Enhancing Medical Large Vision-Language Models via Alignment Distillation

Aofei Chang[1], **Ting Wang**[2], **Fenglong Ma**[1*]

[1]Pennsylvania State University
[2]Stony Brook University
{aofei,fenglong}@psu.edu, twang@cs.stonybrook.edu

Abstract

Medical Large Vision-Language Models (Med-LVLMs) have shown promising results in clinical applications, but often suffer from hallucinated outputs due to misaligned visual understanding. In this work, we identify two fundamental limitations contributing to this issue: insufficient visual representation learning and poor visual attention alignment. To address these problems, we propose **MEDALIGN**, a simple, lightweight alignment distillation framework that transfers visual alignment knowledge from a domain-specific Contrastive Language-Image Pre-training (CLIP) model to Med-LVLMs. MEDALIGN introduces two distillation losses: a spatial-aware visual alignment loss based on visual token-level similarity structures, and an attention-aware distillation loss that guides attention toward diagnostically relevant regions. Extensive experiments on medical report generation and medical visual question answering (VQA) benchmarks show that MEDALIGN consistently improves both performance and interpretability, yielding more visually grounded outputs.

Code — https://github.com/Aofei-Chang/MedAlign

Introduction

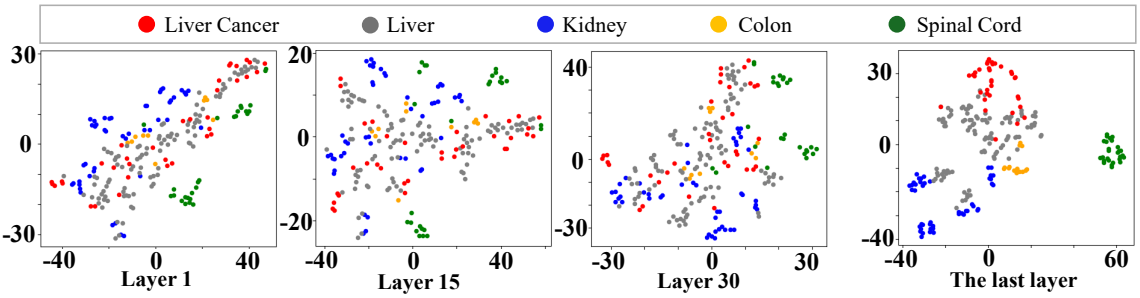
Medical Large Vision-Language Models (Med-LVLMs), such as LLaVA-Med-1.5 and HuatuoGPT-Vision, have shown strong potential in clinical applications (Li et al. 2024; Chen et al. 2024a,c; Thawakar et al. 2024; Moor et al. 2023). However, recent studies (Xia et al. 2024; Gu et al. 2024; Chen et al. 2024b; Chang et al. 2025a; Wang et al. 2024) have revealed that these models often produce inaccurate or hallucinated responses that fail to faithfully reflect the input medical images. To the best of our knowledge, **no** existing work has proposed targeted methods to mitigate hallucinations specifically in Med-LVLMs. Existing hallucination mitigation strategies primarily focus on general-purpose LVLMs and follow two main directions: (1) enhancing visual grounding and reducing over-reliance on textual input through contrastive decoding – applied at either the attention or input level (Leng et al. 2024; Favero et al. 2024; Liu, Zheng, and Chen 2025; Tu et al. 2025; Chen et al. 2025); and (2) correcting attention biases, such as the overemphasis on background elements or "register" tokens among visual inputs (Darcet et al. 2024; Woo et al. 2024; Gong et al. 2024). While these techniques may alleviate hallucinations, they do not explicitly improve the distribution of visual attention or ensure that the model focuses on clinically relevant regions. More critically, they overlook a key contributor to hallucination in Med-LVLMs: *the quality of the learned visual representations.*

Preliminary Analysis. To address these limitations, we conduct a preliminary analysis to investigate two fundamental factors driving hallucinations in Med-LVLMs: (1) the quality of the visual representations learned by the model, and (2) the alignment of visual attention during generation.

(1) Insufficient Visual Representation Learning. Unlike images in the general domain that contain diverse objects, medical images often feature recurring anatomical structures, such as the lungs, heart, and ribs in chest X-rays. Ideally, a well-trained Med-LVLM should learn similar representations for the same organ across different images. To evaluate the quality of visual representation learning in existing Med-LVLMs, we adopt LLaVA-Med-1.5 as a representative model. We randomly sample 100 abdominal CT scans from the SLAKE (Liu et al. 2021) dataset, which includes Region-of-Interest (RoI) annotations for image patches. For analysis, we extract the visual token representations from various layers of the Transformer-based large language model (LLM) used in LLaVA-Med-1.5 and visualize five key entities using t-SNE. The results, shown in Figure 1 (a), illustrate how visual representations evolve across layers. We can observe that LLaVA-Med-1.5 fails to clearly distinguish key entities in medical images, resulting in **entangled** and **dispersed** visual representations. For example, the representations of "*liver cancer*" are heavily mixed with those of "*liver*" and other nearby organs. These results suggest that current Med-LVLMs exhibit insufficient visual representation learning, particularly for clinically critical concepts, which may potentially leads to poor visual reasoning and increased risk of hallucinations.

