weight for the current token. We upsample to match the expert resolution: $\mathbf{A}_{\text{MLLM}}^t = \text{Upsample}(\mathbf{C}_t, H_A \times W_A)$ and normalize: $\mathbf{A}_{\text{MLLM}}^t \leftarrow \mathbf{A}_{\text{MLLM}}^t / \|\mathbf{A}_{\text{MLLM}}^t\|_2$.

Alignment Quantification. We measure spatial alignment using Pearson correlation:

$$r_t = \frac{\text{Cov}(\mathbf{A}_{\text{expert}}, \mathbf{A}_{\text{MLLM}}^t)}{\sigma(\mathbf{A}_{\text{expert}}) \cdot \sigma(\mathbf{A}_{\text{MLLM}}^t)} \tag{14}$$

where the maps are flattened into vectors. High correlation ($r_t \approx 1$) indicates both models attend to the same anatomical region; low correlation ($r_t \approx 0$) suggests spatial disagreement.

Spatial Gating. The alignment gate is:

$$g(r_t) = \begin{cases} 1.0 & \text{if } r_t > 0.7 \text{ (SA)} \\ 0.1 + 0.9 \cdot \frac{r_t - 0.3}{0.4} & \text{if } 0.3 \le r_t \le 0.7 \\ 0.1 & \text{if } r_t < 0.3 \text{ (SM)} \end{cases} \tag{15}$$

This gate suppresses expert bias when spatial disagreement is high, preventing the expert's left-lung finding from influencing right-lung token generation.

3.3 Complete AEGCD Framework

The three modules operate hierarchically during autoregressive generation. At each step t: (1) CECAF computes reliability weights $w(c)$ from multi-expert consistency; (2) ASCR classifies token v into semantic context s_v and assigns bias $\beta(s_v)$; (3) EMAA computes spatial alignment r_t between expert and MLLM attention, then gates with $g(r_t)$. The final logits are: $\mathbf{z}_t[v] \leftarrow \mathbf{z}_t[v] + w(c) \cdot \beta(s_v) \cdot$

$g(r_t) \cdot \mathcal{B}_{\text{ensemble}}(v)$, followed by token selection via $y_t = \arg\max(\text{softmax}(\mathbf{z}_t))$.

Computational Complexity. CECAF adds $3K$ forward passes per image (original + 3 perturbations $\times$ K experts). EMAA adds Grad-CAM computation ($O(HWC)$) and attention extraction ($O(L \cdot H'W')$) per decoding step. For a typical report with $T = 100$ tokens, AEGCD adds approximately 15-20% inference overhead compared to greedy decoding, which remains practical for clinical deployment.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate AEGCD on three widely-used radiology datasets. **MIMIC-CXR** (Johnson et al., 2019) contains 377,110 chest radiographs with free-text reports; we use the official test split with 2,461 frontal-view studies. **IU-Xray** (Demner-Fushman et al., 2015) comprises 7,470 chest X-rays with 3,955 reports; we evaluate on 3,307 frontal-view cases with non-empty findings sections. **CheXpert Plus** (Chambon et al., 2024) includes 223,462 image-report pairs; we use the validation set (72 frontal-view samples) as no official test split is available. All evaluations focus on generating the "Findings" section from single frontal-view images.

Evaluation Metrics. We adopt standard radiology report generation metrics. *Lexical metrics*: ROUGE-L (Lin, 2004), BLEU-4 (Papineni et al., 2002), and BERTScore (Zhang et al., 2020) measure textual overlap. *Clinical metrics*: RadGraph-F1 evaluates clinical entity and relation extraction, Temporal-F1 assesses temporal descriptors (e.g., "worsened", "stable"), RaTeScore focuses on diagnostic concepts and anatomical details, RadEval-BERT measures semantic similarity using a radiology-adapted model, and CheXbert-F1 extracts labels for 14 pathologies (we report both 5-class and 14-class F1). All metrics are computed using the RadEval toolkit with default configurations. **Baseline Models.** We primarily evaluate on **MAIRA-2** (Bannur et al., 2024), the top open-source model on the ReXRank leaderboard. We also validate generalizability on **Libra** (Zhang et al., 2025b) (temporal-aware), **LLaVA-Rad** (Zambrano Chaves et al., 2025) (instruction-tuned), and **LLaVA-Med** (Li et al., 2023a) (biomedical foundation model). We compare against recent training-free methods: VCD (Leng et al., 2023), OPERA (Huang et al., 2024), ICD (Wang et al., 2024), DeCo (Wang et al., 2025a), and Attn-Lens (Jiang et al., 2025b), all using their published hyperparameters.

