CMID: Towards Medical Visual Question Answering via Contrastive Mutual Information Decoding

Zhihong Zhu[1], **Yunyan Zhang**[1], **Fan Zhang**[2], **Bowen Xing**[3], **Xian Wu**[1*]
[1]Tencent Jarvis Lab
[2]The Chinese University of Hong Kong
[3]University of Science and Technology Beijing
{profzhu, kevinxwu}@tencent.com

Abstract

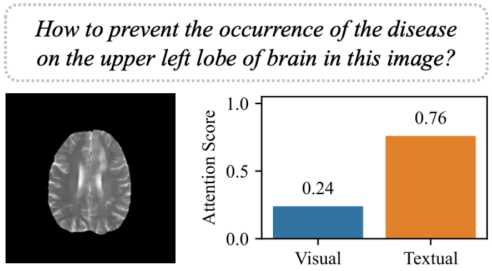
Medical Visual Question Answering (Med-VQA) aims to generate accurate answers for clinical questions grounded in medical images, which has attracted increasing research attention due to its potential to streamline diagnostics and reduce clinical burden. Recent advances in Large Vision-Language Models (LVLMs) have shown great promise for Med-VQA, but still suffer from two inference-time issues: (1) *attention shift*, where the LVLM over-relies on textual priors; and (2) *attention dispersion*, where it fails to focus on critical diagnostic regions. To tackle these issues, we propose Contrastive Mutual Information Decoding (CMID), a training-free inference-time intervention grounded in information theory for Med-VQA. Concretely, CMID first identifies the *Principal Focus Area* (PFA) from decoder attention maps, then constructs focus-preserving and focus-excluding views to derive dual contrastive signals that simultaneously amplify salient visual cues and suppress background noise. Crucially, these corrective signals are adaptively scaled by a reliability-gated self-correction mechanism, based on the distributional shift induced by the PFA. Extensive experiments on three Med-VQA benchmarks demonstrate the effectiveness of CMID. Further analyses showcase its robust generalizability across diverse medical architectures and tasks.

1 Introduction

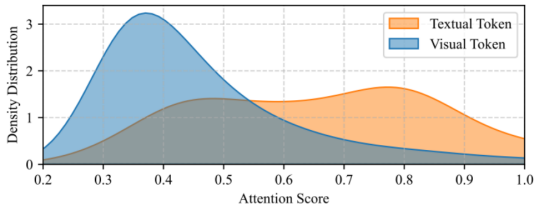
Medical visual question answering (Med-VQA) aims to provide accurate answers to clinical questions regarding medical images, which has attracted increasing research attention (Yan, Duan, and Wang 2024; Liu et al. 2024). By automating the diagnostic process, Med-VQA has the potential to reduce turnaround times and lower medical costs (Zhu et al. 2025a; Wu et al. 2024). With the rapid advancement of large vision language models (LVLMs), their powerful capabilities are unlocking new frontiers in Med-VQA systems (Chang et al. 2025). However, LVLMs still struggle to fully understand multimodal medical information, and often produce inaccurate or hallucinated outputs (Xia et al. 2024).

Recent efforts to adapt LVLMs for Med-VQA can be broadly classified into *fine-tuning* and *training-free* methods. Regarding the former, some works focus on large-scale, domain-specific pre-training to instill medical knowledge (Li et al. 2023a; Chen et al. 2024). Furthermore, advanced training strategies like Mixture-of-Experts (MoE) (Liu et al. 2024; Chang et al. 2025), curriculum learning (Rui et al. 2025), and multi-step reasoning (Gai et al. 2025; Wang et al. 2025) have been introduced. While fine-tuning improves performance, it incurs significant training costs and reduces scalability in medical applications.

In contrast, *training-free* methods serve as an alternative approach to enhance LVLM performance in Med-VQA. Some studies guide LVLMs to pay more attention to the region of interest (ROI) through visual prompting (Zhu et al. 2025b), but such methods require prior annotation of medical images, resulting in additional overhead. Motivated by this, we aim to boost LVLM performance in Med-VQA solely during the inference stage, without relying on additional training or manual annotation. This not only signifi-



(a) A Med-VQA Sample with Typical Attention Shift



(b) Attention Density Distribution for Visual and Textual Tokens

Figure 1: Attention shift in Med-VQA. (a) The example showing a typical image-question pair, with the average attention scores for visual and textual tokens. (b) The attention density distribution for visual (blue) and textual (orange) tokens on the overall VQA-RAD dataset (Lau et al. 2018).

*Corresponding author.

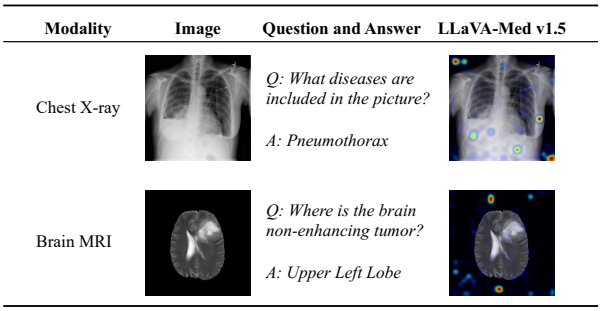


Figure 2: Examples from LLaVA-Med v1.5 with attention maps over Chest X-ray and Brain MRI modalities.

cantly reduces resource consumption, but also improves the flexibility of the model in real-world medical applications.

Despite their promising results, existing *training-free* methods struggle to resolve the underlying inference-time limitations of LVLMs in Med-VQA. Through an in-depth analysis of decoder self-attention patterns across visual and textual tokens, we reveal two main issues: *attention shift*, where the model over-relies on the textual question while under-utilizes visual information from the medical image as quantified in Figure 1; *attention dispersion*, where the model's attention is scattered across irrelevant background or redundant features, hindering accurate focus on key diagnostic details such as pneumothorax as shown in Figure 2.

