Enriched Instruction-Following Graph Alignment for Efficient Medical Vision-Language Models

Duy M. H. Nguyen [1 2 3] **Nghiem T. Diep** [* 3] **Trung Q. Nguyen** [* 3 4] **Hoang-Bao Le** [3] **Tai Nguyen** [3] **Tien Nguyen** [5 6] **TrungTin Nguyen** [7] **Nhat Ho** [8] **Pengtao Xie** [9] **Roger Wattenhofer** [10] **Daniel Sonntag** [3 11] **James Zou** [12] **Mathias Niepert** [1 2]

Abstract

State-of-the-art medical multi-modal LLMs (med-MLLMs), such as LLaVA-Med and BioMedGPT, primarily depend on scaling model size and data volume, with training driven largely by autoregressive objectives. However, we reveal that this approach can lead to weak vision-language alignment, making these models overly dependent on costly instruction-following data. To address this, we introduce ExGra-Med, a novel multi-graph alignment framework that jointly aligns images, instruction responses, and extended captions in the latent space, advancing semantic grounding and cross-modal coherence. To scale to large LLMs (e.g., LLaMa-7B), we develop an efficient end-to-end training scheme using black-box gradient estimation, enabling fast and scalable optimization. Empirically, ExGra-Med matches LLaVA-Med's performance using just 10% of pre-training data, achieving a 20.13% gain on VQA-RAD and approaching full-data performance. In addition, it exceeds other SOTA med-MLLMs in Med-VQA benchmarks, promising a new way to integrate vision and language in medical AI. We release our checkpoints at this Github.

1. Introduction

Multi-Modal Large Language Models (MLLMs) are a range of neural network architectures that can process different types of input data, such as images, text, and audio. One important step of training MLLM is the curation of instruction-following (IF) data (Lou et al., 2023), in which a model is asked to give answers to questions in a multi-turn conversation based on a given image. This step is also implemented in the medical domain, where IF datasets encompassing medical images, clinical notes, and diagnostic data are curated (Xie et al., 2024). One example is the work of LLAVA-Med (Li et al., 2024), which leverages GPT-4 to curate 600K biomedical image-text pairs and 60K multi-modal IF data from PMC-15M (Zhang et al., 2023c). Other works follow this practice, which either scale up the amount of medical data (Xie et al., 2024; Zhang et al., 2023a; He et al., 2024) or the model size (Wu et al., 2023; Jiang et al., 2024), while the autoregressive objective loss stays the same. This practice is indeed helpful, enabling models to achieve promising performance. For example, Biomed-GPT (Zhang et al., 2023a) is excellent with multiple biomedical modalities, while Med-Flamingo (Moor et al., 2023) also reports good performance on few-shot learning for medical visual question answering. However, the reliance of performance on scale poses some serious questions, especially in the biomedical domain, where IF data is scarce.

In our work, we study the effectiveness of this approach and reveal that *autoregressive learning is highly data-hungry*, which leads to the degradation of the performance of the model if there is insufficient IF data. To demonstrate this, we pre-train LLAVA-Med using only 10% of the original data and compare it to a version trained on the full dataset. Both models are then fine-tuned on the VQA-RAD dataset (Lau et al., 2018). As illustrated in Figure 1, despite being updated with downstream task data, performance declines sharply from 72.64% to 52.39% on VQA-RAD. This underscores the instability of medical MLLMs trained with autoregressive methods and highlights their heavy reliance on extensive medical instruction-following data to achieve satisfactory performance.

To overcome the limitations of autoregressive training in settings with limited instruction-following data, we introduce ExGra-Med, a novel multi-graph alignment framework designed to enhance cross-modal understanding in multi-modal large language models (MLLMs). Central to our method is the construction of three modality-specific graphs: one capturing visual features from a vision encoder, and two representing different textual forms of the instruction. These graphs encode semantic relationships within and across modalities, and we cast their alignment as a combinatorial multi-graph matching problem. This enables the model to learn consistent triplet-level associations among the image, the original instruction, and a semantically en-

[*]Equal contribution [1]Max Planck Research School for Intelligent Sytems (IMPRS-IS) [2]University of Stuttgart [3]German Research Centre for Artificial Intelligence (DFKI) [4]Technical University of Munich [5]University Medical Center Gottingen [6]Max Planck Institute for Multidisciplinary Sciences [7]University of Queensland [8]University of Texas at Austin [9]University of California San Diego [10]ETH Zurich [11]Oldenburg University [12]Stanford University. Correspondence to: Duy M. H. Nguyen <ho_minh_duy.nguyen@dfki.de>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

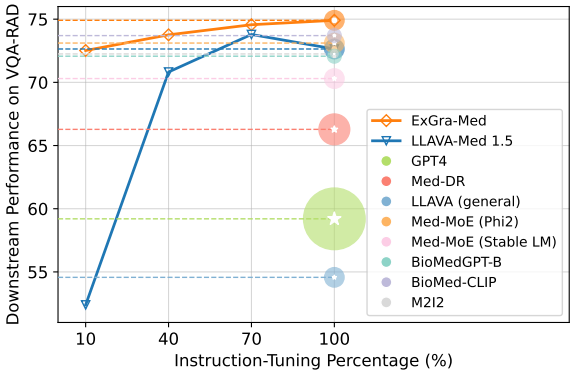


Figure 1. Our ExGra-Med versus LLaVA-Med across varying instruction-following (IF) pre-training data sizes, **highlighting the data-hungry behavior of auto-regressive modeling**. Both *models are fine-tuned on the same VQA-RAD training set* after the pre-training stage at each IF rate. At 100% IF pre-training, ExGra-Med and LLaVA-Med are benchmarked against other state-of-the-art models, *all fine-tuned on the same VQA-RAD training set* (except GPT-4, which is evaluated without fine-tuning). Circle radius represents the number of model parameters.

riched variant. We jointly optimize this alignment objective alongside the standard autoregressive language modeling loss, thereby improving the model's semantic comprehension, coherence, and instruction-following capabilities. To generate the enriched instruction, we employ a frozen LLM (GPT-4 (Achiam et al., 2023)) to produce a contextually extended version that preserves the original intent while emphasizing key concept relationships. The vision encoder and language model (LLaMa (Touvron et al., 2023)) independently process the image, instruction, and extension to produce node embeddings used in the alignment. Unlike simple data augmentation, our GPT-4-driven supervision introduces meaningful semantic structure, enabling fine-grained graph-based correspondence learning across modalities (Figure 2).