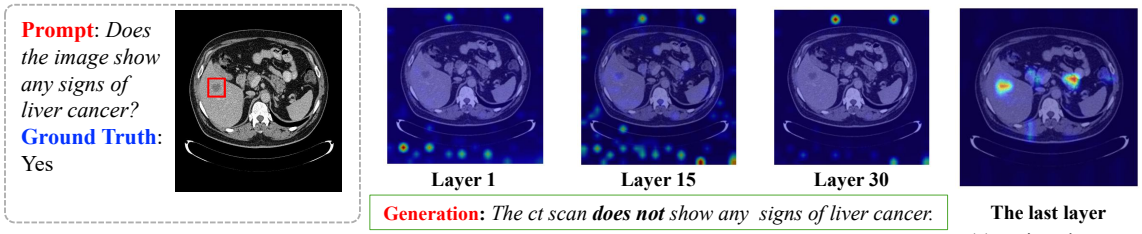
(2) Visual Attention Misalignment. A well-trained Med-LVLM should understand both the input image and the corresponding text prompt and assign higher attention weights to image regions relevant to the medical concepts mentioned in the prompt. However, as previously discussed, the issue of insufficient visual representation learning in Med-LVLMs leads to a secondary problem in the LLM component: visual

*Corresponding author.



(a) LLaVA-Med-1.5 (b) UniMed-CLIP

Figure 1: A t-SNE visualization of visual features derived from sampled abdominal CT scans using LLaVA-Med-1.5 and UniMed-CLIP.



(a) Input prompt and image (b) LLaVA-Med-1.5 (c) UniMed-CLIP

Figure 2: Attention visualizations from LLaVA-Med-1.5 and UniMed-CLIP on an abdominal CT scan. The red box (not part of the input) highlights the liver cancer region.

attention misalignment. As illustrated in Figure 2 (b), we analyze a specific prompt, "*Does the image show any signs of liver cancer*", by visualizing the average visual attention weights across H attention heads in the l-th layer. The visualization reveals that the high attention weights are not aligned with the relevant region (highlighted by the red box in the input image), leading the model to generate a hallucinated response, "*The ct scan **does not** show any signs of liver cancer.*", despite the ground truth being "*Yes*".

Motivations of Incorporating Domain-specific Guidance. We hypothesize that the general CLIP-based image encoder used in Med-LVLMs is a primary contributor to both insufficient visual representation learning and visual attention misalignment. Leveraging a more domain-adapted CLIP-based model, such as UniMed-CLIP (Khattak et al. 2024), which is trained on large-scale medical image–text data, could significantly enhance Med-LVLM performance. To explore this, we use UniMed-CLIP (large) as an example, which takes an image–text pair (I, P) as input. The output of the image encoder consists of a concatenation of the [cls] representation $\mathbf{E}_{[\text{cls}]} \in \mathbb{R}^b$ and patch-wise visual representations $\mathbf{E}_v \in \mathbb{R}^{M \times b}$, where M denotes the number of image patches and b is the hidden dimension.

As shown in Figure 1 (b), the visual features $\mathbf{E}_v$ extracted from the final layer of UniMed-CLIP exhibit improved semantic separation across medical concepts, suggesting more structured and discriminative visual representations. In addition, we extract the attention map $\mathbf{E}_a \in \mathbb{R}^M$, representing attention weights between the [cls] token and the M image patches from the last layer of the visual encoder. The visualization in Figure 2 (c) shows that UniMed-CLIP achieves stronger visual grounding by attending more precisely to diagnostically relevant regions. Together, these observations indicate that UniMed-CLIP provides more better representations and attention distributions than LLaVA-Med-1.5.

Given these findings, a straightforward strategy to enhance Med-LVLMs is to replace the original CLIP encoder with a domain-specific expert encoder like UniMed-CLIP. However, this requires re-training the visual projection layer and adapting the entire Med-LVLM to the new feature space using large-scale data, which is *computationally intensive*. This limitation motivates the design of a lightweight, non-invasive approach to transfer alignment knowledge from expert CLIP models without fully replacing the original visual encoder.

Our Approach. We propose MEDALIGN, a novel alignment distillation framework designed to enhance Med-LVLMs by transferring both visual representations and attention patterns from a domain-specific expert CLIP model. As illustrated in Figure 3, given an input image–prompt pair, we extract two alignment signals from the expert CLIP: (1) visual representations and (2) visual attention maps. These signals are then distilled into intermediate layers of the Med-LVLM to improve its alignment with clinically relevant content. To enable lightweight and non-invasive integration, we introduce two core components. First, a **spatial-aware visual alignment loss** captures the pairwise similarity structure among image patches - reflected in the expert CLIP's visual features - and transfers it to the Med-LVLM's internal representations. Second, an **attention-aware distillation loss** aligns the Med-LVLM's attention distributions with those derived from the

expert model. MEDALIGN does not require retraining or modifying visual encoders and provides a plug-and-play solution that integrates seamlessly into existing Med-LVLMs.

Contributions. In summary, our work makes the following contributions: (1) We conduct a detailed analysis of current Med-LVLMs and identify two key but underexplored sources of misalignment: misaligned visual representations and misaligned attention distributions. (2) We introduce MEDALIGN, a novel alignment distillation framework that transfers alignment knowledge from expert CLIP models to Med-LVLMs without requiring model retraining or fine-grained supervision. The framework features two lightweight loss functions: spatial-aware visual alignment and attention-aware distillation, for effective and interpretable knowledge transfer. (3) We validate MEDALIGN through extensive experiments on two report generation benchmarks and five medical VQA datasets, demonstrating consistent improvements over strong baselines in both task performance and visual interpretability. Our approach leads to better visual grounding and more clinically faithful outputs.