In this paper, we introduce Contrastive Mutual Information Decoding (CMID), a new inference-time intervention method of LVLM for Med-VQA. Concretely, our CMID consists of three core components: (1) it begins by identifying the Principal Focus Area (PFA) from the self-attention matrix of visual-text tokens, which highlights the critical diagnostic regions in the medical image. Building on this, we apply a contrastive decoding strategy (Li et al. 2023c; Leng et al. 2024) grounded in mutual information theory to explicitly address the above two attentional issues. (2) To tackle attention shift, CMID maximizes the information gain from the PFA by contrasting the prediction logits from the *standard full* view and the *focus-excluding* view, encouraging the LVLM to rely more on the visual modality. (3) To tackle attention dispersion, CMID minimizes the redundant information from background regions by contrasting the prediction logits from the *standard full* view and the *focus-preserving* view, encouraging LVLM to filter out background noise. Crucially, CMID integrates a reliability-gated self-correction mechanism that adjusts the correction strength based on the estimated informativeness of the PFA.

Our contributions can be summarized as follows: (1) We present a new inference-time intervention method dubbed CMID for Med-VQA. To our best knowledge, this is the first attempt to bridge contrastive decoding and information-theoretic principles in Med-VQA. (2) We introduce three primary components in the proposed CMID, to explicitly address issues of attention shift and attention dispersion in LVLMs for Med-VQA. (3) Extensive experiments including quantitative, qualitative, cross-task, cross-architecture evaluation demonstrate the effect and generalizability of CMID.

2 Related Work

Medical Visual Question Answering. Medical Visual Question Answering (Med-VQA) aims to provide clinically accurate answers to questions about medical images (Liu et al. 2022). Prior research has primarily focused on designing specialized architectures (Zhan et al. 2020; Cong et al. 2022), to facilitate effective cross-modal fusion between visual and textual features. The advent of LVLMs has marked a significant paradigm shift (Zhang et al. 2025b, 2024a). Models such as LLaVA-Med (Li et al. 2023a) and HuatuoGPT-Vision (Chen et al. 2024) leverage domain-specific pretraining and instruction tuning to boost performance (Zhu et al. 2025c; Zhang et al. 2025a). Other approaches enrich the input context through auxiliary tools (Zhu et al. 2025b; Chang et al. 2025), or improve training dynamics via curriculum learning (Rui et al. 2025) and multi-step reasoning (Gai et al. 2025). Overall, most existing methods rely on resource-intensive fine-tuning, which incurs high computational costs and limits scalability. Moreover, even well-tuned LVLMs often suffer from inference behaviors, including attention shift toward textual priors and attention dispersion over irrelevant visual regions.

In this work, we propose a training-free decoding strategy that enhances diagnostic accuracy by leveraging mutual information. Our method applies contrastive signals between attention-masked views, and incorporates a reliability-gated correction mechanism to refine the attentional focus.

Inference-time Intervention. Inference-time intervention offers a compelling and training-free alternative for enhancing LVLM (Park et al. 2025; Suo et al. 2025). However, most existing strategies are developed for natural images and fail to address the unique challenges of Med-VQA. A prominent line of this paradigm is contrastive decoding (Li et al. 2023c), where methods like VCD (Leng et al. 2024) and PAI (Liu, Zheng, and Chen 2024) globally amplify the visual signal. However, they lack spatial specificity and struggle to distinguish subtle lesions from complex contexts. Conversely, spatially-focused approaches, such as token pruners (Zhang et al. 2025d) and attention calibrators (Woo et al. 2025), rely on risky magnitude-based saliency heuristics. Such heuristics are unreliable in medicine, where critical diagnostic signals are often low-contrast or diffuse, leading to the erroneous pruning of clinically vital features.

In this work, we propose a unified solution grounded in mutual information theory that departs from relying on potentially misleading static attention scores. Instead, our method assesses a region's importance based on its functional impact on the output distribution. It achieves this by applying dual contrastive signals derived from *focus-preserving* and *focus-excluding* views, with the correction adaptively moderated by a novel reliability-gated mechanism to ensure faithfulness to critical diagnostic evidence.

3 Methodology

1 Preliminaries

Consider a general LVLM parameterized by θ, which comprises a vision encoder, a vision-text interface (*e.g.*, a linear layer or Q-Former (Li et al. 2023b)), and an LLM decoder.