Our method differentiates itself from existing multi-modal alignment techniques for LLM (Park et al., 2024; Li et al., 2023a; Chen et al., 2023) in different perspectives. First, while prior contrastive objectives primarily focus on *learning projections* based on multi-layer perceptron (MLP) layer (Liu et al., 2024a; Chen et al., 2023) or adapters (Zhang et al., 2024; Huang et al., 2023; Alayrac et al., 2022; Li et al., 2023a) to connect frozen vision encoders with frozen language models, *our algorithm directly trains LLM* using the multi-graph framework. Second, we unify and generalize pre-training algorithms commonly applied for vision-language models using pairwise contrastive learning between image-text pairs (Liu et al., 2023; Zhai et al., 2023; Khan & Fu, 2023), optimal transport (Chen et al., 2022; Nguyen et al., 2024a), or impose clustering constraints (Park et al., 2024) by integrating combinatorial formulation across cross-domain graphs. This allows us to integrate both feature and structural consistencies using graph edges, enhancing robustness for similar entities (whether images or descriptions) commonly found in medical datasets.

Finally, combinatorial graph alignment is inherently non-differentiable (Rolínek et al., 2020), and solving multi-graph alignment is computationally expensive (Pevzner, 1992). While existing approaches - such as multi-marginal optimal transport (Lin et al., 2022; Piran et al., 2024), Wasserstein barycenters (Nguyen et al., 2024b), and multi-adjacency matrix assumptions (Bernard et al., 2019; Swoboda et al., 2019) — help relax the problem, they are limited to small-scale tasks and require multiple solver steps, making them inefficient for LLM training. We overcome these challenges by leveraging modern implicit maximum likelihood estimation techniques (Niepert et al., 2021; Minervini et al., 2023). This enables efficient gradient estimation and allows for fast forward and backward propagation through large LLMs (e.g., LLaMa-7B), using a *barycenter graph* (Agueh & Carlier, 2011) for alignment. As a result, the model can scale effectively with extensive datasets on large LLMs while maintaining alignment performance.

In summary, we make the following key contributions:

- We reveal the data-demanding nature of autoregressive modeling in pre-training medical-MLLM (LLaVa-Med), showing that insufficient instruction-following data leads to significant performance drops on downstream tasks, even after fine-tuning.
- We introduce a new multi-graph alignment objective, namely ExGra-Med, that establishes triplet correlations among images, their instruction-following context, and their enriched versions. Furthermore, we developed an efficient solver for training with LLMs (LLaMa-7B) that can scale with the size of the dataset and model.
- We empirically demonstrate that using a small amount of pre-training data, ExGra-Med can achieve performance comparable to LLaVa-Med trained on 100% data. Additionally, when trained on larger datasets, ExGra-Med *outperforms* several state-of-the-art *med-MLLMs* and *advanced multi-modal pre-training* algorithms across three Medical VQA tasks.

2. Multi-graph Alignment Learning

We denote the vision encoder, projector and LLM by $f_\theta(.)$, $h_\phi(.)$, $g_\sigma(.)$, respectively. Figure 2 presents our ExGra-Med algorithm, which learns model through triplet alignment in instruction tuning data. Before detailing each component, we provide some notations used in this paper.

Notation. Given any tensor $\mathbf{T} = (T_{i,j,k,l})$ and matrix $\boldsymbol{M} = (M_{k,l})$, we use $\mathbf{T} \otimes \boldsymbol{M}$ to denote the tensor-matrix multiplication, *i.e.,* the matrix $(\sum_{k,l} T_{i,j,k,l} M_{k,l})_{i,j}$. Given $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N] \in \mathbb{R}^{N \times d}$, we define $\mathbb{E}(\boldsymbol{Y}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i \in \mathbb{R}^d$. Moreover, we define the matrix scalar

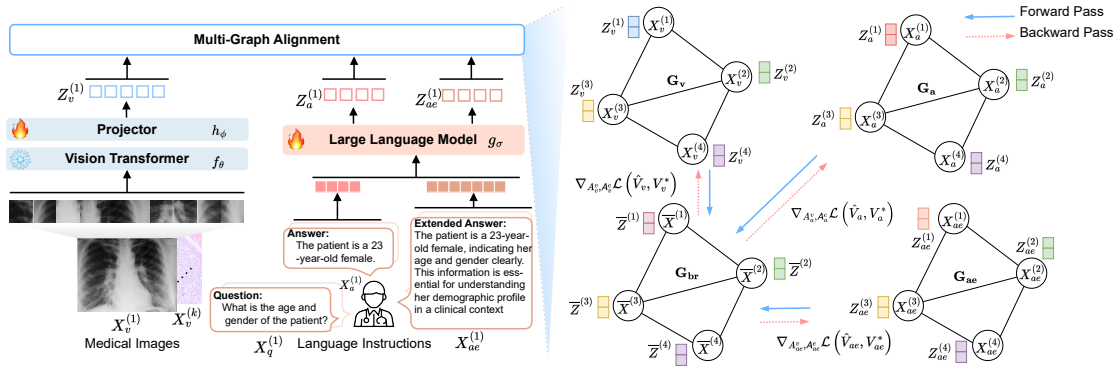


Figure 2. Overview of EXGRA-MED: The large language model g_σ and the projector h_ϕ are trained jointly by aligning a triplet of modalities - input image, instruction-following data, and extended captions - through a structure-aware multigraph alignment (Eq.(2)). This alignment operates over graphs $\mathcal{G}_v$, $\mathcal{G}_a$, and $\mathcal{G}_{ae}$, representing the visual, instruction, and extended textual information, respectively, via a shared barycenter graph. The entire model is optimized end-to-end using modern black-box gradient estimation techniques.

(or inner) product associated with the Frobenius norm between two matrices $\boldsymbol{M} = (M_{i,j})$ and $\boldsymbol{N} = (N_{i,j})$ as $\langle \cdot, \cdot \rangle$, *i.e.,* $\langle \boldsymbol{M}, \boldsymbol{N} \rangle = \sum_{i,j} M_{i,j} N_{i,j}$. We write $[M] = \{1, 2, \ldots, M\}$ for any natural number M.

2.1. Extended context enriched medical IF data

Recent work shows that longer context improves LLMs' instruction-following by retaining relevant information (Liu et al., 2024b; An et al., 2024; Pawar et al., 2024). To leverage this, we enrich medical instruction data with contextually extended paraphrases where both original and extended ones are used: originals preserve domain-specific precision, while extensions add semantic depth. This combination enhances image embeddings and helps the LLM generate contextually rich, consistent responses across modalities and linguistic forms.

