IS LLAVA-MED VISUALLY GROUNDED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical Vision-Language Models (Medical VLMs) offer a scalable pathway to support radiological workflows by automating report generation and clinical question answering. LLaVA-Med is one of the most widely adopted Medical VLMs. It is instruction-tuned on PubMed Central, enabling the original LLaVA model to adapt to the medical domain. However, we identify that LLaVA-Med is susceptible to a critical failure mode: rather than learning genuine visual-textual coherence, it often exploits biased language priors, effectively memorizing statistical patterns in clinical text and replaying them at inference time with limited regard to the actual image content. This undermines diagnostic reliability and poses a direct risk to patient safety in real-world clinical deployment. In this paper, we first propose a diagnostic framework to systematically identify and analyze the underlying causes of this failure mode. Building on these insights, we develop interventions that mitigate linguistic bias and encourage LLaVA-Med to generate responses that are more faithfully grounded in visual evidence. Results on out-of-distribution datasets like VQA-RAD, SLAKE and PathVQA support the claim.

1 INTRODUCTION

The past few years have witnessed a paradigm shift in artificial intelligence with the emergence of Vision-Language Models (VLMs) (Radford et al., 2021; Liu et al., 2023b) which are deep neural networks (Goodfellow et al., 2016) capable of jointly processing and reasoning over visual and textual information within a single unified architecture. Unlike earlier computer vision models that operated purely on pixels, or language models that operated purely on text, VLMs learn a shared representational space in which images and their corresponding linguistic descriptions are mutually grounded. This is typically achieved through large-scale pretraining on web-scale image-text pairs, where a visual encoder (often a Vision Transformer) and a language model are aligned via contrastive objectives (Radford et al., 2021) or autoregressive next-token prediction conditioned on visual input (Alayrac et al., 2022; Liu et al., 2023a). Formally, let $x_v \in \mathcal{X}_v$ denote a medical image drawn from the visual input space and $x_t \in \mathcal{X}_t$ denote the textual prompt or question. Let $y = (y_1, y_2, \ldots, y_T)$ be the generated token sequence. Medical VLMs define a conditional autoregressive distribution: $P_\theta(y_t \mid y_{<t},\, x_v,\, x_t)$ where θ denotes the model parameters.

Foundational models like CLIP, Flamingo (Alayrac et al., 2022), and LLaVA (Liu et al., 2023a) excel at zero-shot and few-shot across tasks of image captioning, visual question answering (VQA), and image grounded instruction following. Despite these capabilities, general-purpose VLMs transfer poorly to the clinical domain. Medical images contain subtle anatomical structures and fine-grained pathological patterns that differ substantially from natural images (Zhang et al., 2023b; Chen et al., 2024a), while radiology reports employ specialised terminology, structured negation, and diagnostic uncertainty absent in conversational text. This limitation motivated development of medical VLMs these models are fine-tuned on biomedical image-report pairs to align visual representations with medical language. Models such as LLaVA-Med (Li et al., 2024) have demonstrated improved performance on report generation, visual question answering, abnormality localisation, and multimodal clinical retrieval (Chen et al., 2024b). However, this raises a fundamental question that this paper seeks to answer: *Does the Medical VLM actually use the image to generate an answer to the medical question?* Powerful language priors acquired from biomedical corpora may allow the model to generate clinically plausible responses without genuinely grounding them in visual evidence. Such behaviour poses a serious patient safety risk, potentially leading to hallucinated findings, missed abnormalities, or confident diagnoses unsupported by the input image. This phenomenon is referred

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Q: What type of carcinoma is present?
A: The image shows a histopathological examination of a tissue sample, which reveals the presence of a carcinoma. Histopathology is a technique used to study the microscopic structure of tissues in order to diagnose diseases, such as cancer. In this case, the presence of carcinoma indicates that there is an abnormal growth of cells in the tissue sample



Q: Can you see carcinoma in this image?
A: The image is a CT scan of the abdomen, and it shows the presence of carcinoma. Carcinoma is a type of cancer that originates in the epithelial cells, which are the cells that line the surfaces of organs and tissues throughout the body. In this case, the carcinoma is located in the abdomen.

Figure 1: LLaVA-Med responses to white and black visual inputs, illustrating responses persist even in absence of meaningful visual information. Additional results are present in Table 6 in appendix.

to as *linguistic bias*, and occurs when model outputs are driven more by learned statistical language patterns than by visual evidence. Unlike classical memorization, the model generates clinically plausible responses based on prior knowledge rather than learning fixed mapping from image to a report. In this work, we pursue three objectives: (i) Identifying whether LLaVA-Med has linguistic bias. (ii) Develop quantitative methods to detect and localize linguistic bias in LLaVA-Med. (iii) Design and evaluate architectural interventions in LLaVA-Med that mitigate linguistic bias and improve its visual grounding. Specifically, our contributions are as follows:

1. Through controlled image-ablation experiments (black (B) and white (W) null inputs), we show that LLaVA-Med produces clinically plausible responses even in the absence of meaningful visual information, revealing a strong reliance on language priors (Figure 1).
2. A diagnostic framework for visual grounding: We propose the Visual Fidelity Index (VFI) and Effective Representation Quality (ERQ), and use cosine similarity to quantify visual grounding of LLaVA-Med and localize linguistic bias during response generation.
3. We propose (LLaVA-MedFormer) LLaVA-MF, which replaces the CLIP encoder with BiomedCLIP and introduce a Q-Former-based alignment module inspired by (Li et al., 2023b), resulting in stronger visual grounding and reduced linguistic bias.
4. Experiments on in-domain datasets like PubMed Central and out-domain datasets like PathVQA, VQA-RAD, and SLAKE demonstrate consistent improvements.