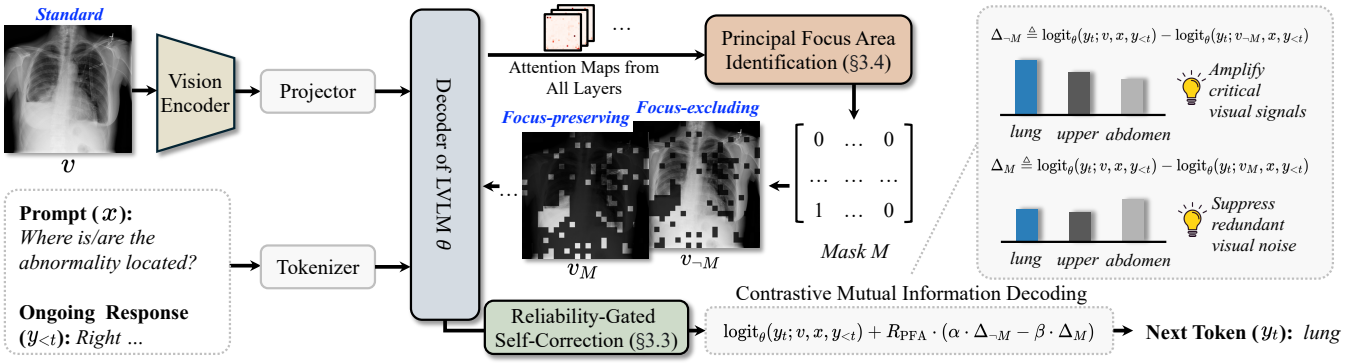


Figure 3: Overview of the CMID. As LVLM autoregressively generates text w.r.t. a medical image input (*e.g.*, a chest X-ray showing Pneumothorax), it may produce incorrect responses due to issues like *attention shift* and *attention dispersion*. However, CMID corrects these by first identifying the Principal Focus Area (PFA) from the decoder's attention maps (§3.4) to create two contrastive views: a *focus-preserving* view (v_M) containing only the critical region and a *focus-excluding* view ($v_{\neg M}$) with the background. By contrasting the logits from these views against the standard one, CMID computes a signal-enhancing $\Delta_{\neg M}$ to amplify critical visual signals and a noise-suppressing Δ_M to suppress redundant visual noise. Crucially, the Reliability-Gated Self-Correction module (§3.3) calculates a score (R_{PFA}) to dynamically modulate the strength of these corrections. This entire gated correction is applied to the standard logits, steering the model to accurately generate the next token *lung*.

For the Med-VQA task, given a medical image v and a question x, the image is first processed by the vision encoder and interface to produce a sequence of visual embeddings. For notational simplicity, we hereafter use v to refer to this sequence of visual embeddings. These embeddings are concatenated with the question x as input to the text decoder, which autoregressively generates a textual answer Y as:

$$y_t \sim p_\theta(y_t|v, x, y_{<t}) \propto \exp(\text{logit}_\theta(y_t|v, x, y_{<t})), \quad (1)$$

where y_t represents the t-th generated token and $y_{<t}$ refers to the sequence of tokens generated prior to step t. The function $\text{logit}_\theta(y_t; \cdot)$ represents the logit distribution function.

During decoding, the key K and value V in the attention heads are derived from preceding decoding steps and stored in a key-value cache. Consequently, the attention for decoding the t-th token with dimension D is calculated as:

$$\text{Attention}(q_t, K_{\leq t}) = \text{Softmax}\left(\frac{q_t K_{\leq t}^\top}{\sqrt{D}}\right), \quad (2)$$

where q_t is the query for the current decoding step, and $K_{\leq t}$ represents the key vectors up to and including step t.

Our primary goal is to address two key attentional issues during inference: (1) *attention shift*, where the LVLM underutilizes the visual modality, and (2) *attention dispersion*, where attention is scattered across irrelevant regions.

2 Contrastive Decoding Objective

To address the above issues, we proposed CMID, whose architecture is shown in Figure 3. Conceptually, we can partition visual embeddings v into two disjoint sets: v_s containing critical diagnostic features, and v_n containing irrelevant background information. Grounded in information theory (Duncan 1970; Latham and Roudi 2009; Belghazi et al. 2018), we aim to maximize the information gain from the visual signal while minimizing that from the noise:

$$\max\left(I(Y; v_s|x) - I(Y; v_n|x)\right). \quad (3)$$

Since this ideal partition (v_s, v_n) is unknown, we cannot optimize this objective directly. Instead, we approximate this partition at inference time using an attention-derived binary mask M (detailed in §4). This mask helps us define a *focus-preserving* view v_M, which serves as a proxy for the signal v_s; and a *focus-excluding* view $v_{\neg M}$ for v_n:

$$v_M \triangleq v \odot M, \quad v_{\neg M} \triangleq v \odot (1 - M). \quad (4)$$

Here, $\odot$ denotes a masking operation that selectively zeros out visual token embeddings based on the binary mask M.

By contrasting predictions from these views, we can derive corrective signals to steer the generation process, pushing it to rely more on visual modality and critical regions.

This leads us to define three generation contexts: the *standard* context $C = (v, x, y_{<t})$, the *focus-preserving* context $C_M = (v_M, x, y_{<t})$, and the *focus-excluding* context $C_{\neg M} = (v_{\neg M}, x, y_{<t})$. From these, we then derive two corrective signals $\Delta_{\neg M}$ and Δ_M to represent the contribution of the PFA and the background, respectively:

$$\Delta_k \triangleq \text{logit}_\theta(y_t; C) - \text{logit}_\theta(y_t; C_k), \quad \text{where } k \in \{M, \neg M\}. \quad (5)$$

The logits for the contrastive contexts C_M and $C_{\neg M}$ are computed only once and the resulting corrective signals Δ_k are cached. This design makes the computational overhead a one-time cost, ensuring high efficiency during generation.

3 Reliability-Gated Self-Correction

Naively applying the corrective signals from Eq. (5) is risky, as an imperfect PFA could amplify noise. To this end, we introduce a reliability-gated self-correction mechanism.

The core intuition is that a reliable PFA should contribute significantly more new information to the prediction than the background. To achieve this, the mechanism adjusts the contribution of the PFA based on its informativeness, ensuring that only reliable focus areas are amplified, while potentially noisy or irrelevant regions are attenuated. We formalize this "information contribution" I_k, by measuring the

distributional shift caused by each view using the Kullback-Leibler (KL) divergence (Joyce 2011):

$$I_k \triangleq D_{\mathrm{KL}}\left(P(y_t|C) \,||\, P(y_t|C_k)\right), \quad \text{where } k \in \{M, \neg M\}. \tag{6}$$

We then formulate a PFA Reliability Score, $R_{\mathrm{PFA}} \in [0, 1]$, as the normalized contribution of the PFA:

$$R_{\mathrm{PFA}} = \frac{I_{\neg M}}{I_{\neg M} + I_M + \epsilon}, \tag{7}$$

where ϵ is a small constant for numerical stability.

This score acts as a continuous gate: $R_{\mathrm{PFA}} \to 1$ when the PFA is informative, and $R_{\mathrm{PFA}} \to 0$ when it is not.