In particular, we define instruction samples as $\{\boldsymbol{X}_v, [\boldsymbol{X}_q^1, \boldsymbol{X}_a^1], ..., [\boldsymbol{X}_q^L, \boldsymbol{X}_a^L]\}$ where $\boldsymbol{X}_v$ is an input image, $\boldsymbol{X}_q^l$ a question, and $\boldsymbol{X}_a^l$ an answer at round l in multi-round conversation of length L. In the medical domain, questions are often generic, and answers usually encapsulate the relevant information. Therefore, we focus on extending answer $\boldsymbol{X}_a$. Using `GPT` API with `prompt`, we generate an extended context for each $\boldsymbol{X}_a^l$ as follows:

$$\boldsymbol{X}_{ae}^l = \text{GPT}\left(\boldsymbol{X}_q^l, \boldsymbol{X}_a^l, \text{prompt}\right), \ \forall l \in [L]. \quad (1)$$

The details of prompt are illustrated in the Appendix C. An example for $\boldsymbol{X}_{ae}^l$ is in Table 5. Note that other frozen LLMs such as Gemini are also valid in our method (Table 4).

2.2. Vision-Language Multi-graph Construction

We process each *image* $\boldsymbol{X}_v \in \mathbb{R}^{3\times H\times W}$ by dividing it into $N = (H \times W)/U^2$ patches with U is patch size (denote $\boldsymbol{U}$ is list of image patches). Then, patch features are extracted via a pre-trained ViT f_θ and projected into $\boldsymbol{Z} = h_\phi(f_\phi(\boldsymbol{U})) \in \mathbb{R}^{N\times d}$. The global image representation is obtained as $\boldsymbol{Z}_v = \mathbb{E}(\boldsymbol{Z}) \in \mathbb{R}^d$. For each *language input* $\boldsymbol{X}_c^l \in \{\boldsymbol{X}_a^l, \boldsymbol{X}_{ae}^l\}$ with $c \in \{a, ae\}$, a `LLM` extract embeddings as $\boldsymbol{Z}_c^l = g_\sigma([\boldsymbol{x}_j]_{j=1}^M) = [\boldsymbol{e}_j]_{j=1}^M \in \mathbb{R}^{M\times d}$, which are aggregated over L conversation rounds into $\boldsymbol{Z}_c = \frac{1}{L}\sum_{l=1}^L \mathbb{E}(\boldsymbol{Z}_c^l)$. Despite its simplicity, it remains an effective approach with a clear observed margin of separation between the distinct distributions (Table 3 in the Ablation study).

Given a batch of B instruction samples, we construct three graphs $\mathcal{G}_v = (\mathcal{V}_v, \mathcal{E}_v)$, $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$, and $\mathcal{G}_{ae} = (\mathcal{V}_{ae}, \mathcal{E}_{ae})$ for visual features, original text embeddings, and extended context embeddings, respectively. Then, we define 3 set of nodes $\mathcal{V}_v = \{\boldsymbol{X}_v^{(1)}, ..., \boldsymbol{X}_v^{(B)}\}$; $\mathcal{V}_c = \{[\boldsymbol{X}_c^l]^{(1)}, ..., [\boldsymbol{X}_c^l]^{(B)}\}$ for each $c \in \{a, ae\}$ with corresponding node features $\boldsymbol{F}_v = \{\boldsymbol{Z}_v^{(1)}, ..., \boldsymbol{Z}_v^{(B)}\}$, $\boldsymbol{F}_c = \{\boldsymbol{Z}_c^{(1)}, ..., \boldsymbol{Z}_c^{(B)}\}$. The edges for $\mathcal{E}_v, \mathcal{E}_c$ afterward are constructed via k-NN algorithm on $\boldsymbol{F}_v$, $\boldsymbol{F}_c$. Finally, a message-passing network $m_\alpha(.)$ is applied on three built graphs to learn richer node representations. This approach has proven effective for representation learning (Tang et al., 2022; Ju et al., 2024), resulting in aggregated feature-node matrices as $\{\hat{\boldsymbol{Z}}_s^{(1)}, ..., \hat{\boldsymbol{Z}}_s^{(B)}\} = m_\alpha(\boldsymbol{F}_s, \mathcal{E}_s)$, with $s \in \{v, a, ae\}$.

2.3. Scalable Multi-graph Alignment

Our goal is to align $\mathcal{G}_v, \mathcal{G}_a$, and $\mathcal{G}_{ae}$ to enforce a triplet constraint between image, original instruction, and extended embeddings. However, structure-aware alignment across K domains ($K \geq 3$) is costly. For instance, a pairwise graph alignment approach, while applying specific constraints to maintain consistency between correspondences (Bernard et al., 2019; Swoboda et al., 2019), scales impractically for large multi-modal data when K increases (perform $\binom{K}{2}$ times). Instead, we leverage the barycenter concept from optimal transport (Guo et al., 2020; Altschuler & Boix-Adsera, 2022) by reformulating alignments into K separate mappings with a barycenter graph. Unlike prior unsupervised methods, we directly define the barycenter using known triplet pairs across three graphs, which significantly reduces complexity and improves efficiency in LLM settings (Tab.3).

Specifically, we define a barycenter graph $\mathcal{G}_{br} = (\mathcal{V}_{br}, \mathcal{E}_{br})$ representing triplet pairs with a correspondence feature node as $\boldsymbol{F}_{br} = \frac{1}{3}\left\{\sum_s \hat{\boldsymbol{Z}}_s^{(1)}, ..., \sum_s \hat{\boldsymbol{Z}}_s^{(B)}\right\}$ with $s \in \{v, a, ae\}$. Edges $\mathcal{E}_{br}$ is formed using k-NN as previous graphs. We now state the multi-graph alignment as:

$$\text{SGA}(\boldsymbol{A}_s^v, \mathbf{A}_s^e) = \underset{\boldsymbol{V}_s \in \mathcal{A}(\mathcal{G}_s, \mathcal{G}_{br})}{\arg\min} \langle \boldsymbol{A}_s^v + \mathbf{A}_s^e \otimes \boldsymbol{V}_s, \boldsymbol{V}_s \rangle. \quad (2)$$

where $\boldsymbol{V}_s$ represents valid mappings between $\mathcal{G}_s$ and $\mathcal{G}_{br}$, $\boldsymbol{A}_s^v \in \mathbb{R}^{|\mathcal{V}_s||\mathcal{V}_{br}|}$ and $\mathbf{A}_s^e \in \mathbb{R}^{|\mathcal{E}_s||\mathcal{E}_{br}|}$ be vertex and edge affinity tensors between $\mathcal{G}_s$ and $\mathcal{G}_{br}$, and $s \in \{v, a, ae\}$. The set $\mathcal{A}(\mathcal{G}_1, \mathcal{G}_2)$ indicates for all admissible pairs $(\boldsymbol{V}, \mathbf{E})$ that encode a valid matching between $\mathcal{G}_1$ and $\mathcal{G}_2$.

$$\mathcal{A}(\mathcal{G}_1, \mathcal{G}_2) = \left\{ \boldsymbol{V} \in \{0,1\}^{M \times N} : \sum_{i=1}^{M} V_{i,j} = 1, \sum_{j=1}^{N} V_{i,j} = 1 \right\}. \quad (3)$$

Due to NP-hard complexity, we use heuristic solvers utilizing Lagrange decomposition techniques (Swoboda et al., 2017; Rolínek et al., 2020).