2 Related Work

Although VLMs have demonstrated remarkable success, they encounter certain critical failures. Yuksekgonul et al. (2023) develop a benchmark to show that VLMs often recognize individual concepts but struggle to bind attributes and relations to the correct objects, revealing weak compositional understanding. Golovanevsky et al. (2025) study the competition between learned priors and visual evidence using Visual CounterFact, a dataset in which visually realistic images deliberately contradict common world-knowledge associations. Their layer-wise analysis shows that predictions initially reflect learned priors but become more influenced by visual evidence in later layers. Deng et al. (2025) further study conflicts between textual and visual evidence across four vision-centric tasks. By systematically varying the textual input and evaluating ten VLMs, they find that models often place greater weight on text than on visual evidence, with token order and text relevance further affecting this bias. Such failures are particularly problematic in the medical domain, where clinically meaningful evidence often consists of subtle spatial, anatomical, or pathological details. MedVH evaluates medical LVLMs by replacing the relevant image with unrelated or adversarial images and testing whether the model recognizes the resulting image-question mismatch (Rohrbach et al., 2018; Gu et al., 2024). More recently, Cao et al. (2025) introduce cross-modal guidance, which masks influential image tokens at selected layers to reduce visual information and uses the resulting change in output distributions to increase the influence of visual evidence during decoding. Jayaraman

et al. (2024) similarly study whether VLMs follow learned associations by manipulating contextual cues and analyzing model responses, showing that such associations can influence the output even when visual evidence is insufficient. Our work differs from these approaches in three aspects. First, existing studies focus on detecting problems which are different from linguistic bias. Second, they mitigate it using conventional strategies such as early stopping and temperature control. Third, their evaluation is conducted on CLIP-based models, which do not incorporate the actual vision-language architecture comprising both a visual encoder and a large language model (LLM). In contrast, we investigate linguistic bias in a full-fledged VLM based on the LLaVA architecture, where visual representations are explicitly coupled with an LLM. To the best of our knowledge, ours the first study to empirically identify and characterize linguistic bias in LLaVA-Med and develop targeted interventions to mitigate it.

3 METHODOLOGY

3.1 PRELIMINARIES

LLaVA-Med is a biomedical adaptation of the LLaVA framework, obtained through instruction fine-tuning on a curated corpus of biomedical image-text pairs derived from PubMed Central figure-caption data. It supports multiple medical imaging modalities, including radiology, pathology, and clinical photography, and has become a widely adopted benchmark for medical visual question answering. In a typical LLaVA-Med pipeline (Figure 4 - Supplementary material), a frozen CLIP vision encoder (E) extracts image embeddings from the input, which are mapped into the LLM's token embedding space via a trainable projection layer (P). The projected embeddings are concatenated with text tokens and passed to the language model (e.g., Vicuna/Mistral-7B) for autoregressive generation. Training proceeds in two stages: Stage 1 pre-trains the P with both the E and LLM frozen, and Stage 2 performs end-to-end fine-tuning of the P and LLM while keeping the E frozen.

3.2 PROBLEM STATEMENT

Despite the impressive performance of LLaVA-Med in tasks including report generation and medical VQA, a fundamental question remains unanswered: *Is LLaVA-Med genuinely grounding its responses in the input image?* We hypothesize that VLMs fine-tuned on biomedical corpora acquire strong statistical priors over the co-occurrence of symptoms, diagnoses, and clinical terminology. Consequently, they may generate radiologically plausible reports and clinically convincing responses by relying primarily on these learned language priors rather than the visual evidence contained in the input image (Leng et al., 2023; Ghosh et al., 2024)(Figure 1). In the clinical setting, such behaviour poses a patient safety risk, leading to hallucinated findings, missed abnormalities, or confident diagnoses that are unsupported by the image (Mahdavi et al., 2026; Sarkar et al., 2025). We refer to this failure mode as linguistic bias, wherein model outputs are driven predominantly by training-time statistical language patterns instead of inference-time visual evidence. To formally characterize this phenomenon, we first define the notions of visual grounding and linguistic bias.

Definition 1 (Visual Grounding) *A Medical VLM is visually grounded at token step t if the generated distribution depends meaningfully on the visual input. Formally, for an informative image x_v and a null or uninformative image $\tilde{x}_v \in \mathcal{X}_v$: $P_\theta(y_t \mid y_{<t}, x_v, x_t) \neq P_\theta(y_t \mid y_{<t}, \tilde{x}_v, x_t)$ where $x_v \in \mathcal{X}_v$ denotes the input medical image drawn from the visual input space, $x_t \in \mathcal{X}_t$ denotes the textual prompt or clinical query drawn from the textual input space, and $y = (y_1, y_2, \ldots, y_T) \in \mathcal{Y}$ denotes the autoregressively generated output token sequence.*

Definition 2 (Linguistic Bias) *A Medical VLM exhibits linguistic bias if its output distribution is governed primarily by language priors encoded in θ rather than by the visual input x_v. Equivalently, the model's predictions depend weakly on the image i.e., $P_\theta(y_t \mid y_{<t}, x_v, x_t) \approx P_\theta(y_t \mid y_{<t}, \tilde{x}_v, x_t)$. In the degenerate case which we term "strict linguistic bias", the output becomes completely independent of the visual input: $P_\theta(y_t \mid y_{<t}, x_v, x_t) = P_\theta(y_t \mid y_{<t}, \tilde{x}_v, x_t)$ and the model reduces to a purely language-conditional generator $P_\theta(y_t \mid y_{<t}, x_t)$, producing outputs that reflect corpus-level statistical patterns rather than visual evidence.*

3.3 PROPOSED SOLUTION

We propose a diagnostic framework with principled metrics to quantify visual grounding and identify where linguistic bias emerges in the LLaVA-Med pipeline. Guided by our diagnostic findings, we then design architectural interventions to enhance visual grounding in LLaVA-Med. We call this variant LLaVA-MedFormer (LLaVA-MF).

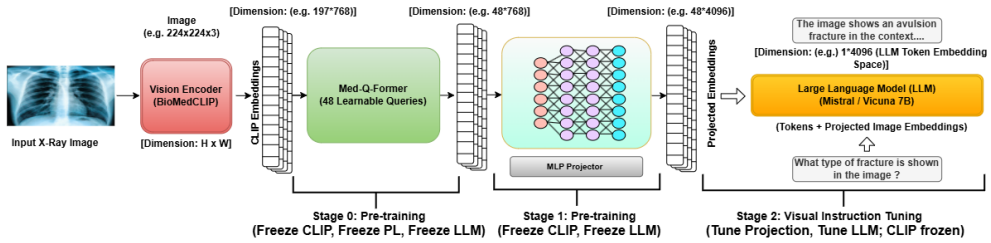


Figure 2: Overview of the proposed LLaVA-MF architecture. A frozen BiomedCLIP vision encoder extracts biomedically-grounded image embeddings, which are processed by Med-Q-Former, a Q-Former variant using 48 learnable query tokens, to produce a compact, text-relevant visual representation. A projection layer (MLP) maps these query embeddings into the LLM's token embedding space, where they are concatenated with text tokens and passed to the language model (e.g., Vicuna/Mistral-7B) for autoregressive generation.