This leads to our CMID objective, where the logit for a token y_t is computed by integrating the base logits with the reliability-gated corrective signals as follows:

$$\mathrm{logit}_{\mathrm{CMID}}(y_t; C) = \mathrm{logit}_\theta(y_t; C) + R_{\mathrm{PFA}} \cdot (\alpha \cdot \Delta_{\neg M} - \beta \cdot \Delta_M), \tag{8}$$

where $\alpha, \beta > 0$ denote trade-off hyperparameters.

Since CMID relies on the PFA for defining its contrastive views, we next describe how the mask M is estimated.

4 Principal Focus Area Identification

To construct the PFA mask M, we aim to produce a reliable focus map by building a consensus across decoder layers. Since naive averaging would conflate decisive signals with noise, we employ an entropy-aware reweighting scheme.

The first step is to quantify the focus of each layer. For each decoder layer l, we extract its attention map A_l over the N_v visual tokens and compute its Shannon Entropy (Araabi, Niculae, and Monz 2024), a measure of uncertainty:

$$H(A_l) = -\sum_{i=1}^{N_v} p_{l,i} \log_2 p_{l,i}, \tag{9}$$

where $\{p_{l,i}\}_{i=1}^{N_v}$ are the attention probabilities for layer l.

Since lower entropy signifies a more focused layer, which is likely to contain more relevant information, we assign each layer an inverse entropy weight w_l and normalize them across all N layers to obtain the final weights w_l':

$$w_l \propto \frac{1}{H(A_l) + \epsilon} \quad \text{and} \quad w_l' = \frac{w_l}{\sum_{j=1}^{N} w_j}. \tag{10}$$

We then compute the final consensus attention map $\mathcal{A}$ as the weighted average of the attention maps from each layer:

$$\mathcal{A} = \sum_{l=1}^{N} w_l' A_l. \tag{11}$$

Finally, to produce the binary mask M from the continuous map $\mathcal{A}$, we employ an adaptive threshold τ, calculated as the mean of all scores in $\mathcal{A}$.[1] The mask value M_i for each visual token i is then determined as:

$$M_i = \begin{cases} 1 & \text{if } \mathcal{A}_i > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

[1]We tested alternative thresholds such as Otsu and quantile-based methods, but found the mean-based one to be more stable.

Dataset	# Images	# QA Pairs	# Open	# Closed
VQA-RAD	203	451	179	272
SLAKE	96	1,061	645	416
PathVQA	858	6,719	3,357	3,362

Table 1: Dataset statistics. For SLAKE, only the English subset is considered for comparison with existing methods.

5 Theoretical Analysis

CMID can be interpreted as a principled update derived from information geometry (Amari and Nagaoka 2000). We first define the contrastive mutual information objective as:

$$\mathcal{L}_{\mathrm{CMID}} \triangleq D_{\mathrm{KL}}\big(P \,||\, P_{\neg M}\big) - D_{\mathrm{KL}}\big(P \,||\, P_M\big), \tag{13}$$

where $P = \mathrm{Softmax}(\mathbf{l})$ and $P_k = \mathrm{Softmax}(\mathbf{l}_k)$ for $k \in \{M, \neg M\}$. This encourages CMID to suppress noisy visual context while amplifying critical visual information.

The gradient with respect to logits is:

$$\nabla_{\mathbf{l}} \mathcal{L}_{\mathrm{CMID}} = P_M - P_{\neg M}. \tag{14}$$

A more principled update follows the *natural gradient* (Amari 1998):

$$\tilde{\nabla}_{\mathbf{l}} \mathcal{L} = \mathcal{I}(\mathbf{l})^{-1} \nabla_{\mathbf{l}} \mathcal{L}, \tag{15}$$

where $\mathcal{I}(\mathbf{l})$ denotes the Fisher Information Matrix (FIM). Using second-order Taylor approximation:

$$D_{\mathrm{KL}}\big(P \,||\, P'\big) \approx \frac{1}{2}(\mathbf{l} - \mathbf{l}')^\top \mathcal{I}(\mathbf{l})(\mathbf{l} - \mathbf{l}'), \tag{16}$$

we interpret logit differences as approximate natural gradient directions. Since computing $\mathcal{I}(\mathbf{l})^{-1}$ is intractable during decoding, we instead perform a single-step logit-space update with reliability gating as follows:

$$\begin{aligned} \mathbf{l}_{\mathrm{CMID}} &\leftarrow \mathbf{l} + \eta \cdot \tilde{\nabla}_{\mathbf{l}} \mathcal{L}_{\mathrm{CMID}} \\ &\approx \mathbf{l} + R_{\mathrm{PFA}} \left(\alpha \Delta_{\neg M} - \beta \Delta_M\right). \end{aligned} \tag{17}$$

This frames CMID as a geometrically-aware, tractable approximation to natural gradient optimization in logit space.

4 Experiments

1 Experimental Settings

Datasets and Evaluation Metrics. Following previous works (Li et al. 2023a; Zou and Yin 2025; Zhu et al. 2025b), we evaluate our model on three open medical VQA datasets. (1) VQA-RAD (Lau et al. 2018) is a radiology VQA dataset featuring diverse, clinician-generated questions spanning 11 categories. (2) SLAKE (Liu et al. 2021), is a bilingual radiology dataset with questions covering a wide range of human body parts. And we evaluated on its English subset for a fair comparison. (3) PathVQA (He et al. 2020) is a large-scale pathology VQA dataset, where questions focus on the visual attributes of microscopy images. Detailed statistics of these three datasets are reported in Table 1. All datasets contain open-ended questions (free-form answers) and closed-ended questions (yes/no answers), and we follow the official data splits for all datasets.

For evaluation metrics, we report Accuracy for closed and Recall for open questions, following Li et al. (2023a).