Related Work

While prior efforts such as A^{3}Tune (Chang et al. 2025b) and CoMT (Jiang et al. 2024) reduce hallucinations in Med-LVLMs, most focus on surface-level attention adjustments and overlook the core issue of multimodal misalignment. Since Med-LVLMs inherit architectures from general LVLMs, they exhibit similar hallucination behaviors, making several inference-time techniques transferable to the medical domain—such as improving visual grounding (Leng et al. 2024; Favero et al. 2024; Liu, Zheng, and Chen 2025; Yuan et al. 2024; Liang et al. 2024; Chen et al. 2025; Tu et al. 2025), correcting visual attention biases (Woo et al. 2024; Gong et al. 2024), and refining decoding (Huang et al. 2024; Chuang et al. 2024). However, resolving modality misalignment in Med-LVLMs remains largely unsolved.

The Proposed MEDALIGN

Overview

Based on the preliminary observations presented in the introduction, we propose a novel, lightweight, and straightforward distillation-based framework, named MEDALIGN, aimed at improving visual-text alignment in Med-LVLMs. MEDALIGN achieves this by transferring fine-grained alignment knowledge from a domain-specialized expert CLIP model to the target Med-LVLM. Specifically, two forms of knowledge extracted from the last layer of the expert CLIP image encoder are used to guide the learning process: the visual representations $\mathbf{E}_v \in \mathbb{R}^{M\times b}$ and the visual attention vector $\mathbf{E}_a \in \mathbb{R}^{M}$.

An overview of MEDALIGN is illustrated in Figure 3. Med-LVLMs take paired inputs of an image and a text prompt, denoted as (I, P), similar to general LVLMs (Liu et al. 2024). The medical image I is typically divided into N patches and encoded by a CLIP-based image encoder and a visual projection layer, producing visual embeddings $\mathbf{X}_v \in \mathbb{R}^{N\times d}$, where d is the token dimension used by the large language model (LLM). Meanwhile, the text prompt P is processed through a text embedding layer, yielding $\mathbf{X}_p \in \mathbb{R}^{T\times d}$, where T is the number of text tokens. The model then concatenates $\mathbf{X}_v$ and $\mathbf{X}_p$ and feeds the combined sequence into the L-layer Transformer-based LLM to generate the output. To address the issue of dispersed visual representations, we introduce a **spatial-aware visual alignment distillation** loss that transfers the relative similarity structure among image patches from the expert CLIP model to the Med-LVLM, using $\mathbf{E}_v$. To further align the attention distribution, we leverage $\mathbf{E}_a$ from the expert model to guide the visual attention learning in the Med-LVLM through an **attention-aware distillation** loss. Importantly, both forms of alignment supervision are applied only at a designated layer l within the Med-LVLM. Specifically, visual representation distillation is applied to the input of layer l, intervening at the output visual representations of layer $l-1$. This design enables a lightweight and non-invasive distillation process, avoiding full encoder replacement or costly end-to-end retraining.

Spatial-aware Visual Alignment via $\mathbf{E}_v$

A straightforward approach to aligning the visual representations from layer $l-1$ of the Med-LVLM (i.e., $\mathbf{X}_v^{l-1} \in \mathbb{R}^{N\times d}$) with those from the expert model (i.e., $\mathbf{E}_v \in \mathbb{R}^{M\times b}$) is to minimize the distance between them. However, this is impractical due to the mismatch in dimensionality between the two matrices. Moreover, even if the dimensions were aligned, directly forcing the representations to match could adversely affect the LLM generation, leading to degraded performance.

Representation Rotation. To address these challenges, we introduce a trainable rotation matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ to adaptively adjust the Med-LVLM visual representations. The transformed representation is computed as:

$$\tilde{\mathbf{X}}_v^{l-1} = (\mathbf{I} + \mathbf{W})\mathbf{X}_v^{l-1}, \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix. This formulation allows for a lightweight and smooth transformation, gently steering $\mathbf{X}_v^{l-1}$ without drastic changes to the original representation space. The goal of this rotation is to bring visually similar concepts closer in the feature space, thus mitigating the representation dispersion issue. As demonstrated in Figure 1 (a), the visual representations suffer from poor semantic structure. Inspired by techniques such as word embedding steering in LLMs (Han et al. 2024), which apply targeted modifications through small perturbations, our learnable matrix $\mathbf{W}$ serves as a controlled adjustment mechanism to enhance visual coherence in a minimally invasive manner.

Spatial-aware Visual Alignment. After obtaining the rotated visual representations $\tilde{\mathbf{X}}_v^{l-1} \in \mathbb{R}^{N\times d}$, we still face a dimensional mismatch with the expert visual features $\mathbf{E}_v \in \mathbb{R}^{M\times b}$. To resolve this, we first apply an interpolation function to adjust $\mathbf{E}_v$ such that it matches the number of patches of $\tilde{\mathbf{X}}_v^{l-1}$:

$$\tilde{\mathbf{E}}_v = \text{interpolate}(\mathbf{E}_v, (N, b)). \tag{2}$$

However, directly minimizing the distance between $\tilde{\mathbf{X}}_v^{l-1}$ and $\tilde{\mathbf{E}}_v \in \mathbb{R}^{N\times b}$ may still negatively affect LLM performance due to rigid alignment. Instead, we propose leveraging the