3.3.1 DIAGNOSTIC FRAMEWORK

Our proposed diagnostic framework comprises of two complementary components: a behavioral measure: Visual Fidelity Index (VFI), which evaluates whether predictions are genuinely influenced by the image; and a structural measures like cosine similarity and Effective Representation Quality (ERQ) which trace the feature separation and qualitative information preservation across dimensions. Together, these enable detection and localization of linguistic bias in the LLaVA-Med.

Visual Fidelity Index (VFI): VFI quantifies the degree of linguistic bias by measuring the extent to which model predictions change when the input image is replaced by a null image or omitted entirely. If the model produces nearly identical predictions under these conditions, the generated response cannot reasonably be attributed to the visual input. For each generated token y_t, VFI is defined as $\text{VFI}(y_t) = \frac{L_{\text{real}}(y_t) - L_{\text{null}}(y_t)}{|L_{\text{real}}(y_t)|}$, where $L_{\text{real}}(y_t)$ and $L_{\text{null}}(y_t)$ denote the logits assigned to token y_t under the real-image and null-image conditions (complete black or white images), respectively. A VFI value close to zero corresponds to the linguistic bias condition formalized in definition 2, where model predictions become approximately invariant to the visual input, implying that generation is driven predominantly by language priors rather than visual evidence. Conversely, larger positive values indicate greater dependence on the visual input.

Although VFI establishes whether grounding failure has occurred, it does not reveal where that failure originates in the LLaVA-Med pipeline. A model may ignore the image because clinical details are lost in the visual pathway or because the language model fails to utilize them. We therefore use the cosine similarity metric to localize the failure.

Cosine Similarity of Representations: If the vision encoder fails to produce embeddings that are geometrically distinguishable (representational collapse) between real medical images and null images, then the visual grounding problem originates at the very first processing stage. Conversely, if the encoder produces discriminative embeddings but the projection layer collapses them into a concentrated region of the embedding space, the bottleneck is localised to the projection. Separating these two cases requires direct comparison of mean feature representations in the embedding spaces at each stage (encoder and projection).

Effective Representation Quality (ERQ): While cosine similarity measures the alignment between real and null representations, it does not characterize how information is distributed across different principal directions. Consequently, it provides only a partial indication of representation collapse. Visually distinct medical images should occupy distinguishable regions of the feature space; when representations collapse into a few dominant directions, distinct images may become increasingly

indistinguishable before reaching the language model. However, the distribution of information across dimensions alone does not reveal whether the retained dimensions preserve discriminative visual information. We therefore introduce ERQ, a metric that combines effective dimensionality (Roy & Vetterli, 2007) with the discriminative quality of visual representations, motivated by Fisher's discriminant criterion. ERQ thus quantifies the trade-off between the dimensionality of the visual representation and the quality of the information preserved within it.

Definition 3 (Effective Rank) *Given a feature matrix $X \in \mathbb{R}^{N\times D}$, where N denotes the number of images and D the embedding dimensionality with singular values $\sigma_1, \sigma_2, \ldots, \sigma_k$, where $k = \min(N, k)$. We define the normalized singular value distribution as $p_j = \frac{\sigma_j}{\sum_{\ell=1}^{k}\sigma_\ell}$, for $j = 1, \ldots, k$ and compute the Effective Rank (ERank) as entropy across all singular directions i.e.,* $\mathrm{ERank}(X) = \exp\left(-\sum_{j=1}^{k} p_j \log p_j\right)$ *. Thus ERank measures the visual spread into the singular dimensions.*

Definition 4 (Effective Representation Quality) *For N evaluation images with visual representations $\{x_v\}_{v=1}^{N} \in \mathbb{R}^D$, ERQ evaluates whether the visual representation preserves discriminative information despite being compressed into a lower-dimensional subspace. Specifically, ERQ jointly considers two complementary properties: (i) the effective dimensionality of the representation, quantified by ERank, and (ii) the qualitative discriminability of the retained dimensions. Let v_j denote the j-th principal singular direction. The discriminability along this direction is quantified as:*

$$q_j = \frac{|\mu_{j,\mathrm{real}} - \mu_{j,\mathrm{null}}|}{\sqrt{\frac{1}{2}\left(\sigma^2_{j,\mathrm{real}} + \sigma^2_{j,\mathrm{null}}\right) + \epsilon}}, \qquad Q = \frac{1}{k}\sum_{j=1}^{k} q_j, \qquad \rho_Q = \frac{Q}{\mathrm{ERank}}. \tag{1}$$

Here, $\mu_{j,\mathrm{real}}$ and $\mu_{j,\mathrm{null}}$ denote the mean features of the real and null-image representations, respectively, after projection onto the j-th principal direction across the N images. Similarly, $\sigma_{j,\mathrm{real}}$ and $\sigma_{j,\mathrm{null}}$ denote the corresponding standard deviations, and ϵ is a small constant for numerical stability. The quantity q_j therefore measures the standardized separation between real and null representations along an individual principal direction, while Q summarizes their average directional separation over the k retained directions. Finally, ρ_Q normalizes this discriminative separation by the effective dimensionality, thereby measuring the amount of discriminative visual information retained per effective dimension. A higher ρ_Q indicates that the representation retains stronger real-versus-null separation with fewer effective dimensions, whereas a lower value suggests that the representation either loses discriminative information or distributes it across a less informative effective subspace.

Existing methods for analyzing visual grounding in medical VLMs mainly use attention maps (Zhang et al., 2025; Kang et al., 2025), gradient-based saliency (Selvaraju et al., 2020), or causal perturbation (Fong & Vedaldi, 2017; Hooker et al., 2019). They do not directly measure how strongly predictions rely on visual evidence or where visual information is lost in the pipeline. Our framework addresses both by combining VFI, cosine similarity and ERQ.

3.3.2 Architectural Interventions

We examine the vision encoder, projection layer, and language model of LLaVA-Med and introduce targeted architectural interventions to improve visual grounding, resulting in LLaVA-MF (Figure 2).

Vision Encoder (E): dictates what visual information is retained for all downstream components. The choice of pretraining data strongly influences both visual representations and their alignment with clinical semantics. LLaVA-Med employs a CLIP-ViT-L/14 encoder pre-trained on natural-image datasets, which often biases the model towards general visual concepts rather than clinically meaningful representations. We replaced CLIP with BiomedCLIP which being domain specific preserves subtle pathological patterns and semantic alignment that natural-image encoders ignore.