Method	Reference	VQA-RAD		SLAKE		PathVQA	
		Open	Closed	Open	Closed	Open	Closed
Base	-	34.39	61.76	42.72	54.09	9.86	68.32
Greedy Decoding	-	33.56	62.87	43.72	54.81	10.04	68.65
Beam Search (Sutskever, Vinyals, and Le 2014)	NeurIPS'14	32.76	<u>65.07</u>	41.90	56.49	10.15	<u>71.00</u>
Nucleus Sampling (Holtzman et al. 2020)	ICLR'20	27.87	63.60	38.56	51.20	8.70	64.28
VCD (Leng et al. 2024)	CVPR'24	29.75	62.50	40.84	56.01	10.49	66.87
PAI (Liu, Zheng, and Chen 2024)	ECCV'24	33.02	63.24	43.87	55.05	10.20	67.55
M3ID (Favero et al. 2024)	CVPR'24	30.93	59.19	40.57	52.88	9.95	64.63
AvisC (Woo et al. 2025)	ACL'25	<u>37.89</u>	62.87	41.19	55.05	10.15	69.48
SparseVLM (Zhang et al. 2025d)	ICML'25	32.98	61.76	43.10	<u>56.79</u>	10.63	68.91
VisPruner (Zhang et al. 2025c)	ICCV'25	33.02	61.40	<u>44.89</u>	56.25	<u>10.76</u>	67.82
CMID (Ours)	-	**38.75**†	**66.87**†	**46.56**†	**58.01**†	**11.41**†	**71.97**†
Improv. (%)	-	2.27↑	2.77↑	3.72↑	2.15↑	6.04↑	1.37↑

Table 2: Main results of the competitive baseline methods and the proposed CMID framework on three medical VQA benchmarks. The best results are in **bold**, while the second-best results are in <u>underlined</u>. † denotes the improvements over the best baseline are statistically significant with $p < 0.05$ under t-test.

Variant	VQA-RAD		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
CMID (Full Model)	**38.75**	**66.87**	**46.56**	**58.01**	**11.41**	**71.97**
w/ Direct Enhancement	36.78 (↓1.97)	64.51 (↓2.36)	44.62 (↓1.94)	55.95 (↓2.06)	10.61 (↓0.80)	70.13 (↓1.84)
w/o Attention Shift Correction ($\alpha = 0$)	36.57 (↓2.18)	63.98 (↓2.89)	44.49 (↓2.07)	55.73 (↓2.28)	10.53 (↓0.88)	69.85 (↓2.12)
w/o Attention Dispersion Correction ($\beta = 0$)	37.12 (↓1.63)	65.04 (↓1.83)	45.10 (↓1.46)	56.48 (↓1.53)	10.80 (↓0.61)	70.59 (↓1.38)
w/o Reliability Gate ($R_{\text{PFA}} = 1$)	36.89 (↓1.86)	64.66 (↓2.21)	44.78 (↓1.78)	56.09 (↓1.92)	10.69 (↓0.72)	70.24 (↓1.73)
w/o Entropy-aware Reweighting	37.41 (↓1.34)	65.35 (↓1.52)	45.43 (↓1.13)	56.82 (↓1.19)	10.92 (↓0.49)	70.83 (↓1.14)

Table 3: Ablation Study. 'w/o' is short for 'without'.

Baselines. We compare the proposed CMID model with 10 competitive inference-time intervention baselines, which can be classified into three main categories: (1) *Conventional decoding methods*: Base, Greedy Decoding, Beam Search (Sutskever, Vinyals, and Le 2014), and Nucleus Sampling (Holtzman et al. 2020). (2) *Contrastive decoding methods*: VCD (Leng et al. 2024), PAI (Liu, Zheng, and Chen 2024), M3ID (Favero et al. 2024), and AvisC (Woo et al. 2025). (3) *Token pruning methods*: SparseVLM (Zhang et al. 2025d) and VisPruner (Zhang et al. 2025c).

Implementation Details. We utilize the checkpoint from llava-med-v1.5-mistral-7b (Li et al. 2023a) and follow the template provided by the LLaVA model to design basic system instructions for inference *without further training*. The hyperparameters α and β are empirically set to 1.0 and 0.3, respectively. To ensure a fair comparison, we re-implemented all baselines following their official settings. For statistical stability, all reported results are averaged over 5 runs using different random seeds.

2 Main Results

The main results of CMID and baselines are shown in Table 2, from which we have the following observations:

(1) CMID consistently outperforms all baselines across datasets and question types, with average gains of +2.52% on VQA-RAD and +2.94% on SLAKE. These improvements are statistically significant (p-value < 0.05), demonstrating the effectiveness of our information-theoretic objective in mitigating attention shift and dispersion. (2) General-domain inference-time methods yield suboptimal results. Methods like VCD and M3ID struggle to adapt to subtle features of medical images, showing limited generalizability to Med-VQA. In contrast, the proposed CMID explicitly operates the medical visual modality, leading to robust improvements. (3) The improvements of open-type questions on PathVQA are sharper. We suspect that PathVQA focus on pathology images, where subtle diagnostic signals are embedded in visually similar healthy tissue. Our CMID not only identifies the PFA, but also penalizes attention to misleading background regions via its contrastive terms.

3 Ablation Study

We perform thorough ablation studies on all dataset to understand the necessity of different designs and strategies in CMID. Table 3 shows the following vital observations.