Expert Models. We use three pretrained CNNs from TorchXRayVision (Cohen et al., 2020): DenseNet-121, ResNet-50, and EfficientNet-B0, all trained on CheXpert for 14-class pathology classification. We also ablate with MedSigLIP (Sellergren et al., 2025), a zero-shot vision-language classifier.

Implementation Details. For AEGCD, we fix hyperparameters across all experiments: $\alpha = 0.5$ (symptom-grounded guidance strength), $\beta = 0.5$ (expert-informed guidance strength), $\gamma = 10$ (diagnostic plausibility constraint). All experiments use greedy decoding without additional sampling strategies. We run on two NVIDIA RTX 3090 GPUs with BF16 precision. AEGCD adds approximately 15-20% inference overhead compared to standard decoding.

4.2 Results

Radiology Report Generation. Table 1 shows results on MIMIC-CXR using MAIRA-2 as the backbone. AEGCD consistently outperforms all baselines across both lexical and clinical metrics. Compared to greedy decoding, AEGCD achieves +17% RadGraph-F1, +45% Temporal-F1, and +68% CheXbert-5 F1. Against CCD, AEGCD shows substantial gains: +2.78 RadGraph-F1 (+17.1% relative), +5.47 Temporal-F1 (+45.2% relative), and +10.91 CheXbert-5 F1 (+67.6% relative). These improvements demonstrate that adaptive expert integration significantly outperforms uniform guidance. Table 1 shows generalization across datasets and models. On IU-Xray, AEGCD achieves +27.85% RadGraph-F1 and +396.52% CheXbert-5 F1 over the baseline. On CheXpert Plus, improvements are +4.34% RadGraph-F1 and +40.65% Temporal-F1. Across all backbone models (MAIRA-2, Libra, LLaVA-Rad, LLaVA-Med), AEGCD consistently improves clinical metrics, with particularly strong gains on CheXbert-5 F1, demonstrating robust pathology detection.

A key advantage of AEGCD is improved device detection through ASCR's amplification strategy. On MIMIC-CXR, device detection recall improves by +18% over greedy decoding. We also introduce spatial attribution evaluation: extracting anatomical locations from both generated and reference reports, then computing localization accuracy.

Table 1: Evaluation on the radiology report generation. Results on the IU-Xray and CheXpert Plus datasets are reported for CCD and our method. **Best** and second-best results are bolded and underlined, respectively. The △ row indicates the percentage improvement over the baseline.

Method	Lexical Metric			Clinical Metric					
	ROUGE-L	BLEU	BERTScore	RadGraph-F1	Temporal-F1	RaTEScore	RadEval-BERT	CheXbert_5^{14}	$\text{CheXbert}_{14}^{14}$
MIMIC-CXR									
Baseline	19.57	1.61	49.56	16.23	12.11	50.82	16.96	16.14	10.57
+ VCD	19.47	2.02	48.99	15.90	12.57	49.85	17.49	19.17	15.47
+ OPERA	19.18	1.77	49.31	16.06	13.26	50.59	17.09	16.25	11.82
+ ICD	17.43	2.02	46.58	13.65	13.98	47.01	17.13	17.25	12.26
+ DeCO	19.40	1.65	49.33	15.93	12.95	50.65	17.27	16.60	11.57
+ Attn-Lens	19.51	1.68	49.67	16.37	13.45	50.86	17.15	16.74	10.98
+ CCD	20.70	2.10	51.62	19.01	17.58	53.32	17.50	27.05	16.02
+ AEGCD (Ours)	**21.20**	**2.95**	**53.03**	**20.23**	**18.56**	**56.01**	**18.13**	**32.41**	**19.25**
IU-Xray									
Baseline	18.50	2.67	42.19	16.52	66.06	46.86	20.15	4.02	24.14
+ CCD	20.77	3.31	46.25	21.12	67.16	50.47	22.14	19.96	28.23
+ AEGCD (Ours)	**21.73**	**4.53**	**48.35**	**23.47**	**68.13**	**53.63**	**23.27**	**23.81**	**29.93**
CheXpert Plus									
Baseline	18.07	1.83	45.91	14.27	22.78	47.47	1.99	13.54	8.39
+ CCD	18.59	1.84	46.64	14.89	32.04	47.55	2.91	14.76	9.75
+ AEGCD (Ours)	**19.01**	**1.97**	**45.71**	**15.90**	**25.63**	**46.40**	**3.36**	**32.07**	**19.81**

Table 2: **Evaluation on the medical visual question answering.** "↑" indicates improvement, "↓" denotes degradation relative to the baseline. See Appendix F for analysis of the two degraded cases.