Fusion: determines how visual and textual representations are integrated, influencing how much visual information is retained and the stage at which it is incorporated into the language generation process. LLaVA-Med does not incorporate a multimodal fusion mechanism; instead, it maps each visual patch token directly into the language model via a multilayer projection. In medical images, many tokens represent an uninformative background or noise. An multilayer projection cannot dynamically query, filter or compress these features. We adopt Q-Former fusion (BLIP-2) (Li et al., 2023b), which compresses visual tokens into a small set of $N_q = 48$ learnable queries

and selectively extracts only the significant and clinically relevant features before passing them to the projection layer (Figure 2). While multilayer projection is computationally cheaper to train due to its uncompressed visual tokens but Q-Former has a favorable trade-off between computational overhead and representational efficiency. The Q-Former is pre-trained with three jointly optimised losses operating over the image-text pairs (x_v, x_t): $\mathcal{L}_{\mathrm{ITC}}$ (Image-Text Contrastive Loss), $\mathcal{L}_{\mathrm{ITM}}$ (Image-Text Matching) and $\mathcal{L}_{\mathrm{ITG}}$ (Image-Grounded Text Generation).The detailed explanation of these losses is presented in the supplementary material.

Projection Layer (P): aligns visual features into the language model's embedding space and can become an information bottleneck despite rich incoming representations. While computationally efficient, its limited feature capacity may discard fine-grained visual details. Exploring alternative interventions to reduce information loss at this bottleneck is an interesting future direction.

Language Model: is the final stage LLaVA-Med pipeline. Despite receiving rich visual representations, it may still rely on linguistic priors due to extensive training of a large parameter set. LLaVA-Med fully fine-tuned the P and language model for its BiomedCLIP variant. We therefore investigate LoRA-based adaptation to mitigate this overreliance and improve visual grounding.

4 Experiments

4.1 Datasets and Protocol

Fine-tuning Datasets. Similar to LLaVA-Med, we use the 60K biomedical instruction-tuning data derived from image-caption data from PubMed Central (Li et al., 2024). It contains 60,000 biomedical image-caption pairs spanning diverse clinical modalities, including chest radiography, computed tomography (CT), magnetic resonance imaging (MRI), histopathology, and gross clinical photography, paired with descriptive diagnostic captions and structured visual instruction dialogues.

Evaluation Datasets. We evaluate LLaVA-MF against the LLaVA-Med on three unseen medical VQA benchmarks: VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020). VQA-RAD contains 315 radiology images and 3,515 question answer pairs. SLAKE contains 642 images and 14,028 question answer pairs covering radiology, pathology, and clinical photography. PathVQA contains 4,998 pathology images and 32,799 question answer pairs. Each benchmark includes both open-ended (OE) questions that require free-form text responses and closed-ended (CE) questions that require binary answers, such as yes or no.

Diagnostic Set. We use a diagnostic set of 193 evaluation prompts associated with 50 unique images from the LLaVA-Med evaluation set. The prompts cover five imaging modalities: chest X-ray, MRI, histology, gross pathology, and CT, and include 143 conversation prompts and 50 description prompts. This set is used only for diagnostic evaluation using VFI, ERQ, and cosine similarity.

Null conditions. To evaluate linguistic bias, we define null conditions by replacing the original visual input with null baselines (solid black (B) and solid white images (W)), effectively stripping all visual and anatomical semantic content. We also add Gaussian noise to the diagnostic set with $\sigma \in \{0.1, 0.5, 1.0\}$ to probe LLaVA-Med and LLaVA-MF's robustness against visual degradation. Following prior perturbation-based attribution studies (Fong & Vedaldi, 2017; Hooker et al., 2019), we define a visual fidelity threshold $\tau = 0.05$. A VFI score below this value indicates that the generated text has negligible dependence on the visual input.

Evaluation Metrics. For downstream VQA, we use token-level exact-match (EM) accuracy for both OE and CE questions, following the LLaVA-Med evaluation protocol. We first convert both predicted and ground truth answers to lowercase and remove whitespaces. Following LLaVA-Med, we use GPT-4 as an automatic judge to assess clinical accuracy, visual grounding, specificity, and fluency of OE responses. The GPT-4 evaluation score for model P_θ is normalized relative to the score assigned to GPT-4's own reference responses: $\mathrm{Score}(P_\theta) = \frac{\sum_{i=1}^{N} s_i^{(P_\theta)}}{\sum_{i=1}^{N} s_i^{(\mathrm{GPT4})}} \times 100$ where N is the number of evaluation prompts, $s_i^{(P_\theta)} \in [1, 10]$ is the score assigned by GPT-4 to the response generated by model P_θ for prompt i, and $s_i^{(\mathrm{GPT4})}$ is the score assigned by GPT-4 to its own reference response for the same prompt. Thus, the GPT-4 reference provides a normalized score of 100.

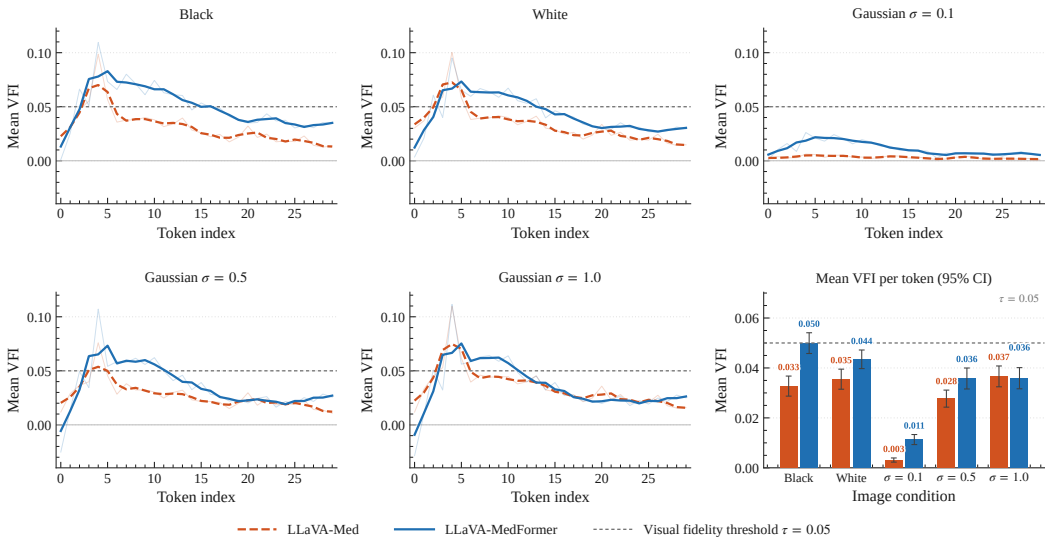


Figure 3: VFI trajectories for LLaVA-Med and LLaVA-MF under B and W null conditions, as well as for Gaussian-noise-perturbed images with noise levels ($\sigma = 0.1, 0.5, 1.0$) on LLaVA-Med evaluation set. The degradation in VFI is more gradual for LLaVA-MF than for LLaVA-Med.