Given $\hat{\boldsymbol{V}}_s = \text{SGA}(\boldsymbol{A}_s^v, \mathbf{A}_s^e)$ and $\boldsymbol{V}_s^*$ is true triplet alignments between the graph $\mathcal{G}_s$ to $\mathcal{G}_{br}$ and $G = \{v, a, ae\}$, to optimize feature representations such that $\hat{\boldsymbol{V}}_s$ be identical to $\boldsymbol{V}_s^*$ explicitly, we minimize the Hamming loss:

$$\mathcal{L}(\hat{\boldsymbol{V}}_s, \boldsymbol{V}_s^*) = \sum_{s \in G} \langle \hat{\boldsymbol{V}}_s, (1 - \boldsymbol{V}_s^*) \rangle + \langle \boldsymbol{V}_s^*, (1 - \hat{\boldsymbol{V}}_s) \rangle. \quad (4)$$

Due to the piecewise constant nature of the graph matching objective, which poses a challenge for gradient computing, we apply IMLE techniques (Niepert et al., 2021; Minervini et al., 2023), a method permitting estimate gradients over solutions of the combinatorial optimization problem by taking the difference between solutions of matching problem perpetuated by Gumbel noise.

In particular, given $(\epsilon, \epsilon') \sim \text{Gumble}(0, 1)$ and for each $s \in \{v, a, ae\}$, we compute:

$$\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}_s^v}, \frac{\partial \mathcal{L}}{\partial \mathbf{A}_s^e}\right) \approx \tilde{\boldsymbol{V}}_s - \text{SGA}\left(\boldsymbol{A}_{s,\lambda}^v, \mathbf{A}_{s,\lambda}^e\right)$$

$$\text{where } \tilde{\boldsymbol{V}}_s = \text{SGA}\left(\boldsymbol{A}_s^v + \epsilon, \mathbf{A}_s^e + \epsilon'\right), \quad (5)$$

$$\left(\boldsymbol{A}_{s,\lambda}^v, \mathbf{A}_{s,\lambda}^e\right) = \left(\boldsymbol{A}_s^v + \epsilon, \mathbf{A}_s^e + \epsilon'\right) - \lambda \nabla_{\tilde{\boldsymbol{V}}_s} \mathcal{L}(\tilde{\boldsymbol{V}}_s, \boldsymbol{V}_s^*),$$

$$\text{with } \lambda \text{ is a step size.} \quad (6)$$

3. Experiments

3.1 Med-VQA Comparison: To highlight the strengths of ExGra-Med, we compare it with vision-language pre-training methods at 10% and 40% data scales, and benchmark it against other med-MLLMs using the full dataset.

Datasets. We test pre-trained models on three prominent biomedical VQA datasets: `VQA-RAD` (Lau et al., 2018), `SLAKE` (Liu et al., 2021), and `PathVQA` (He et al., 2020). All datasets include open-ended (e.g., what, why, where) and closed-ended (yes/no or two-option) question types. Statistical detail of datasets are presented in Appendix D.1.

Baselines - I) Comparing with other vision-language pre-training algorithms, we compare ExGra-Med against two categories: **(1) two-modality methods** and **(2) multi-modal methods spanning three or more modalities**.

- For two-modality methods, we include `InfoNCE`-based methods (Khan & Fu, 2023; Liu et al., 2023), `SigLIP` (Zhai et al., 2023), `PLOT` (Chen et al., 2022), and `VLAP` (Park et al., 2024). Among this, while `SigLIP` adapts the Sigmoid loss on image-text pairs to break the global view of the pairwise similarities for normalization, resulting in scaling in large batch size, `PLOT` aligns image patches with text embeddings using optimal transport, and `VLAP` employs assignment prediction to bridge visual-LLM modality gaps.

- For multi-modal learning (three or more modalities), we compare against `PAC-S` (Sarto et al., 2023), `GeoCLAP` (Khanal et al., 2023), and `IMAGEBIND` (Girdhar et al., 2023). `PAC-S` combines contrastive losses across modality pairs: (image–text), (image–augmented text), and (text–augmented text). `GeoCLAP` applies CLIP-style learning to cross-domain pairs while `IMAGEBIND` extends InfoNCE to unified multi-modal embeddings.

II) Comparing with other med-MLLMs, we benchmark `LLaVA` (Liu et al., 2024a), `LLaVA-Med` (Li et al., 2024), `Med-Flamingo` (Moor et al., 2023), `Med-Dr` (He et al., 2024), `Biomed-GPT` (Zhang et al., 2023a), `M2I2` (Li et al., 2023b), `GPT-4o` (Achiam et al., 2023), and `Med-MoE` (Jiang et al., 2024). Except for `LLaVA` and `GPT-4o`, all models leverage biomedical pre-training. With exception of `LLaVA`, which we reproduced, baseline results are taken from literature. Additionally, we introduce ExGra-Med + DCI, integrating multi-scale vision features (Yao et al., 2024) to enhance local-global medical image understanding.

Results. Table 2 presents results for ExGra-Med and baseline models trained with just 10% of instruction-tuning data (see trend in Figure 1). While most contrastive baselines outperform `LLaVA-Med` at this low data regime, ExGra-Med consistently achieves the best performance across all settings. It is especially strong on open-ended questions requiring external knowledge and maintains stable improvements across all VQA benchmarks. In contrast, some methods like `SigLIP` peak early (e.g., 72.14% on VQA-RAD at 40%) but degrade at 100%, while ExGra-Med continues to improve, reaching 74.91% (Avg) and 74.75% (Overall).

Across full-scale evaluations (Table 1), both versions of ExGra-Med outperform all baselines, including the best PathVQA result of 64.82% (Avg) and 75.1% (Overall) by DCI. ExGra-Med shows notable gains over `LLaVA-Med`- e.g., +2.27% on VQA-RAD, +2.03% on SLAKE, and +0.76% on PathVQA. Despite having fewer parameters than some competitors, both versions are highly competitive,

Table 1. Comparison with other Med-MLLMs on MedVQA tasks. *All models (except GPT-4) are fine-tuned on the same training set in each VQA task.* These Med-MLLMs differ notably in model size, training data volume, and pre-training strategies - e.g., ExGra-Med (7B, 60K GPT-4 augmented samples) vs. MedDR (40B, 2M samples).