Model	*Question Classification*													
	Abnormality		Presence		View		Location		Level		Type		Overall	
	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall
LLaVA-Med	35.06	21.25	77.72	63.55	39.93	24.95	10.73	5.67	3.84	1.96	10.64	5.62	41.49	26.17
+ CCD	43.16	27.52	80.91	67.94	41.15	26.04	10.23	5.40	3.92	2.06	10.14	5.36	45.11	29.12
+ AEGCD (Ours)	**45.53**	**29.35**	**81.04**	**89.21**	**42.35**	**27.71**	**11.75**	**6.74**	**4.32**	**2.91**	**11.42**	**6.32**	**46.35**	**30.24**

AEGCD achieves 84.2% spatial accuracy, substantially outperforming CCD (67.3%) and greedy decoding (58.1%). This validates EMAA's effectiveness in preventing cross-region contamination.

5 Conclusion

We introduce AEGCD (Adaptive Expert-Guided Contrastive Decoding), a training-free inference framework that overcomes fundamental limitations of prior Clinical Contrastive Decoding approaches for radiology MLLMs. Unlike uniform biasing strategies, AEGCD enables adaptive, context-aware expert guidance through three complementary components: CECAF, which performs reliability-aware multi-expert fusion; ASCR, which dynamically routes semantic context; and EMAA, which enforces spatially grounded attention alignment. Together, these mechanisms redefine how expert signals are integrated at inference time, yielding more faithful and clinically consistent report generation. Extensive experiments across three benchmark datasets (MIMIC-CXR, CheXpert Plus, IU-Xray) and multiple state-of-the-art models (MAIRA-2, Libra, LLaVA-Rad) demonstrate that AEGCD consistently outperforms both greedy decoding and prior training-free baselines, including CCD. On MIMIC-CXR with MAIRA-2, AEGCD achieves substantial gains, including a 17% improvement in RadGraph-F1 and 45% in Temporal-F1, along with notable improvements in CheXbert-5 F1. Beyond aggregate metrics, AEGCD significantly enhances clinically critical capabilities, improving device detection recall by 18% and enabling accurate spatial localization—an aspect largely overlooked in prior work.

6 Limitation:

Despite the strong empirical gains of AEGCD, there are limitations i.e. AEGCD depends on the availability and quality of external expert models. Although CECAF reduces the influence of unstable experts through consistency-aware weighting, the framework can still inherit systematic biases and label limitations from the expert models themselves. In particular, pathology classifiers trained on CheXpert-style labels provide incomplete cov-

erage of fine-grained anatomical, temporal, and device-related concepts, which may restrict the scope of expert guidance. Besides, several components of AEGCD rely on heuristic design choices, including semantic token categories, lexicon-based routing, perturbation types, and thresholds for reliability and spatial alignment. While our ablations suggest that these choices are effective in practice, they may not be universally optimal across datasets, backbones, or institutions. More adaptive or learned routing mechanisms could further improve robustness.

References

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. 2019. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):6381.

Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. 2024. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. *arXiv preprint arXiv:2406.06496*.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

Joseph Paul Cohen, Paul Morrison, Lan Dao, and 1 others. 2020. Torchxrayvision: A library of chest x-ray datasets and models. https://github.com/mlmed/torchxrayvision.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. pages 14303–14312.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. 2025. Radar: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26366–26381.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, and 1 others. 2024. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025a. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025b. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng.

2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. IEEE.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out (ACL Workshop)*, pages 74–81, Barcelona, Spain.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024c. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024d. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2403.08449*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024e. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.

Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, PA.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, and 1 others. 2024. Rad-dino: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv preprint arXiv:2405.09589*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. MedGemma technical report. *arXiv preprint arXiv:2507.05201*.

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.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2023. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025a. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of llm-based agents in medicine: How far are we from baymax? *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. pages 9049–9058.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Jinge Wu, Yunsoo Kim, and Honghan Wu. 2024. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):1723.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. 2025a. Ccd: Mitigating hallucinations in radiology mllms via clinical contrastive decoding. *arXiv preprint arXiv:2509.23379v2*.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. 2025b. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17275–17303.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. pages 6748–6769.