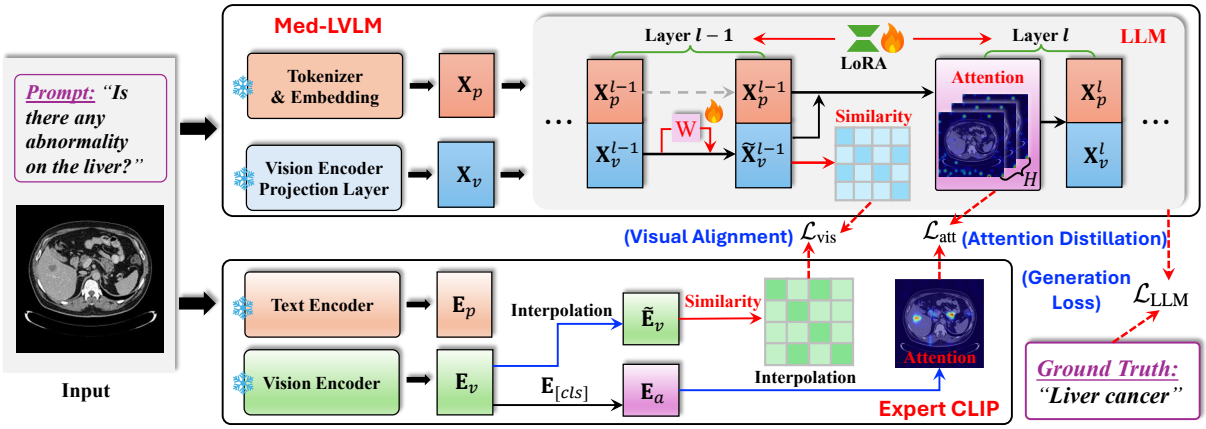


Figure 3: The overview of the proposed MEDALIGN that uses an expert CLIP model as a reference to guide the fine-tuning of Med-LVLMs via a spatial-aware visual alignment loss and an attention-aware distillation loss.

pairwise similarity structure among patch representations as a softer, more semantically meaningful alignment signal. Intuitively, patches depicting the same organ should have similar visual embeddings, resulting in higher similarity scores, while unrelated patches should be dissimilar.

To capture these structural relationships, we compute pairwise cosine similarity matrices for both the expert features $\tilde{\mathbf{E}}_v$ and the rotated Med-LVLM features $\tilde{\mathbf{X}}_v^{l-1}$:

$$\mathbf{S}^e_{i,j} = \cos(\tilde{\mathbf{E}}_v[i], \tilde{\mathbf{E}}_v[j]), \text{and } \mathbf{S}^x_{i,j} = \cos(\tilde{\mathbf{X}}_v^{l-1}[i], \tilde{\mathbf{X}}_v^{l-1}[j]). \tag{3}$$

The spatial-aware visual alignment distillation loss is then defined as the mean squared error between the two similarity matrices:

$$\mathcal{L}_{\text{vis}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\mathbf{S}^e_{i,j} - \mathbf{S}^x_{i,j}\right)^2. \tag{4}$$

This loss encourages the model to capture the same structural similarity patterns among image patches as the expert model, promoting semantically coherent and spatially aware visual representations without enforcing direct value matching.

Attention-aware Alignment via $\mathbf{E}_a$

In addition to refining visual representations, Med-LVLMs also suffer from visual attention misalignment, where attention weights fail to properly focus on diagnostically relevant regions. In each Transformer layer l of a Med-LVLM, H attention heads are used, and each head h produces a visual attention vector $\mathbf{M}^l_{a,h} \in \mathbb{R}^N$, where N is the number of image patches. The average visual attention across all heads at layer l is computed as $\tilde{\mathbf{M}}^l_a = \frac{1}{H}\sum_{h=1}^{H} \mathbf{M}^l_{a,h}$.

However, aligning this attention distribution with the expert attention vector $\mathbf{E}_a \in \mathbb{R}^M$ is not straightforward due to a mismatch in patch numbers (i.e., M may not be equal to N). To address this, we apply interpolation to resize $\mathbf{E}_a$ to match the Med-LVLM patch resolution, yielding an interpolated expert attention vector $\tilde{\mathbf{E}}_a \in \mathbb{R}^N$.

To guide attention learning, we treat $\tilde{\mathbf{E}}_a$ as the teacher and $\tilde{\mathbf{M}}^l_a$ as the student. We then define the attention-aware alignment loss using the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\text{att}} = \text{KL}(\tilde{\mathbf{E}}_a || \tilde{\mathbf{M}}^l_a), \tag{5}$$

where both distributions are normalized using softmax. This loss encourages the Med-LVLM to mimic the expert model's attention focus, promoting more accurate grounding of visual information and improving the quality of generated responses.

Final Objective

We implement MEDALIGN using parameter-efficient fine-tuning based on LoRA (Hu et al. 2022), applied to all linear layers in the LLM of the Med-LVLM. During fine-tuning, we jointly optimize the proposed alignment losses alongside the standard language modeling objective $\mathcal{L}_{\text{LLM}}$, defined as:

$$\mathcal{L}_{\text{LLM}} = -\sum_{t=1}^{T} \log p_t(y_t \mid \mathbf{X}_v, \mathbf{X}_p, y_{<t}; \Theta, \Phi), \tag{6}$$

where Θ denotes the *frozen* parameters of the base LLM, and Φ denotes the trainable parameters introduced by LoRA and our designed alignment modules. The final training target is:

$$\mathcal{L} = \mathcal{L}_{\text{LLM}} + \alpha\mathcal{L}_{\text{vis}} + \beta\mathcal{L}_{\text{att}}, \tag{7}$$

where α and β are hyperparameters that balance alignment distillation with the language modeling loss $\mathcal{L}_{\text{LLM}}$.

Experiments

Experimental Setups

Med-LVLMs and Expert CLIPs. We evaluate our method on two representative Med-LVLMs: LLaVA-Med-1.5 and HuatuoGPT-Vision-7B, with Beam search as the default decoding strategy. In our main experiments, we use UniMed-CLIP (ViT-L/14, 336px) as the expert CLIP model. To further examine the impact of expert model selection, we also investigate additional configurations including UniMed-CLIP (ViT-B/16, 224px) and BiomedCLIP (Zhang et al. 2023).