(1) We first evaluate a baseline that bypasses contrastive deltas by directly injecting logits from PFA-preserving and PFA-suppressed inputs: $\text{logit}_{\text{Direct}} \leftarrow \text{logit}_\theta(C) + R_{\text{PFA}}\left(\alpha \cdot \text{logit}_\theta(C_M) - \beta \cdot \text{logit}_\theta(C_{\neg M})\right)$. This naive alternative results in a significant performance drop. This confirms that contrastive deltas ($\Delta_{\neg M}$, Δ_M) inspired from mutual information are essential for measuring and correcting

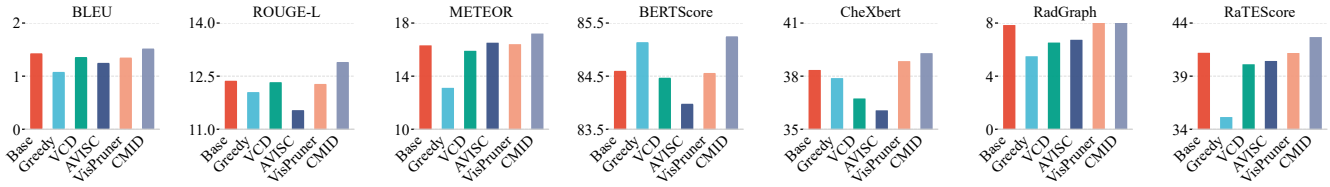


Figure 4: Comparison of medical report generation across seven evaluation metrics on the IU-Xray dataset.

Method	VQA-RAD		SLAKE	
	w/o CMID	w/ CMID	w/o CMID	w/ CMID
LLaVA-Med v1.5	48.08	52.81	48.41	52.29
HuatuoGPT-V	52.76	54.50	53.24	54.98
Med-Flamingo	46.10	48.87	46.45	49.13
RadFM	44.82	47.26	45.28	48.04
MedVInT	47.27	49.79	47.84	50.25

Table 4: Comparison of using different medical LVLMs on VQA-RAD and SLAKE datasets with and without CMID.

Method	VQA-RAD	SLAKE	PathVQA
Masking Strategy			
Soft Mask	52.15	51.68	41.03
Hard Mask	**52.81**	**52.29**	**41.69**
Layer Source			
w/ Early Layers (1-10)	51.34	51.21	40.25
w/ Middle Layers (11-21)	52.56	51.67	41.53
w/ Late Layers (22-32)	52.47	51.78	41.42
w/ All Layers	**52.81**	**52.29**	**41.69**

Table 5: Ablation study on the design of the Principal Focus Area (PFA) on three Med-VQA benchmarks. We report the average of open-ended Recall and closed-ended Accuracy.

modality-specific dependencies. (2) We ablate each correction term to assess its contribution. Removing $\Delta_{\neg M}$ (attention shift correction) leads to a substantial drop, as it enforces visual information and mitigates cross-modal imbalance; removing Δ_M (attention dispersion correction) also degrades performance, underscoring its role in suppressing irrelevant visual content. Notably, the former proves more critical: without initial visual attention, refining intra-modal focus yields limited benefit. (3) We further examine two PFA components. First, disabling the reliability gate reduces robustness, highlighting its role in filtering noisy or inaccurate PFAs. Second, replacing entropy-based reweighting with naive averaging degrades performance, suggesting that entropy effectively prioritizes informative attention layers.

4 Method Analysis

Attention Map Evaluation. We conduct ablation studies in Table 5 to validate the design choices for PFA, focusing on two key factors: the masking strategy and the selection of decoder attention layers. (1) Hard Masking consistently outperforms soft masking, due to its sharper separation between focus-preserving (C_M) and focus-excluding ($C_{\neg M}$)

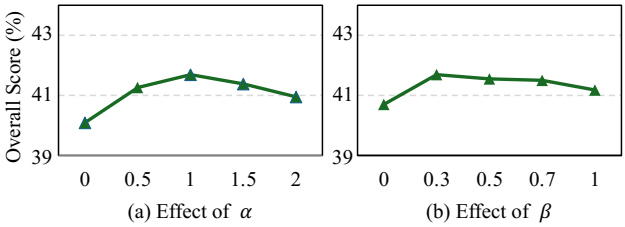


Figure 5: Hyper-parameter analysis of (a) α and (b) β on the PathVQA dataset. "Overall Score" represents the average of open-ended Recall and closed-ended Accuracy.

views. This yields stronger contrastive signals (Δ_k), which are critical for the corrective objective in Eq. 8. (2) Regarding attention source, middle layers (11–21) yield the best single-range performance, likely due to their balanced cross-modal alignment. Late layers (22–32) underperform slightly, as they emphasize abstract semantics over visual grounding. Our All Layers strategy achieves the highest scores by aggregating entropy-weighted attention across all layers (Eq. 10), adaptively emphasizing reliable signals.

Applicability on other Medical LVLMs. A natural question that arises is whether the proposed CMID is effective for diverse medical LVLMs. To answer the question, we conduct experiments with another four powerful medical LVLMs including HuatuoGPT-Vision-7B (Chen et al. 2024), Med-Flamingo (Moor et al. 2023), RadFM (Wu et al. 2023), and MedVInT (Zhang et al. 2024b). From the results in Table 4, the performance gains achieved by CMID consistently increase with stronger backbones. These results demonstrate the generalizability of CMID, confirming its effectiveness without reliance on any specific medical LVLM.

Generalizability on Medical Report Generation. To further assess the generalizability of CMID beyond Med-VQA, we conduct preliminary experiments on the medical report generation task using the IU-Xray dataset (Demner-Fushman et al. 2015; Xia et al. 2024). We adopt seven evaluation metrics to comprehensively evaluate model performance. As shown in Figure 4, our CMID consistently outperforms all baselines, demonstrating its strong generalization ability across different medical vision-language tasks.