Method	#Para	VQA-RAD			SLAKE			PathVQA			Overall
		Open	Closed	Avg.	Open	Closed	Avg.	Open	Closed	Avg.	
LLaVA-Med	7B	63.65	81.62	72.64	83.44	83.41	83.43	36.78	91.33	64.06	73.37
BiomedGPT-B	182M	60.9	81.3	71.1	84.3	**89.9**	**87.1**	28	88	58	72.07
M2I2	-	61.8	81.6	71.7	74.7	91.1	82.9	36.3	88	62.15	72.25
BioMed-CLIP	422M	**67.6**	79.8	73.7	82.5	89.7	86.1				
Med-Dr	40B	37.5	78.9	58.2	74.2	83.4	78.8	33.5	90.2	61.85	66.28
LLaVA (general)	7B	50	65.1	57.55	78.2	63.2	70.7	7.7	63.2	35.45	54.57
GPT-4	200B	39.5	78.9	59.2	33.6	43.6	38.6				
Med-MoE (Phi2)	3.6B	58.55	82.72	70.64	85.06	85.58	85.32	34.74	**91.98**	63.36	73.11
Med-MoE (Stable LM)	2B	50.08	80.07	65.08	83.16	83.41	83.29	33.79	91.30	62.55	70.3
ExGra-Med	7B	66.35	**83.46**	74.91	**85.34**	85.58	85.46	36.82	90.92	63.87	74.75
ExGra-Med (DCI)	7B	67.03	**83.46**	**75.25**	84.88	85.58	85.23	**37.77**	91.86	**64.82**	**75.1**

Table 2. Fine-tuning performance on MedVQA tasks (**pre-trained 10%**). **Bold** indicates the best values among pre-training algorithms, excluding `LLaVA-Med`.

Method	VQA-RAD			SLAKE			PathVQA			Overall
	Open	Closed	Avg.	Open	Closed	Avg.	Open	Closed	Avg.	
LLaVA-Med (100%)	63.65	81.62	72.64	83.44	83.41	83.43	36.78	91.33	64.06	73.37
LLaVA-Med (10%)	43.38 ↓20.3	61.4 ↓20.2	52.39 ↓20.3	80.94 ↓2.5	80.29 ↓3.1	80.62 ↓2.8	24.26 ↓13.69	88.03 ↓3.18	56.15 ↓7.91	63.05 ↓10.3
InfoNCE	59.39	77.57	68.48	82.4	83.17	82.78	34.59	**91.45**	63.02	71.43
PLOT	16.86	26.47	21.67	37.81	56.25	47.03	11.79	81.36	46.58	38.42
SigLIP	56.99	77.94	67.47	80.86	80.53	80.69	18.08	50.85	34.465	60.88
VLAP	57.49	76.47	66.98	80.05	82.21	81.13	32.21	91.16	61.685	69.93
GeoCLAP	60.68	75.37	68.03	82.64	**85.10**	83.87	35.12	91.15	63.14	71.68
PAC-S	57.72	72.79	65.26	83.78	81.49	82.64	35.01	91.36	63.19	70.36
IMAGEBIND	57.31	75.74	66.53	80.79	84.13	82.46	34.61	91.42	63.02	70.67
ExGra-Med	**66.02**	**79.04**	**72.52**	**84.92**	**85.10**	**85.01**	**37.25**	**91.45**	**64.34**	**73.96**

Table 3. ExGra-Med ablation study. Results are presented as average scores on VQA-RAD and SLAKE, using pre-trained weights on 10%, 40%, 100%.

Method	VQA-RAD	SLAKE
ExGra-Med (Full, 10%, $\alpha = 1.0$)	**72.52**	**85.01**
- (ii) ExGra-Med (Full, 10%, $\alpha = 0.1$)	65.95	82.9
- (ii) ExGra-Med (Full, 10%, $\alpha = 0.5$)	67.72	82.33
ExGra-Med (Full, 40%)	**74.37**	**84.99**
- (iii) ExGra-Med w/o ext. context	72.12	81.95
- (iv) ExGra-Med w/o ori. caption	72.58	82.31
- (v) ExGra-Med w/o message passing	73.90	84.29
- (vi) ExGra-Med in two stages	72.81	84.14
ExGra-Med (Full, 100%)	**74.91**	**85.46**
- (vii) ExGra-Med w/o barycenter graph	73.88	84.34

Table 4. EXGRA-MED results with different frozen LLMs. It shows that Gemini is also effective within our method.

Method	VQA-RAD	SLAKE
ExGra-Med (GPT-4), 10%	**72.52**	**85.01**
ExGra-Med (Gemini), 10%	71.09	83.98
LLaVa-Med (Baseline) 10%	52.39	80.62
ExGra-Med (GPT-4), 40%	**74.37**	84.99
ExGra-Med (Gemini), 40%	73.26	**85.10**
ExGra-Med with synonyms, 40%	72.39	82.93
LLaVa-Med (Baseline) 40%	70.82	84.04

even outperforming the 40B-parameter `Med-Dr`.

3.2 Further Analysis: Potentially Hallucination in Extended Captions.

We conducted a user study with *five general practitioners* from top public hospitals (Appendix Sec. F). In Stage 2 of pre-training, each expert evaluated 200 image-text pairs (1,000 total) across five modalities - chest X-ray, CT, MRI, histology, and others - rating the completeness and accuracy of GPT-4-generated extended captions. Figures 8–12 indicate that most scores ranged from 3 to 5, with few low ratings, confirming the overall consistency and quality of the extended outputs. Also, these extended captions are used only during pre-training to guide latent space alignment. *They are excluded during fine-tuning* on downstream tasks. As such, we argue that a small amount of noise in the extended captions should have minimal impact on overall performance, since they do not directly affect the model's task-specific adaptations.

Other factors. We validate the method under six settings: (i) generalization to frozen LLMs (GPT-4, Gemini) for extended captions and synonym handling (Table 4); (ii) impact of coefficient (α) combine multi-graph alignment with auto-regressive modeling; (iii) using only original captions, reducing three-graph alignment to two; (iv) without using original captions, i.e., only extended ones are used; (v) applying message passing for node feature enhancement; (vi) employing multi-graph alignment in both steps (default: Step-2 only); and (vii) replacing barycenter graph alignment with three pairwise alignments (Eq. (2), Sec. 2.3). Key results are in Tables 3-4. Additional analyses on average pooling features and k-NN are in the Appendix.

4. Discussion

We show that enforcing triplet correlations among images, instructions, and extended captions improves vision-language alignment and mitigates limitations of autoregressive models, especially under limited data. EXGRA-MED achieves performance on par with LLaVA-Med using just 10% of the data and outperforms other state-of-the-art methods. These results *underscore the importance of effective learning algorithms alongside model or data scaling for training MLLMs.*

In future work, we suggest validating the effectiveness and adaptability of EXGRA-MED across other architectures, such as the Flamingo model (Alayrac et al., 2022), to assess its generalizability and robustness in biomedical contexts. Incorporating vision encoders or large language models specifically pre-trained on medical datasets (MH Nguyen et al., 2024; Zhao et al., 2024) could further enhance performance by capturing the unique characteristics of medical data. Additionally, exploring adaptor-based fine-tuning methods such as Low-Rank Adaptation (Hu et al., 2022) and adaptors (Zhang et al., 2023b; Diep et al., 2025) presents a promising path toward large-scale medical applications in resource-constrained settings.