Evaluation Tasks and Datasets. We evaluate our method on two core tasks in medical applications of Med-LVLMs:

Dataset	Metric	Method										
		Greedy	Beam	Nucleus	VCD	DoLa	OPERA	AVISC	M3ID	DAMRO	PAI	MEDALIGN
IU-Xray	BLEU	9.34	10.21	8.19	9.10	9.03	10.01	7.47	8.35	9.10	6.92	**10.73**
	ROUGE-L	28.17	28.64	26.28	27.80	27.50	28.57	24.78	27.05	27.80	24.36	**29.10**
	METEOR	31.76	34.23	30.72	32.13	31.24	34.10	30.88	32.17	32.13	30.84	**36.30**
	BERTScore	88.53	88.60	88.19	88.36	88.39	88.51	87.77	88.19	88.36	87.44	**88.67**
	CheXbert	55.16	55.84	53.86	54.34	53.89	55.01	52.11	53.78	54.34	50.09	**56.27**
	RadGraph	21.86	22.47	20.17	21.65	21.04	22.59	19.71	21.19	21.65	18.26	**23.51**
	RaTEScore	58.66	59.78	58.29	58.29	57.84	59.33	56.49	57.86	58.29	55.16	**60.49**
MIMIC-CXR	BLEU	4.22	4.16	3.61	3.73	3.98	4.04	3.55	3.74	3.73	3.87	**4.76**
	ROUGE-L	18.11	18.26	16.93	17.25	17.37	18.03	16.15	17.08	17.25	16.92	**19.32**
	METEOR	20.54	19.79	19.56	20.20	20.90	19.80	19.71	20.74	20.20	20.86	**22.02**
	BERTScore	85.79	85.84	85.35	85.53	85.47	85.73	84.70	85.25	85.53	84.59	**86.02**
	CheXbert	27.49	27.56	26.72	25.80	26.33	28.15	25.84	26.91	25.80	26.80	**29.53**
	RadGraph	11.75	10.85	9.96	10.69	11.39	10.60	10.09	10.67	10.69	10.57	**12.82**
	RaTEScore	43.39	42.38	41.66	42.54	42.71	41.93	42.07	42.37	42.54	42.48	**44.96**

Table 1: Report generation results of HuatuoGPT-Vision-7B fine-tuned with LoRA. Best results are highlighted in **bold**.

Dataset	Metric	Method										
		Greedy	Beam	Nucleus	VCD	DoLa	OPERA	AVISC	M3ID	DAMRO	PAI	MEDALIGN
IU-Xray	BLEU	9.36	9.54	7.80	8.83	8.93	9.23	5.57	8.44	8.21	8.52	**10.31**
	ROUGE-L	27.57	28.41	26.72	27.36	26.94	27.48	21.71	26.21	25.77	26.97	**29.01**
	METEOR	27.91	35.40	30.33	31.77	25.74	34.17	26.84	30.86	30.58	28.63	**35.22**
	BERTScore	88.55	88.45	88.28	88.30	88.42	88.17	87.34	88.20	88.09	88.42	**88.66**
	CheXbert	52.44	53.70	52.73	51.86	52.27	51.65	47.32	51.13	50.10	52.22	**55.62**
	RadGraph	21.28	22.43	20.85	22.02	20.63	21.37	16.87	20.77	22.33	20.99	**23.29**
	RaTEScore	58.61	59.65	57.84	58.93	58.10	57.89	53.66	59.37	57.31	58.21	**59.99**
MIMIC-CXR	BLEU	3.50	3.66	3.48	3.74	3.48	3.56	3.31	3.14	3.42	3.63	**4.51**
	ROUGE-L	16.49	16.85	16.35	16.88	16.45	16.77	16.36	16.13	16.63	16.65	**18.43**
	METEOR	18.71	20.68	18.93	19.03	18.66	20.10	18.64	18.52	18.87	18.61	**20.80**
	BERTScore	85.54	85.51	85.50	85.56	85.54	85.46	85.48	85.39	85.48	85.60	**86.00**
	CheXbert	23.43	25.00	22.21	22.98	23.34	24.31	23.31	22.42	23.30	24.51	**25.67**
	RadGraph	9.63	9.91	9.27	9.56	9.52	9.81	9.02	9.15	9.46	9.60	**10.92**
	RaTEScore	40.49	41.46	40.08	40.93	40.49	41.33	40.36	39.74	40.80	40.49	**42.03**

Table 2: Performance on report generation benchmarks using LLaVA-Med-1.5 fine-tuned with LoRA.

medical report generation and medical visual question answering (VQA). For the report generation task, we use MIMIC-CXR (Johnson et al. 2019) and IU-Xray (Demner-Fushman et al. 2016). For the VQA task, we adopt a diverse set of benchmark datasets, including SLAKE (Liu et al. 2021), VQA-RAD (Lau et al. 2018), PathVQA (He et al. 2020), IU-Xray, and OmniMedVQA (Hu et al. 2024).

Baselines. We compare against popular hallucination mitigation methods, including decoding strategies and contrastive decoding techniques. The decoding baselines include Greedy decoding, Nucleus sampling, and Beam search. For contrastive decoding methods, we evaluate against recent techniques including VCD (Leng et al. 2024), OPERA (Huang et al. 2024), DoLa (Chuang et al. 2024), AVISC (Woo et al. 2024), M3ID (Favero et al. 2024), DAMRO (Gong et al. 2024) and PAI (Liu, Zheng, and Chen 2025).