Hyper-parameter Analysis. The hyper-parameters α and β in Eq.8 control the strength of the attention shift correction term $\Delta_{\neg M}$ and the attention dispersion correction term Δ_M, respectively. Here, we conduct a sensitivity analysis by varying $\alpha \in [0.0, 2.0]$ and $\beta \in [0.0, 1.0]$, with results illustrated in Figure 5. We observe that increasing either co-

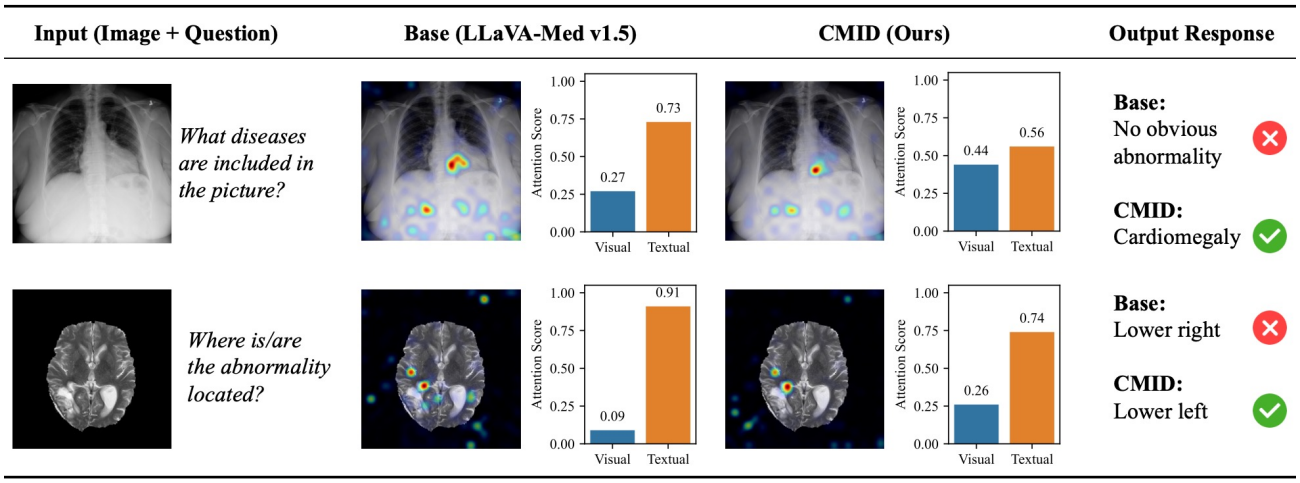


Figure 6: Cases on Chest X-ray and Brain MRI of the baseline LLaVA-Med v1.5 and CMID. For each case, we present the visual attention maps and corresponding bar charts that quantify the allocation of attention between visual and textual modalities.

Method	Latency ↓ (ms/token)	Memory ↓ Cost (MB)
Greedy Decoding	121.48 (× 1.00)	15219 (× 1.00)
VCD (Leng et al. 2024)	128.88 (× 1.06)	15224 (× 1.00)
AVISC (Woo et al. 2025)	193.34 (× 1.59)	16236 (× 1.07)
VisPruner (Zhang et al. 2025c)	119.25 (× 0.98)	15256 (× 1.00)
CMID (ours)	136.23 (× 1.12)	16387 (× 1.08)

Table 6: Comparison of SOTA baselines and the proposed CMID in latency (ms/token) and memory cost (MB).

efficient initially improves the overall score, reaching peak performance at $\alpha = 1.0$ and $\beta = 0.3$. Beyond these points, further increases lead to performance degradation. Notably, the model exhibits greater sensitivity to α than β, indicating that attention shift correction requires more careful tuning.

Inference Latency. We evaluate the inference latency of CMID against several SOTA baselines in Table 6. Our method introduces a modest 12% latency and 8% memory increase over greedy decoding, due to additional forward passes and PFA identification. Notably, CMID is more efficient than complex strategies like AVISC (× 1.59 latency) and competitive with VCD. This efficiency comes from computing corrective signals for contrastive views only once at the start then caching them, thus making the overhead a one-time cost. Similarly, the memory increase is due to caching these signals. Given the performance gains, this minor cost increase is a worthwhile trade-off for Med-VQA.

5 Case Study

Qualitative Comparison. To qualitatively demonstrate how CMID addresses the attentional failures in Med-VQA, we present two case studies in Figure 6. The first case (top row) shows a chest X-ray where the baseline model exhibits both a significant attention shift, with attention primarily on the textual input (0.73 vs. 0.27), and attention dispersion, with visual focus scattered across the abdomen. As a result, the baseline fails to detect the enlarged heart and generates the incorrect response "*No obvious abnormality*". In contrast, CMID restores balance between modalities (0.44 vs. 0.56) and relocates the spatial focus to the heart through PFA, leading to the correct answer "*Cardiomegaly*".

A similar pattern arises in the more subtle case of brain abnormalities. The baseline exhibits extreme modality bias (0.91 vs. 0.09) and produces vague and misaligned localization of the lesion, resulting in the wrong answer "*Lower right*". CMID not only restores cross-modal balance but also sharpens intra-modal focus, aligning attention precisely with the lesion. This enables the model to output the accurate response "*Lower left*". Thanks to the proposed contrastive mutual information decoding, the model corrects both inter-modality (*attention shift*) and intra-modality (*attention dispersion*) failures, improving performance in Med-VQA.

Analysis of Failure Cases. We have collected failure cases from VQA-RAD and SLAKE datasets, categorizing them into two main types. Type 1 occurs when CMID fails while the backbone performs correctly. This is often caused by a noisy PFA identified from an ambiguous medical image, which leads the contrastive decoding to inadvertently disrupt a valid reasoning process. Future work may explore more robust strategies to mitigate such disruptions. For Type 2, it arises when both outputs fail, typically due to fundamental issues in the base LVLM, like knowledge gaps.

5 Conclusion

We presented CMID, an inference-time intervention designed to mitigate attention shift and attention dispersion in LVLMs for Med-VQA. Grounded in mutual information, CMID identifies Principal Focus Areas and a reliability-gated mechanism to amplify signals from critical diagnostic regions while suppressing irrelevant background noise. Experiments demonstrate that CMID not only improves diagnostic accuracy but also exhibits strong robustness and generalizability across diverse medical LVLMs and tasks.