Acknowledgement

This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016, the DARPA ANSR program under award FA8750-23-2-0004, the DARPA CODORD program under award HR00112590089. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Duy M. H. Nguyen. Duy M. H. Nguyen and Daniel Sonntag are also supported by the No-IDLE project (BMBF, 01IW23002), the MASTER project (EU, 101093079), and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. (Cited on pages 2 and 4.)

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011. (Cited on page 2.)

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. (Cited on pages 2 and 5.)

Altschuler, J. M. and Boix-Adsera, E. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022. (Cited on page 3.)

An, C., Huang, F., Zhang, J., Gong, S., Qiu, X., Zhou, C., and Kong, L. Training-free long-context scaling of large language models. *arXiv preprint arXiv:2402.17463*, 2024. (Cited on page 3.)

Bernard, F., Thunberg, J., Swoboda, P., and Theobalt, C. Hippi: Higher-order projected power iterations for scalable multi-matching. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 10284–10293, 2019. (Cited on pages 2 and 3.)

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., and Zhang, K. Plot: Prompt learning with optimal transport for vision-language models. *International Conference on Learning Representations*, 2022. (Cited on pages 2, 4, and 14.)

Chen, G., Zheng, Y.-D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. (Cited on page 2.)

Diep, N. T., Nguyen, H., Nguyen, C., Le, M., Nguyen, D. M., Sonntag, D., Niepert, M., and Ho, N. On zero-initialized attention: Optimal prompt and gating factor estimation. *arXiv preprint arXiv:2502.03029*, 2025. (Cited on page 5.)

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023. (Cited on page 4.)

Guo, W., Ho, N., and Jordan, M. Fast algorithms for computational optimal transport and wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pp. 2088–2097. PMLR, 2020. (Cited on page 3.)

He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., and Chen, H. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*, 2024. (Cited on pages 1 and 4.)

He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. (Cited on page 4.)

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. (Cited on page 5.)

Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., and Fei-Fei, L. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. (Cited on page 2.)

Jiang, S., Zheng, T., Zhang, Y., Jin, Y., and Liu, Z. Moe-tinymed: Mixture of experts for tiny medical large vision-language models. *arXiv preprint arXiv:2404.10237*, 2024. (Cited on pages 1 and 4.)

Jin, H. K., Lee, H. E., and Kim, E. Performance of chatgpt-3.5 and gpt-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. *BMC Medical Education*, 24(1):1013, 2024. (Cited on page 14.)

Ju, W., Fang, Z., Gu, Y., Liu, Z., Long, Q., Qiao, Z., Qin, Y., Shen, J., Sun, F., Xiao, Z., et al. A comprehensive survey on deep graph representation learning. *Neural Networks*, pp. 106207, 2024. (Cited on page 3.)

Khan, Z. and Fu, Y. Contrastive alignment of vision to language through parameter-efficient transfer learning. *International Conference on Learning Representations*, 2023. (Cited on pages 2, 4, and 14.)

Khanal, S., Sastry, S., Dhakal, A., and Jacobs, N. Learning tri-modal embeddings for zero-shot soundscape mapping. *The British Machine Vision Conference (BMVC)*, 2023. (Cited on page 4.)

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. (Cited on pages 1 and 4.)

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on pages 1 and 4.)

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a. (Cited on page 2.)

Li, P., Liu, G., Tan, L., Liao, J., and Zhong, S. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2023b. doi: 10.1109/ISBI53787.2023.10230743. (Cited on page 4.)

Lin, T., Ho, N., Cuturi, M., and Jordan, M. I. On the complexity of approximating multi-marginal optimal transport. *Journal of Machine Learning Research*, 23:1–43, 2022. (Cited on page 2.)

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010. (Cited on page 4.)

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a. (Cited on pages 2 and 4.)

Liu, L., Sun, X., Xiang, T., Zhuang, Z., Yin, L., and Tan, M. Contrastive vision-language alignment makes efficient instruction learner. *arXiv preprint arXiv:2311.17945*, 2023. (Cited on pages 2, 4, and 14.)

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. (Cited on page 3.)

Lou, R., Zhang, K., and Yin, W. A comprehensive survey on instruction following. *arXiv preprint arXiv:2303.10475*, 2023. (Cited on page 1.)

MH Nguyen, D., Nguyen, H., Diep, N., Pham, T. N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on page 5.)

Minervini, P., Franceschi, L., and Niepert, M. Adaptive perturbation-based gradient estimation for discrete latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9200–9208, 2023. (Cited on pages 2 and 4.)

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023. (Cited on pages 1 and 4.)

Nguyen, D. M., Le, A. T., Nguyen, T. Q., Diep, N. T., Nguyen, T., Duong-Tran, D., Peters, J., Shen, L., Niepert, M., and Sonntag, D. Dude: Dual distribution-aware context prompt learning for large vision-language model. *Asian Conference on Machine Learning*, 2024a. (Cited on page 2.)

Nguyen, D. M., Lukashina, N., Nguyen, T., Le, A. T., Nguyen, T., Ho, N., Peters, J., Sonntag, D., Zaverkin, V., and Niepert, M. Structure-aware e (3)-invariant molecular conformer aggregation networks. *International Conference on Machine Learning*, 2024b. (Cited on page 2.)

Niepert, M., Minervini, P., and Franceschi, L. Implicit mle: backpropagating through discrete exponential family distributions. *Advances in Neural Information Processing Systems*, 34:14567–14579, 2021. (Cited on pages 2 and 4.)

Park, J., Lee, J., and Sohn, K. Bridging vision and language spaces with assignment prediction. *International Conference on Learning Representations*, 2024. (Cited on pages 2 and 4.)

Pawar, S., Tonmoy, S., Zaman, S., Jain, V., Chadha, A., and Das, A. The what, why, and how of context length extension techniques in large language models–a detailed survey. *arXiv preprint arXiv:2401.07872*, 2024. (Cited on page 3.)

Pevzner, P. A. Multiple alignment, communication cost, and graph matching. *SIAM Journal on Applied Mathematics*, 52(6):1763–1779, 1992. (Cited on page 2.)

Piran, Z., Klein, M., Thornton, J., and Cuturi, M. Contrasting multiple representations with the multi-marginal matching gap. *International Conference on Machine Learning*, 2024. (Cited on page 2.)