Metrics. For the *medical report generation* task, we adopt standard text generation metrics, including BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), and BERTScore (Zhang et al. 2020). Additionally, we include domain-specific evaluation metrics tailored to medical report generation: CheXbert (Smit et al. 2020), RadGraph (Jain et al. 2021), and RaTEScore (Zhao et al. 2024). For the *medical VQA* task, we follow the evaluation of LLaVA-Med (Li et al. 2024), reporting Accuracy for close-ended VQA and Recall for open-ended VQA.

Medical Report Generation Results

Table 1 and Table 2 present the evaluation of MEDALIGN on medical report generation using HuatuoGPT-Vision-7B and LLaVA-Med-1.5, respectively. We report both traditional generation metrics (e.g., BLEU, ROUGE-L) and domain-specific metrics such as RaTEScore. All baselines are applied to LoRA-tuned models, as the original Med-LVLMs perform poorly on report generation tasks. It is shown that MEDALIGN outperforms all baselines across both models and most evaluation metrics. Notably, with HuatuoGPT-Vision-7B, MEDALIGN achieves the best scores on all metrics, including substantial improvements in report-specific metrics, demonstrating stronger generation quality.

Med-LVLM	Method	SLAKE		VQA-RAD		PathVQA		IU-Xray	OmniMedVQA
		Open	Close	Open	Close	Open	Close	Close	Close
HuatuoGPT-Vision + LoRA	Greedy	85.57	90.14	41.37	76.77	37.08	93.31	85.33	91.33
	Beam	86.03	90.14	43.75	76.77	37.16	93.16	85.33	91.33
	Nucleus	84.75	90.42	38.98	77.56	34.03	93.45	85.59	90.50
	VCD	84.73	89.58	40.81	77.95	34.97	93.01	85.08	90.50
	DoLa	85.30	89.86	42.08	77.95	35.92	91.71	85.71	91.33
	OPERA	86.01	90.14	43.57	76.77	37.40	93.16	85.20	91.37
	AVISC	84.16	91.27	38.97	78.35	35.48	93.34	85.33	90.76
	M3ID	85.00	89.86	41.79	77.95	35.93	93.39	85.33	91.52
	DAMRO	84.73	89.58	40.81	77.95	34.97	93.01	85.08	90.50
	PAI	84.40	90.14	41.83	77.17	35.87	92.95	85.71	90.99
	MEDALIGN	**86.85**	**92.39**	**43.75**	**78.74**	**38.49**	**93.63**	**86.22**	**93.60**
LLaVA-Med-1.5 + LoRA	Greedy	82.97	88.45	36.70	74.41	37.98	93.22	84.69	91.03
	Beam	83.27	88.45	36.95	74.41	38.32	92.98	84.82	90.99
	Nucleus	82.75	86.76	36.58	73.62	35.13	92.95	85.20	90.69
	VCD	82.89	86.76	35.12	73.62	35.76	92.92	85.59	90.69
	DoLa	82.97	88.45	36.70	74.41	37.95	93.22	84.69	91.03
	OPERA	83.07	88.45	36.91	74.80	38.64	93.28	84.69	91.03
	AVISC	83.26	87.32	36.58	74.41	36.52	93.34	84.82	91.14
	M3ID	83.20	87.04	35.90	74.02	35.53	92.63	84.44	90.69
	DAMRO	82.89	86.76	35.12	73.62	35.76	92.92	85.59	90.69
	PAI	83.50	**89.01**	34.54	74.41	37.46	93.24	84.69	90.92
	MEDALIGN	**84.85**	**89.01**	**39.62**	**74.80**	**38.65**	**93.51**	**85.84**	**93.38**

Table 3: Performance comparison on medical VQA benchmarks using HuatuoGPT-Vision-7B and LLaVA-Med-1.5. "Open" denotes open-ended answers; "Close" refers to close-ended (e.g., yes/no or multiple-choice) responses.

Medical VQA Results

Table 3 reports results on five medical VQA benchmarks using HuatuoGPT-Vision-7B and LLaVA-Med-1.5. MEDALIGN consistently surpasses all baselines on both open- and close-ended questions, demonstrating strong and stable alignment across datasets. Although some methods excel in isolated cases (e.g., PAI and AVISC on SLAKE close-ended), their performance varies widely, revealing limited generalization. In contrast, MEDALIGN maintains balanced gains and shows particularly notable improvements on open-ended tasks—for example, on LLaVA-Med-1.5, +1.58 on SLAKE and +2.67 on VQA-RAD over the strongest baselines. This trend echoes its strong report-generation results, indicating that open-ended generation is especially sensitive to alignment quality and highlighting the effectiveness of our alignment distillation.

Model Design Analysis

Given that open-ended tasks better reflect the model's ability to interpret and describe medical images and are more sensitive to alignment quality, we focus our analysis primarily on the SLAKE open-ended subset.

Ablation Study on Distillation Design. Our proposed distillation framework employs two alignment loss terms, $\mathcal{L}_{\text{vis}}$ and $\mathcal{L}_{\text{att}}$, as defined in Eq. (7). To evaluate the individual contributions of each component, we conduct an ablation study by selectively removing each loss term. Specifically, we denote the variants as MEDALIGN−att (removing $\mathcal{L}_{\text{vis}}$) and MEDALIGN−vis (removing $\mathcal{L}_{\text{att}}$). We also include a baseline using LoRA-based fine-tuning with Beam search, denoted as LoRA-FT. As shown in Figure 4 (A), both loss terms individually improve performance over LoRA-FT, while combining them yields the best results, confirming their complementary benefits. In Figure 4 (B), directly using UniMed-CLIP features without distillation leads to performance degradation, likely due to feature mismatch. These results show the effectiveness of our distillation design in enhancing Med-LVLM performance without directly perturbing the feature space.