4.2 Implementation Details

Our LLaVA-MF replaces the general-domain encoder with frozen BiomedCLIP ViT-B/16 (Zhang et al., 2023a). BiomedCLIP extracts 197 tokens of dimension 768 (comprising 196 spatial grid patch tokens and 1 [CLS] token), preserving spatial information for cross-attention. A 6-layer Q-Former is used after the encoder which is initialised with BioClinicalBERT (Alsentzer et al., 2019) encoder weights. Q-Former compresses these 197 visual tokens into 48 visual query tokens using 32 general and 16 clinical learnable queries. The resulting 768-dimensional query representations are mapped into the 4096-dimensional Vicuna-7B token embedding space via a two-layer GELU MLP projector.

Training proceeds in three stages. Stage 0 pre-trains the Q-Former with ITC, ITM and ITG, following BLIP-2. Stage 1 (S1) trains the projection with the language model frozen. In stage 2 (S2) we adapted Vicuna-7B using LoRA (Hu et al., 2022) with (rank $r = 16$ and scaling factor $\alpha = 32$; $\sim$58M of 7B parameters are trainable). We evaluate on 60K instruction-tuning variants and vary the LoRA fine-tuning depth as $FT \in \{0, 1, 3\}$ epochs. Here, $FT = 0$ denotes no LoRA fine-tuning, while $FT = 1$ and $FT = 3$ denote one and three LoRA epochs, respectively. The vision encoder is frozen throughout the training process and contributes no trainable parameters. All experiments are performed in FP16 on a single NVIDIA H100 GPU. Code is available at [1].

4.3 Measuring Linguistic Bias: VFI

To analyze how strongly the generated response depends on the visual input, we use the VFI to measure token-level visual dependence during generation. Higher VFI indicates stronger dependence on the image and, consequently, less reliance on language priors. So, we evaluate model responses against both uninformative visual baselines and noise-perturbed inputs. Figure 3 compares LLaVA-Med and LLaVA-MF over the 193 prompts paired with 50 images sampled from PMC-OA under five such conditions: null baselines (black and white images) and images with Gaussian noise with standard deviations $\sigma \in \{0.1, 0.5, 1.0\}$. Both models show a two-phase pattern in which visual grounding peaks during early token gener-

Representation	R-B	R-W	B-W
LLaVA-Med (E)	0.938	0.937	0.994
LLaVA-MF (E)	**0.313**	**0.317**	**0.829**
LLaVA-Med (P)	0.990	0.990	1.000
LLaVA-MF(P)	**0.889**	**0.900**	**0.975**

Table 1: Cosine similarity between mean embeddings of real images (R) and null inputs. Lower is better.

[1] https://huggingface.co/annonymousHF/LLaVA-MedFormer

ation and gradually attenuates as autoregressive decoding advances. This is natural, as the later parts/sentences depend higly on the preceding parts/sentences. Across four of the five control conditions, LLaVA-MF consistently maintains higher VFI compared to LLaVA-Med and decays more gradually than LLaVA-Med over a 30-token horizon. These findings demonstrate that LLaVA-MedFormer generated tokens remain more dependent on visual input for a longer portion of the sequence and thus do not overly rely on language priors.

4.4 Cosine Similarity Analysis of Visual Representations

To assess the possibility of representation collapse in LLaVA-Med, we evaluate the quality of the encoder (E) and projection layer (P) embeddings, as the visual input may not contain sufficient signal for the language model to ground on, leading to overreliance on the language prior. Table 1 shows the cosine similarity between the mean embeddings of real (R), black (B), and white (W) images for both models at the encoder (E) and projection (P) stages. In the LLaVA-Med E space, real images are highly similar to black and white images (0.938 and 0.937), and the null inputs are nearly identical (0.994). After LLaVA-Med's projection, all three pairwise similarities are at or above 0.99, indicating that the language model receives visual tokens that are collapsed. This indicates that the encoder and projection layer of the pipeline may be responsible for linguistic bias. In contrast, the LLaVA-MF encoder separates R from B and W images at 0.313 and 0.317, and also distinguishes black from white images at 0.829. However, the projection layer embedding similarity of LLaVA-MF for R and B and W images 0.889 and 0.900 and for Black-White image 0.975 which is less than LLaVA-Med, indicating better quality of visual input entering the language model. However, these results also reflect that the grounding failure present at the E is further amplified by the P in both methods, although less for LLaVA-MF. Designing a projection layer that reduces the representation collapse in the output embedding is an interesting future direction.

Real vs.	E $\left(\frac{\rho_Q^{MF}}{\rho_Q^{Med}}\right)$	P $\left(\frac{\rho_Q^{MF}}{\rho_Q^{Med}}\right)$
Black	**1.24**	**15.53**
White	**1.34**	**15.89**
Gauss $\sigma = 0.1$	**1.73**	**13.73**
Gauss $\sigma = 0.5$	**2.45**	**62.20**
Gauss $\sigma = 1.0$	**2.63**	**105.92**

Table 2: Ratio of ERQ for LLaVA-MF to LLaVA-Med across five null conditions at the E and P. Values > 1 indicate greater discriminative quality preserved per effective dimension.

4.5 Evaluating Representation Quality

To quantify the qualitative effective dimensions retained across the pipeline stages E and P in LLaVA-Med and LLaVA-MF, we compute the ratio $\left(\frac{\rho_Q^{MF}}{\rho_Q^{Med}}\right)$ separately for each stage. Table 2 shows that LLaVA-MF maintains higher discriminative quality per effective dimension. Specifically, encoder ratio ranges from $1.24\times$ to $2.63\times$, while the projection ratio ranges from $13.73\times$ to $105.92\times$, with the gap increasing under stronger Gaussian (Gauss) perturbations. These results indicate that LLaVA-MF retains more discriminative visual information per effective dimension.