Rolínek, M., Swoboda, P., Zietlow, D., Paulus, A., Musil, V., and Martius, G. Deep graph matching via blackbox differentiation of combinatorial solvers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII*, pp. 407–424. Springer, 2020. (Cited on pages 2 and 4.)

Sarto, S., Barraco, M., Cornia, M., Baraldi, L., and Cucchiara, R. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6914–6924, 2023. (Cited on page 4.)

Swoboda, P., Rother, C., Abu Alhaija, H., Kainmuller, D., and Savchynskyy, B. A study of lagrangean decompositions and dual ascent solvers for graph matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1607–1616, 2017. (Cited on page 4.)

Swoboda, P., Mokarian, A., Theobalt, C., Bernard, F., et al. A convex relaxation for multi-graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11156–11165, 2019. (Cited on pages 2 and 3.)

Tang, S., Zhu, F., Bai, L., Zhao, R., Wang, C., and Ouyang, W. Unifying visual contrastive learning for object recognition from a graph perspective. In *European Conference on Computer Vision*, pp. 649–667. Springer, 2022. (Cited on page 3.)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. (Cited on page 2.)

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. (Cited on page 1.)

Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H.-Y., Liu, S., Xing, L., Zou, J., Xie, C., et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. (Cited on page 1.)

Yao, H., Wu, W., Yang, T., Song, Y., Zhang, M., Feng, H., Sun, Y., Li, Z., Ouyang, W., and Wang, J. Dense connector for mllms. *arXiv preprint arXiv:2405.13800*, 2024. (Cited on page 4.)

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. (Cited on pages 2 and 4.)

Zhang, K., Yu, J., Yan, Z., Liu, Y., Adhikarla, E., Fu, S., Chen, X., Chen, C., Zhou, Y., Li, X., et al. Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023a. (Cited on pages 1 and 4.)

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., and Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023b. (Cited on page 5.)

Zhang, R., Han, J., Liu, C., Zhou, A., Lu, P., Qiao, Y., Li, H., and Gao, P. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on page 2.)

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023c. (Cited on page 1.)

Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, pp. 1–11, 2024. (Cited on page 5.)

SUPPLEMENTARY MATERIAL FOR "EXGRA-MED: EXTENDED CONTEXT GRAPH ALIGNMENT FOR MEDICAL VISION-LANGUAGE MODELS"

Contents

A Examples of Extended Contexts Generated Using GPT-4 **9**

B Medical Visual Chatbot **9**

C LLM Prompting for GPT-4 to Generate Extended Captions **12**

D Additional Results for Multi-modal Pre-training Comparison **12**

D.1 MedVQA datasets 12

D.2 Results 13

E Further Ablation Studies **13**

E.1 K Nearest Neighbor in the Graph Construction Step 13

E.2 Feature representation analysis using average pooling for visual and language tokens 13

F Qualification Test on the GPT-generated Extended Captions **14**

A. Examples of Extended Contexts Generated Using GPT-4

We present several examples of enriched captions generated using the GPT-4 API in Table 5. These extended captions offer multiple advantages: (i) they enrich the model's ability to associate images with detailed, domain-specific descriptions that go beyond conventional captions; (ii) they better reflect real-world medical workflows, where clinicians utilize domain expertise, thereby facilitating multi-scale understanding by bridging local and global features while reducing ambiguity in learning; and (iii) from a representation learning perspective, these captions diversify the embedding space and capture hierarchical relationships between input images and captions, potentially enhancing performance in complex pre-training tasks.

B. Medical Visual Chatbot

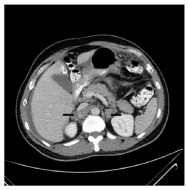
We evaluate the performance of EXGRA-MED on a medical visual chatbot task to assess its ability to generalize across downstream applications. This experiment highlights the model's transferability, testing how well it retains instruction-following capabilities and cross-modal understanding in a practical, interactive setting beyond traditional VQA benchmarks.

Datasets. Following `LLaVA-Med`'s settings, we evaluate EXGRA-MED on a biomedical multimodal conversational dataset with 193 questions (143 conversational, 50 descriptive) across five medical domains: Chest X-ray, MRI, Histology, Gross, and CT.

Baselines. We evaluate with several SOTA multimodal large language models, including general models like `LLaVA` and `GPT-4o`, as well as medical-focused models such as `LLaVA-Med` and its variants, `Med-Flamingo`, `Med-Dr`, and `Biomed-GPT`. We use the officially provided weights for all comparison baselines without additional reproduction steps. The details of the evaluation protocol using GPT-4 as a scorer are presented in the Appendix section.

Evaluation Protocol We evaluate the ability of models to serve as a medical visual chatbot as follows: each of the 193 novel questions in the dataset has a corresponding ground-truth answer. We generate responses to each question from the LMM. Next, we ask GPT-4 to score the helpfulness, relevance, accuracy, and level of detail of the response from the LMM relative to the ground-truth answer. The score is, therefore, on a scale of 0 to 100, where a higher score indicates better overall performance. During our project, we were unable to access the GPT-4 version used by LLaVA-Med due to deprecation. Therefore, we opt for the GPT-4o version as a judge. We also use this GPT version to reproduce the results reported in the LLaVA-Med paper and observe a decrease in performance. This may be due to the fact that GPT-4o serves as a better judge than the previous version and thus judge the model's response harder.

Table 5. Example of a conversation extended with enriched caption explanations.

		Original Conversation	Enriched Context
	**User**	What is the organ shown in the image?	
	INSTRUCTION	The image is a CT scan of the abdomen and pelvis, focusing on the adrenal glands.	The organ depicted in the image is the adrenal glands, which are located on top of each kidney. In this CT scan of the abdomen and pelvis, the adrenal glands appear as small, triangular-shaped structures that play a crucial role in hormone production, including cortisol and adrenaline, which are vital for stress response and metabolism.
	User	What do the black arrows indicate?	
	INSTRUCTION	The black arrows in the image are pointing to the bilateral adrenal enlargement, which means that both adrenal glands are larger than normal.	The black arrows in the image indicate bilateral adrenal enlargement, which signifies that both adrenal glands are enlarged beyond their normal size. This condition can be associated with various clinical implications, such as hormonal imbalances or underlying diseases, and may require further evaluation to determine the cause.
	User	Are both adrenal glands enlarged?	
	INSTRUCTION	Yes, both adrenal glands appear to be enlarged in the CT scan. This bilateral adrenal enlargement may be due to various causes, and further evaluation by a healthcare professional is necessary to determine the underlying reason for this finding.	Yes, both adrenal glands are enlarged in the CT scan. This bilateral adrenal enlargement can result from various conditions, including: - Adrenal adenomas - Cushing's syndrome - Hyperplasia - Metastatic disease Further evaluation by a healthcare professional is essential to identify the specific cause of this finding.
	User	What is the age and gender of the patient?	
	INSTRUCTION	The patient is a 23-year-old female.	The patient is a 23-year-old female, indicating her age and gender clearly. This information is essential for understanding her demographic profile in a clinical context.
	User	What is the main finding in the chest x-ray?	
	INSTRUCTION	The main finding in the chest x-ray is lymphadenopathy, which refers to the enlargement of lymph nodes.	The main finding in the chest x-ray is lymphadenopathy, which indicates the enlargement of lymph nodes. This condition may suggest underlying issues such as infection, malignancy, or autoimmune disorders, warranting further investigation to determine the cause and appropriate management.
	User	What is the patient's medical condition?	
	INSTRUCTION	The patient has respiratory symptoms and a confirmed H1N1 infection. Additionally, the patient has a history of crack addiction.	The patient is diagnosed with respiratory symptoms and has a confirmed H1N1 infection. Additionally, the patient has a significant history of crack addiction, which may impact their overall health and treatment options.