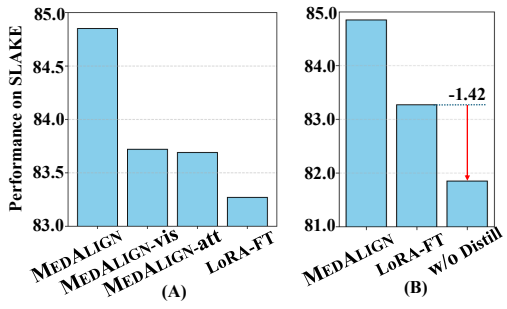


Figure 4: Ablation study on distillation design. (A) Impact of removing loss designs. (B) Performance when directly using UniMed-CLIP features without distillation ("w/o Distill").

Qualitative Analysis. To evaluate the effectiveness of our loss design, we examine changes in visual representations and attention after alignment distillation. For visual representations, we visualize t-SNE projections of features from layers 1–30 (Figure 5). Because alignment is applied at layer 20, feature separability improves noticeably from that layer onward,

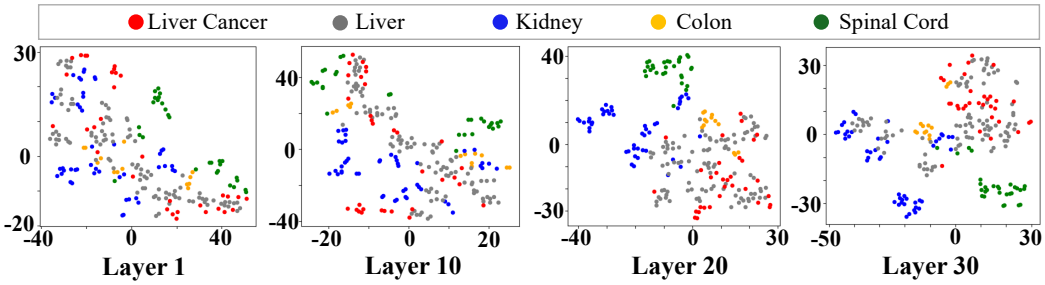


Figure 5: A t-SNE visualization of visual features from multiple layers of LLaVA-Med-1.5 after applying MEDALIGN.

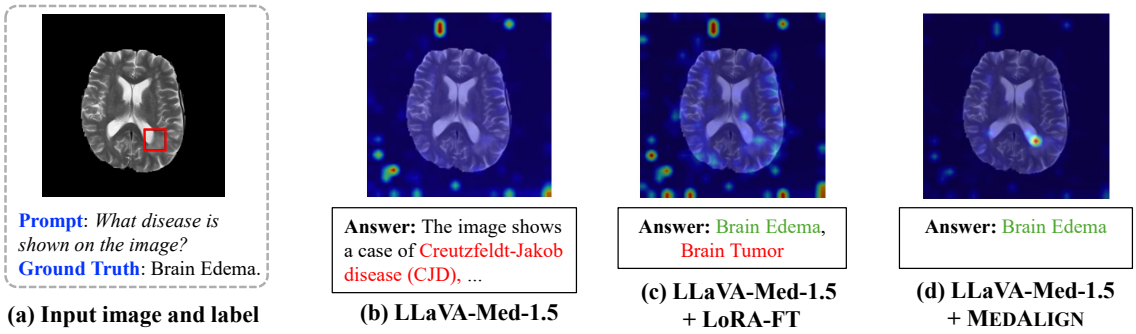


Figure 6: Comparison of overall attention distributions on the visual input averaged across all layers on a Brain MRI VQA example from SLAKE. Red text indicates hallucinated answers.

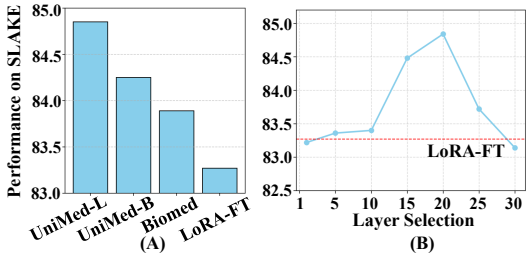


Figure 7: (A) Comparison of expert CLIP models. (B) Results of layer selection.

with enhancements even in earlier layers (e.g., layer 10), indicating backward propagation of alignment benefits. For visual attention, we visualize average attention on a Brain MRI VQA example (Figure 6). Models trained with MEDALIGN produce more accurate answers and concentrate attention on clinically relevant regions, confirming the effectiveness of our attention alignment.

Expert CLIP Model Selection

We use UniMed-CLIP (ViT-L/14, 336px), denoted **UniMed-L**, as our default expert. To assess expert choice and distillation generalizability, we also test UniMed-CLIP (ViT-B/16), **UniMed-B**, and BiomedCLIP (ViT-B/16), **Biomed**. Both UniMed-B and Biomed produce only 196 visual tokens—significantly fewer than the 576 used in LLaVA-Med-1.5 and HuatuoGPT-Vision-7B. To ensure compatibility, we apply interpolation on both the image representations and attention maps.

As shown in Figure 7 (A), all expert CLIP models outperform LoRA-FT and prior baselines (Table 3), with UniMed-L achieving the highest gains due to its finer token granularity. In contrast, UniMed-B and Biomed offer weaker supervision. These results demonstrate the flexibility of our framework and the value of strong visual alignment priors.