4.6 Benchmark Results

We also evaluate LLaVA-Med and LLaVA-MF across three out-of-distribution medical VQA benchmarks VQA-RAD, SLAKE and PathVQA, under both open-ended and closed-ended settings. As summarized in Table 3, LLaVA-MF achieves substantial average accuracy gains on PathVQA ($+19.98\%$) and SLAKE ($+8.72\%$), with consistent improvements across both question modalities. However, in VQA-RAD, a 10% decrease in open-ended questions outweighs a 2.20% gain in closed-ended questions, leading to a net decrease of 3.90%. Nevertheless, LLaVA-MF increases the cumulative average accuracy

Benchmark	LLaVA-Med	LLaVA-MF	Δ
VQA, OE	**54.12**	44.12	-10.00
VQA, CE	64.71	**66.91**	$+2.20$
SLAKE, OE	79.33	**89.78**	$+10.45$
SLAKE, CE	64.90	**71.88**	$+6.98$
PathVQA, OE	17.18	**37.97**	$+20.79$
PathVQA, CE	71.37	**90.54**	$+19.17$
Average	58.60	**66.86**	$+8.26$

Table 3: Exact-match accuracy (%) of LLaVA-Med and LLaVA-MF for configuration (60K, S2=3, FT=3). Results of LLaVA-Med were taken from their paper.

from 58.60% to 66.86% (a net improvement of $+8.26\%$). This, together with Table 1 and 2 , and Figure 3 indicates better grounding of visual information and, therefore, a performance boost.

4.7 Ablation Study

Encoder and Q-Former contributions. To determine whether the two architectural interventions reduce linguistic bias, we disentangle the individual contributions of the visual encoder (BiomedCLIP) and the Q-Former. Following (Li et al., 2023a), we evaluate open-ended responses using GPT-4 with a 1-10 rubric covering clinical accuracy, visual grounding, specificity, and fluency, normalized against the judge's reference answers (Table 4). Replacing CLIP with the domain-matched BiomedCLIP encoder and increases the score from 78.8 to 96.6 (+17.8). Adding the Q-Former provides a further gain of 15.0 points, reaching 111.6. Thus, domain-matched visual encoding provides the greater gain, while Q-Former provides an additional improvement through alignment-guided compression.

	Conv	Desc	CXR	MRI	Hist	Gross	CT	All
LLaVA-Med	82.7	67.5	106.3	65.3	67.5	69.6	86.5	78.8
LLaVA-MF								
w/o Q-Former	103.4	77.0	120.9	79.1	81.0	99.8	105.2	96.6
LLaVA-MF	**124.1**	75.8	**130.8**	**104.8**	**101.2**	**115.2**	108.5	**111.6**

Table 4: Ablations on BiomedCLIP and Q-former. GPT-4-judged response quality over 193 prompts on LLaVA-Med evaluation set.

Effect of Finetuning Steps. Finetuning Steps have a large effect on performance. On both SLAKE and PathVQA under OE and CE questions, accuracy increases from 39.29 at FT = 0 to 89.78 at FT = 3, a 50.49-point gain with all other settings fixed for SLAKE and similarly for PathVQA a 31.23-point gain (Table 5). This indicates that, without stage 2 fine-tuning, the language model does not adapt to the visual prefix and therefore relies largely on its language priors. The upstream changes to the visual encoder alone do not transfer effectively to the language model. The low-rank LoRA update provides the required downstream adaptation while preserving the model's generalizability.

	SLAKE		**PathVQA**	
	Open	**Closed**	**Open**	**Closed**
FT=0	39.29	42.28	7.79	59.31
FT=1	89.54	70.43	37.47	90.54
FT=3	**89.78**	**71.88**	**37.97**	**90.54**

Table 5: Ablations on fine-tuning depth for LLaVA-MF.

5 Discussion

The three proposed diagnostics together provide a consistent explanation of the observed linguistic bias. The out-of-domain CLIP encoder maps real and null images to nearly the same region of feature space. The projection layer then removes 6% of the remaining variation (section 4.4). As a result, the language model receives a visual prefix that is nearly constant and quickly falls back on its learned language priors, within five generated tokens (section 4.3). Thus, the main issue is not that the language model ignores a useful visual signal; rather, much of that signal has already been lost before it reaches the language model. This also differs from conventional hallucination (Rohrbach et al., 2018; Gu et al., 2024). The generated answers are fluent, clinically plausible, and consistent with the training distribution, but they are largely independent of the input image. Standard benchmark accuracy may not reveal this behavior, motivating the use of grounding diagnostics.

6 Conclusion

In summary, LLaVA-Med uses visual information only briefly during generation. It first uses the image to establish a domain-level anchor, but its sensitivity to the visual input drops below the threshold within five generated tokens, after which generation is increasingly driven by language priors. Our analysis identifies two structural factors behind this behavior: the general-domain encoder does not adequately separate medical images from null inputs, and the projection layer reduces the effective visual dimensionality. Replacing the encoder with a in-domain BiomedCLIP and using an alignment-pretrained Q-Former improving visual grounding during generation to some extent; however, the grounding remains transient despite these architectural modifications. This suggests that the problem is not only due to the visual encoder and projection bottleneck, but is also partly inherent to autoregressive multimodal generation.

AI USE STATEMENT

In this work, we used generative AI tools for paraphrasing and editing text, exploration and review of the literature, and generating code for data analysis and plots. We have not used generative AI tools for conducting experiments, generating experimental results, or making scientific claims. Additionally, generative AI tools were used to assist with code development and presentation of the results. All AI-assisted text and code were reviewed, verified, and, where applicable, tested by the authors. The authors independently evaluated the relevant literature, experimental methodology, results, and scientific claims. We take full responsibility for the final content of this work, including all text, claims, analyses, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Not Applicable.

REPRODUCIBILITY STATEMENT

The PubMed Central dataset is available at (`https://pmc.ncbi.nlm.nih.gov/tools/openftlist/`) The fine-tuned LLaVA-Med model and BiomedCLIP model weights can be downloaded from(`https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b`) and `https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224`, respectively. The VQA-RAD, PathVQA, and SLAKE datasets are available at (`https://osf.io/89kps/overview`) (`https://huggingface.co/datasets/flaviagiammarino/path-vqa`) (`https://huggingface.co/datasets/Voxel51/SLAKE`). The initial Q-Former weights can be obtained from the official BLIP-2 repository (`https://huggingface.co/Salesforce/blip2-opt-2.7b`). Detailed implementation settings and hyperparameters required to reproduce our experiments are provided in Section 4.2. The source code of proposed LLaVA-MF and the diagnostic set is available at (`https://huggingface.co/annonymousHF/LLaVA-MedFormer`). The weights of finetuned LLaVA-MF are available at (`https://huggingface.co/annonymousHF/LLaVA-MedFormer`).