Acknowledgments

Bowen Xing was supported by the National Natural Science Foundation of China (Grant No. 62506033).

References

Amari, S.-I. 1998. Natural gradient works efficiently in learning. *Neural computation*.

Amari, S.-i.; and Nagaoka, H. 2000. *Methods of information geometry*, volume 191. American Mathematical Soc.

Araabi, A.; Niculae, V.; and Monz, C. 2024. Entropy–and Distance-Regularized Attention Improves Low-Resource Neural Machine Translation. In *ACL*.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *ICML*.

Chang, A.; Huang, L.; Boyd, A. J.; Bhatia, P.; Kass-Hout, T.; Xiao, C.; and Ma, F. 2025. Focus on What Matters: Enhancing Medical Vision-Language Models with Automatic Attention Alignment Tuning. *ACL*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv*.

Cong, F.; Xu, S.; Guo, L.; and Tian, Y. 2022. Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension. In *ACM MM*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA*.

Duncan, T. E. 1970. On the calculation of mutual information. *SIAM Journal on Applied Mathematics*.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-modal hallucination control by visual information grounding. In *CVPR*.

Gai, X.; Zhou, C.; Liu, J.; Feng, Y.; Wu, J.; and Liu, Z. 2025. Medthink: A rationale-guided framework for explaining medical visual question answering. In *NAACL*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *ICLR*.

Joyce, J. M. 2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*.

Latham, P. E.; and Roudi, Y. 2009. Mutual information. *Scholarpedia*.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023c. Contrastive decoding: Open-ended text generation as optimization. *ACL*.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *ISBI*. IEEE.

Liu, B.; Zhan, L.-M.; Xu, L.; and Wu, X.-M. 2022. Medical visual question answering via conditional reasoning and contrastive learning. *TMI*.

Liu, J.; Wang, Y.; Du, J.; Zhou, J. T.; and Liu, Z. 2024. Medcot: Medical chain of thought via hierarchical expert. *EMNLP*.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*. Springer.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*.

Park, W.; Kim, W.; Kim, J.; and Do, J. 2025. SECOND: Mitigating Perceptual Hallucination in Vision-Language Models via Selective and Contrastive Decoding. *ICML*.

Rui, S.; Chen, K.; Ma, W.; and Wang, X. 2025. Improving Medical Reasoning with Curriculum-Aware Reinforcement Learning. *arXiv*.

Suo, W.; Zhang, L.; Sun, M.; Wu, L. Y.; Wang, P.; and Zhang, Y. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *CVPR*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. *NeurIPS*.

Wang, Y.; Liu, J.; Gao, S.; Feng, B.; Tang, Z.; Gai, X.; Wu, J.; and Liu, Z. 2025. V2T-CoT: From Vision to Text Chain-of-Thought for Medical Reasoning and Diagnosis. *arXiv*.

Woo, S.; Kim, D.; Jang, J.; Choi, Y.; and Kim, C. 2025. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models. *ACL*.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv*.

Wu, X.; Zhao, Y.; Zhang, Y.; Wu, J.; Zhu, Z.; Zhang, Y.; Ouyang, Y.; Zhang, Z.; Wang, H.; Yang, J.; et al. 2024. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *NeurIPS*.

Xia, P.; Chen, Z.; Tian, J.; Gong, Y.; Hou, R.; Xu, Y.; Wu, Z.; Fan, Z.; Zhou, Y.; Zhu, K.; et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *NeurIPS*.

Yan, Q.; Duan, J.; and Wang, J. 2024. Multi-modal concept alignment pre-training for generative medical visual question answering. In *ACL*.

Zhan, L.-M.; Liu, B.; Fan, L.; Chen, J.; and Wu, X.-M. 2020. Medical visual question answering via conditional reasoning. In *ACM MM*.

Zhang, F.; Chen, H.; Zhu, Z.; Zhang, Z.; Lin, Z.; Qiao, Z.; Zheng, Y.; and Wu, X. 2025a. A survey on foundation language models for single-cell biology. In *ACL*.

Zhang, F.; Liu, T.; Chen, Z.; Peng, X.; Chen, C.; Hua, X.-S.; Luo, X.; and Zhao, H. 2024a. Semi-supervised knowledge transfer across multi-omic single-cell data. *NeurIPS*.

Zhang, F.; Liu, T.; Zhu, Z.; Wu, H.; Wang, H.; Zhou, D.; Zheng, Y.; Wang, K.; Wu, X.; and Heng, P.-A. 2025b. CellVerse: Do Large Language Models Really Understand Cell Biology? *NeurIPS*.

Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2025c. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. *ICCV*.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2024b. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D. A.; Okuno, T.; Nakata, Y.; Keutzer, K.; and Zhang, S. 2025d. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *CVPR*.

Zhu, H.; Liu, Y.; Fang, X.; Lu, G.; and Chen, B. 2025a. Cause-Effect Driven Optimization for Robust Medical Visual Question Answering with Language Biases. *IJCAI*.

Zhu, K.; Qin, Z.; Yi, H.; Jiang, Z.; Lao, Q.; Zhang, S.; and Li, K. 2025b. Guiding Medical Vision-Language Models with Diverse Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations. In *NAACL*.

Zhu, Z.; Zhang, Y.; Zhuang, X.; Zhang, F.; Wan, Z.; Chen, Y.; QingqingLong, Q.; Zheng, Y.; and Wu, X. 2025c. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *ACL*.

Zou, Y.; and Yin, Z. 2025. Alignment, Mining and Fusion: Representation Alignment with Hard Negative Mining and Selective Knowledge Fusion for Medical Visual Question Answering. In *CVPR*.