Distillation Layer Selection

In our experiments, we set the distillation layer to $l = 20$. To investigate the effect of layer selection, we varied the position of alignment. As shown in Figure 7 (B), applying MEDALIGN at early layers leads to performance degradation, likely due to disruption of low-level representations. Performance peaks in the middle layers, precisely where fine-grained multimodal interactions are known to occur (Jiang et al. 2025; Neo et al. 2024). Applying the alignment losses at later layers also results in a drop in performance, likely due to limited representational flexibility in these stages.

Conclusion

In this work, we present MEDALIGN, a lightweight alignment distillation framework that enhances Med-LVLMs by transferring visual representation and attention alignment knowledge from expert medical CLIP models. Through designed distillation objectives, MEDALIGN improves visual grounding and output fidelity without requiring fine-grained annotations. Experiments on multiple benchmarks demonstrate consistent gains in both performance and interpretability, offering a practical path toward more reliable Med-LVLMs.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and helpful suggestions. Dr. Ma is partially supported by the National Science Foundation under Grant No. 2238275 and the National Institutes of Health under Grant No. R01AG077016. Dr. Wang is partially supported by the National Science Foundation under Grant No. 2405136 and 2406572.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chang, A.; Huang, L.; Bhatia, P.; Kass-Hout, T.; Ma, F.; and Xiao, C. 2025a. MedHEval: Benchmarking Hallucinations and Mitigation Strategies in Medical Large Vision-Language Models. *arXiv preprint arXiv:2503.02157*.

Chang, A.; Huang, L.; Boyd, A. J.; Bhatia, P.; Kass-Hout, T.; Xiao, C.; and Ma, F. 2025b. Focus on What Matters: Enhancing Medical Vision-Language Models with Automatic Attention Alignment Tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9357–9372. Vienna, Austria: Association for Computational Linguistics.

Chen, B.; Lyu, X.; Gao, L.; Song, J.; and Shen, H. T. 2025. Attention Hijackers: Detect and Disentangle Attention Hijacking in LVLMs for Hallucination Mitigation. *arXiv preprint arXiv:2503.08216*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; et al. 2024a. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7346–7370.

Chen, J.; Yang, D.; Wu, T.; Jiang, Y.; Hou, X.; Li, M.; Wang, S.; Xiao, D.; Li, K.; and Zhang, L. 2024b. Detecting and Evaluating Medical Hallucinations in Large Vision Language Models. *arXiv preprint arXiv:2406.10185*.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Veen, D. V.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; Tsai, E.; Johnston, A.; Olsen, C.; Abraham, T. M.; Gatidis, S.; Chaudhari, A. S.; and Langlotz, C. 2024c. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; and He, P. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multimodal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14303–14312.

Gong, X.; Ming, T.; Wang, X.; and Wei, Z. 2024. DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7696–7712.

Gu, Z.; Yin, C.; Liu, F.; and Zhang, P. 2024. MedVH: Towards Systematic Evaluation of Hallucination for Large Vision Language Models in the Medical Context. *arXiv preprint arXiv:2407.02730*.

Han, C.; Xu, J.; Li, M.; Fung, Y.; Sun, C.; Jiang, N.; Abdelzaher, T.; and Ji, H. 2024. Word Embeddings Are Steers for Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16410–16430.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Jiang, Y.; Chen, J.; Yang, D.; Li, M.; Wang, S.; Wu, T.; Li, K.; and Zhang, L. 2024. CoMT: Chain-of-Medical-Thought Reduces Hallucination in Medical Report Generation.

Jiang, Z.; Chen, J.; Zhu, B.; Luo, T.; Shen, Y.; and Yang, X. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25004–25014.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz,

S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Khattak, M. U.; Kunhimon, S.; Naseer, M.; Khan, S.; and Khan, F. S. 2024. UniMed-CLIP: Towards a Unified Image-Text Pretraining Paradigm for Diverse Medical Imaging Modalities. *arXiv preprint arXiv:2412.10372*.

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*, 5(1): 1–10.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Liang, X.; Yu, J.; Mu, L.; Zhuang, J.; Hu, J.; Yang, Y.; Ye, J.; Lu, L.; Chen, J.; and Hu, H. 2024. Mitigating Hallucination in Visual-Language Models via Re-balancing Contrastive Decoding. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 482–496. Springer.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

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 *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, S.; Zheng, K.; and Chen, W. 2025. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, 125–140. 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)*, 353–367. PMLR.

Neo, C.; Ong, L.; Torr, P.; Geva, M.; Krueger, D.; and Barez, F. 2024. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1500–1519.

Thawakar, O. C.; Shaker, A. M.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. 2024. XrayGPT: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, 440–448.

Tu, C.; Ye, P.; Zhou, D.; Bai, L.; Yu, G.; Chen, T.; and Ouyang, W. 2025. Attention Reallocation: Towards Zero-cost and Controllable Hallucination Mitigation of MLLMs. *arXiv preprint arXiv:2503.08342*.

Wang, J.; Luo, J.; Ye, M.; Wang, X.; Zhong, Y.; Chang, A.; Huang, G.; Yin, Z.; Xiao, C.; Sun, J.; et al. 2024. Recent advances in predictive modeling with electronic health records. In *IJCAI: proceedings of the conference*, volume 2024, 8272.

Woo, S.; Kim, D.; Jang, J.; Choi, Y.; and Kim, C. 2024. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models. *arXiv preprint arXiv:2405.17820*.

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. *Advances in Neural Information Processing Systems*, 37: 140334–140365.

Yuan, F.; Qin, C.; Xu, X.; and Li, P. 2024. HELPD: Mitigating Hallucination of LVLMs by Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1768–1785.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. 2023. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. RaTEScore: A Metric for Radiology Report Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15004–15019.