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. *NeurIPS*, 2022. doi: 10.48550/arXiv.2204.14198.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019. doi: 10.18653/v1/W19-1909.

Jinjin Cao, Zhiyang Chen, Zijun Wang, Liyuan Ma, Weijian Luo, and Guojun Qi. When images speak louder: Mitigating language bias-induced hallucinations in vlms through cross-modal guidance, 2025. URL `https://arxiv.org/abs/2510.10466`.

Richard J. Chen et al. A generalist foundation model for pathology. *Nature Medicine*, 2024a.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024b.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, June 2025.

Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3429–3437, 2017. doi: 10.1109/ICCV.2017.371. URL https://doi.org/10.1109/ICCV.2017.371.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. *arXiv preprint arXiv:2405.15683*, 2024. doi: https://doi.org/10.48550/arXiv.2405.15683. Accepted to ICLR 2025.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1262.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

Zishan Gu, Changchang Yin, Fenglin Liu, and Ping Zhang. MedVH: Towards systematic evaluation of hallucination for large vision language models in the medical context. *arXiv preprint arXiv:2407.02730*, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. doi: 10.48550/arXiv.2106.09685.

Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision-language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.02103.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. *arXiv preprint arXiv:2503.06287*, 2025. doi: 10.48550/arXiv.2503.06287. URL https://arxiv.org/abs/2503.06287.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. doi: https://doi.org/10.48550/arXiv.2311.16922.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS Datasets and Benchmarks Track*, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS Datasets and Benchmarks*, 2024. doi: 10.52202/075280-1240.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

B. Liu, L. Zhan, and et al. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *ICCV Workshops*, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a. doi: 10.48550/arXiv.2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, Sina Bakhshandeh Taleshani, Erfan Hashemi, Javad Mirzapour Kaleybar, and Omid Nejati Manzari. Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 2026. doi: https://doi.org/10.1016/j.compbiomed.2025.111347. arXiv:2512.01922.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pp. 4035–4045, 2018. doi: 10.18653/v1/D18-1437.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, Poznan, Poland, 2007.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12481–12500, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.631. URL https://aclanthology.org/2025.emnlp-main.631/.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007/s11263-019-01228-7.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2502.17422.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a. doi: https://doi.org/10.48550/arXiv.2303.00915.

Sheng Zhang et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023b.

A Appendix

A.1 Architectural Interventions

A.1.1 Fusion:

Here, we provide additional details on the Q-Former-based multimodal fusion mechanism incorporated into LLaVA-MF. The Q-Former is pre-trained with three jointly optimised losses operating

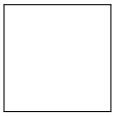
Input	Output
 Black image *Q: What type of tissue is shown in the image?*	The image shows a histopathology of a tumor, which is a microscopic examination of the tissue. Histopathology is a crucial diagnostic tool that helps in understanding the nature of the tumor, its cellular composition, and the extent of the disease. It can provide valuable information for determining the appropriate treatment and prognosis for the patient.
 White image *Q: What type of tissue is shown in the image?*	The image shows a histopathological examination of a tissue sample. Histopathology is the study of diseased tissue under a microscope to examine the manifestations of disease. It is an important diagnostic tool that helps healthcare professionals identify the type and severity of a disease, as well as determine the most appropriate treatment options.
N/A No image *Q: What type of tissue is shown in the image?*	The image shows a histological section of a tissue sample. Histological sections are thin slices of tissue that are prepared for microscopic examination. These sections are typically stained with specific dyes to highlight different structures and cell types within the tissue. The purpose of examining histological sections is to study the microscopic structure of the tissue and to identify any abnormalities or pathological . . .

Table 6: LLaVA-Med responses for inputs without visual content. Despite receiving an all B image, an all W image, or no image, the model describes histopathological tissue, indicating reliance on language priors rather than visual evidence.

Condition	LLaVA-Med		LLaV-MF		LLaVA-MF/LLaVA-Med	
Real vs.	E ρ_Q	P ρ_Q	E ρ_Q	P ρ_Q	E ρ_Q^{MF}/ρ_Q^{Med}	P ρ_Q^{MF}/ρ_Q^{Med}
Black	6.98	7.84	**8.64**	**122**	**1.24**	**15.53**
White	7.54	9.22	**10.10**	**147**	**1.34**	**15.89**
Gauss σ=0.1	0.978	1.44	**1.69**	**19.8**	**1.73**	**13.73**
Gauss σ=0.5	1.63	1.83	**4.00**	**114**	**2.45**	**62.20**
Gauss σ=1.0	1.92	2.15	**5.06**	**228**	**2.63**	**105.92**

Table 7: Stage-wise ERQ ρ_Q at the encoder and projection stages across five real-versus-null conditions. Values in the first four columns are reported in $\times 10^{-3}$; the final two columns show the LLaVA-MF to LLaVA-Med ERQ ratio.

over the image-text pairs (x_v, x_t): $\mathcal{L}_{\mathrm{ITC}}$ (Image-Text Contrastive Loss), $\mathcal{L}_{\mathrm{ITM}}$ (Image-Text Matching) and $\mathcal{L}_{\mathrm{ITG}}$ (Image-Grounded Text Generation).

1) Image-Text Contrastive (ITC) Loss. The image-text contrastive objective $\mathcal{L}_{\mathrm{ITC}}$ trains the model to assign higher similarity to matched image-text pairs than to mismatched pairs. Let B denote the mini-batch size, and $i, j \in \{1, \ldots, B\}$ index image-text pairs within the mini-batch, and $k \in \{1, \ldots, N_q\}$ index the query tokens. The Q-Former produces N_q visual query representations $Z = \{z_1, \ldots, z_{N_q}\} \in \mathbb{R}^{N_q \times d}$, while the text encoder produces a single global text representation $u \in \mathbb{R}^d$ for each image-text pair (x_{v_i}, x_{t_i}) in a mini-batch. Since the image is represented by multiple query tokens whereas the text is represented by a single vector, we define the image-text similarity as the maximum cosine similarity between the text representation and the visual queries: $sim(Z_i, u_j) =$

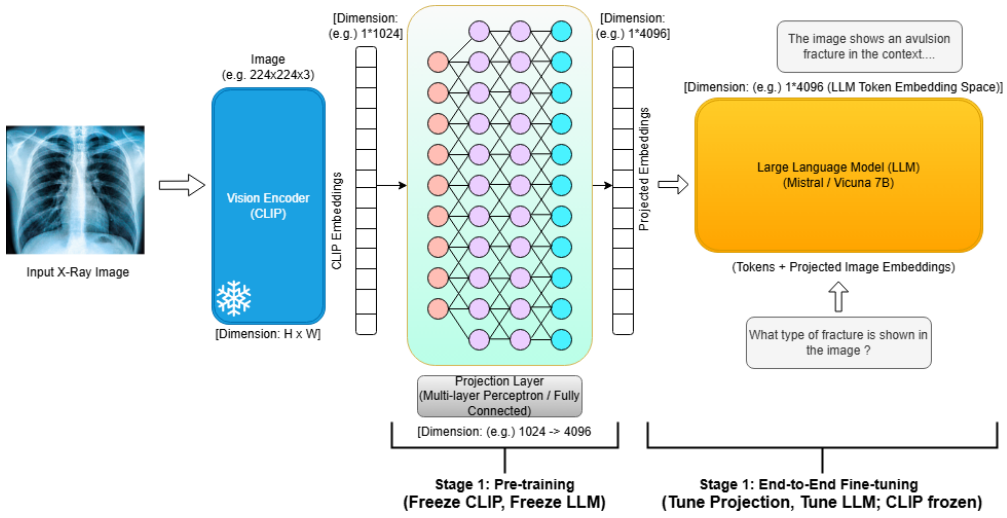


Figure 4: Overview of the LLaVA-Med pipeline, adapted from the architecture proposed by (Li et al., 2024). A frozen CLIP vision encoder extracts image embeddings from the input, which are mapped into the LLM's token embedding space via a trainable projection layer (MLP). The projected embeddings are concatenated with text tokens and passed to the language model (e.g., Vicuna/Mistral-7B) for autoregressive generation. Training proceeds in two stages: Stage 1 pre-trains the projection layer with both the vision encoder and LLM frozen, and Stage 2 performs end-to-end fine-tuning of the projection layer and LLM while keeping the vision encoder frozen.

$\max_{k=1,\ldots,N_q} z_{i,k}^{\top} u_j$ The bidirectional contrastive loss is then

$$\mathcal{L}_{\mathrm{ITC}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp\big(sim(Z_i,u_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(sim(Z_i,u_j)/\tau\big)} + \log\frac{\exp\big(sim(Z_i,u_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(sim(Z_j,u_i)/\tau\big)}\right] \tag{2}$$

where τ is a learnable temperature parameter. A unimodal self-attention mask is employed so that queries and text tokens cannot attend to each other, preventing information leakage.

2) Image-Text Matching (ITM) Loss. The image-text matching objective $\mathcal{L}_{\mathrm{ITM}}$ optimizes a binary classification task to predict whether an image-text pair is matched ($m_i = 1$) or unmatched ($m_i = 0$). To achieve this, model uses a bidirectional self-attention mask, allowing all queries and text tokens to attend to each other. Each output query vector $z_{i,k}$ is projected via a linear classification head parameterized by weight vector w and bias b. The resulting query logits averaged to produce the pair matching score:

$$\hat{m}_i = \phi\left(\frac{1}{N_q}\sum_{k=1}^{N_q} w^{\top} z_{i,k} + b\right), \tag{3}$$

where $\phi(\cdot)$ denotes the sigmoid function and $w^{\top}$ is a learned weight vector that maps d-dimensional query $z_{i,k} \in \mathbb{R}^d$ into a scalar value. Positive pairs (matched) are the original image-text pairs, while negative pairs (unmatched) are formed by pairing an image with text from a different image within the mini-batch. The ITM loss is the binary cross-entropy over the matched and unmatched pairs: $\mathcal{L}_{\mathrm{ITM}} = -\frac{1}{B}\sum_{i=1}^{B}\Big[m_i \log \hat{m}_i + (1-m_i)\log(1-\hat{m}_i)\Big]$

3) Image-Grounded Text Generation (ITG) Loss. The image-grounded text generation objective $\mathcal{L}_{\mathrm{ITG}}$ encourages the generated text to be conditioned on the visual input through Q-Former visual query representations.

$$\mathcal{L}_{\mathrm{ITG}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{t=1}^{T_i}\log P_{\theta_{\mathrm{QF}}}\left(y_{i,t} \mid y_{i,<t},\ Z_i\right). \tag{4}$$

Here, T_i is the target sequence length for sample i, Z_i denotes the visual query representations extracted from input image x_{v_i}, and $P_{\theta_{\mathrm{QF}}}(\cdot)$ denotes the conditional probability distribution parameterized by the trainable Q-Former parameters θ_{QF}.

Finetuning Procedure:

We finetune proposed LLaVA-MF in three sequential stages: Q-Former pre-training, projection layer training, and LoRA-based fine-tuning of the language model. The vision encoder is a BiomedCLIP, which is pretrained on biomedical image-text data from PubMed Central.

- **Stage 0: Q-Former Pre-training.** We first train the Q-Former, which extracts visual information through a fixed set of learnable queries. The vision encoder remains frozen while the Q-Former is optimized using the $\mathcal{L}_{\mathrm{ITC}}$, $\mathcal{L}_{\mathrm{ITM}}$, and $\mathcal{L}_{\mathrm{ITG}}$ objectives. These losses jointly encourage the queries to capture visual information that is aligned with the corresponding medical text.
- **Stage 1: Projection Layer Training.** We then train the projection MLP to map the Q-Former outputs into the input embedding space of the Vicuna-7B, the language model used in our framework. Vicuna remains frozen during this stage, and the projection layer is optimized using the language-modeling objective $P_\theta(y_t \mid y_{<t},\, x_v,\, x_t)$.
- **Stage 2: LoRA Fine-Tuning.** Finally, we fine-tune Vicuna using Low-Rank Adaptation (LoRA), a parameter-efficient method that learns low-rank updates instead of updating the full language model. The Q-Former and vision encoder remain frozen, while the LoRA parameters are optimized on the 60K medical instruction-tuning pairs from LLaVA-Med. The same 60K data is used for the corresponding LLaVA-Med baseline for fair comparison.