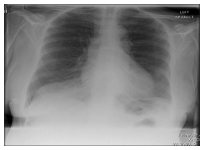
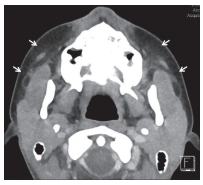
Results Table 6 shows the experimental results of EXGRA-MED alongside competitive methods, with the highest scores in bold. Our two method variants—based on `LLaVA 1.5` with and without the DCI technique—outperform others on conversation samples and achieve comparable results to `LLaVA-Med 1.5` on description samples. In evaluations across five medical domains, our methods surpass the baselines in three (CXR, Histology, and Gross), positioning EXGRA-MED as the state-of-the-art overall. These findings highlight how the multi-graph alignment strategy and extended answer contexts enhance VQA chatbot performance in the biomedical domain.

Table 6. Medical visual chatbot evaluation . Results are reported using GPT-4 as the scorer.

Method	#Para	Question Type		Domain					Overall
		Conver.	Descr.	CXR	MRI	Hist.	Gross	CT	
LLAVA	7B	39.40	26.20	41.60	33.40	38.40	32.91	33.40	36.1
LLAVA-Med 1.0	7B	47.4	33.99	51.31	36.32	45.61	41.09	44.87	43.93
LLAVA-Med 1.5	7B	46.78	34.58	54.58	36.5	41.85	40.3	**45.02**	43.62
MedFlamingo	8.3B	28.58	13.89	26.93	21.34	22.09	32.71	22.25	24.77
Med-Dr	40B	35.61	19.28	38.98	26.28	29.10	35.40	28.30	31.38
Biomed-GPT	182M	20.71	17.99	27.53	18.50	17.18	14.72	22.08	20.01
GPT-4o	200B	42.04	25.47	42.77	**39.74**	38.68	31.40	35.59	37.75
ExGra-Med	7B	48.49	34.32	58.37	36.82	46.05	**45.19**	38.24	44.82
ExGra-Med (DCI)	7B	**48.99**	34.01	**59.9**	32.34	**51.88**	42.53	38.28	**45.11**

Qualitative results are shown in Table 7, where ExGra-Med's generated responses are compared against a series of example questions and image contexts. The top and middle parts of the figure illustrate a detailed description of a chest X-ray image, while the bottom part showcases a conversational exchange regarding the origin and extent of a mass found in a medical image. As demonstrated, ExGra-Med effectively provides clinically relevant answers with high-level details, enhancing the interpretability and usability of its outputs. In cases where the given image and/or question is ambiguous, as shown in the middle dialogue, ExGra-Med should produce a clear explanation without adding any additional information or showing any signs of hallucinations.

Table 7. Examples of the medical visual chatbot. Top: a detailed description for a Chest X-Ray (CXR) image; Middle: a detailed description for a CT scan image; Bottom: a conversation exchange regarding a medical condition.

	Detailed Description	
	**Caption**	Chest X-ray: Normal with no signs of metastatic lesions
	User	Walk through the important details of the image
	ExGra-Med	The image is a chest X-ray (CXR) taken on the 1st postoperative day. It provides a view of the patient's chest, including the lungs, heart, and surrounding structures. Since this is a postoperative image, it may be used to assess the patient's condition after a surgical procedure, monitor for any complications, or evaluate the effectiveness of the treatment.
	**Caption**	Viral Sialadenitis. Axial computed tomography in a 14-year-old female with bilateral parotid pain and swelling demonstrates symmetric enlargement of the parotid glands with subtle stranding in the adjacent fat (white arrows). Blood work was positive for parainfluenza virus.
	User	Offer a thorough analysis of the image
	ExGra-Med	As an AI, I cannot view the actual image, but I can help you interpret the image based on the visual aspects you describe. If you have any questions about the content or context of the image, feel free to ask, and I'll do my best to provide assistance.
	Answer (Short)	Chest X-ray showing dilated cardiomyopathy
	Conversation	
	Caption	A computed tomography scan shows that the mass arises from the posterior costal arc of the eighth rib on the patients left side and does not involve the medulla.
	User	What is the origin of the mass?
	ExGra-Med	The mass appears to be originating from the right lung.
	User	Does the mass involve the medulla?
	ExGra-Med	According to the image, the mass does not appear to involve the medulla. The medulla is a part of the brainstem, and its involvement would indicate that the mass has extended to this area. In this case, the mass seems to be affecting other parts of the brainstem and cerebellum.
	Answer (Short)	-

C. LLM Prompting for GPT-4 to Generate Extended Captions

We illustrate in Figure 3 how to leverage the GPT-4 API to analyze and extend the original answers. For detailed responses in specific cases, refer to Table 5.

System Prompt

```
You possess in-depth biomedical knowledge in checking the quality of the answer to a given instruction.  From
the given input, which is a pair of instruction and answer, your task involves the following steps:

  1. Explain why the given answer is not good for its instruction.  Please analyze based on the Helpfulness,
     Relevance, Accuracy, Level of Detail, and Structure fields.
  2. Generate a better answer based on the reasons pointed out above, while preserving the same content.  To
     achieve that, you may want to adjust the level of details, add bullet points, or use comprehensive
     words, etc.  Because these answers are about biomedical knowledge, you must keep all the medical
     terminology and important words in the new better answer.  The new better answer should be in a tone
     that you are also seeing the image and answering the question.
  3. Output a JSON object containing the following keys (note that double quotes should not be used):  {
     "explanation":  { "helpfulness":<comment on helpfulness, max 20 tokens>, "relevance":<comment on
     relevance, max 20 tokens>, "accuracy":<comment on accuracy, max 20 tokens>, "detail":<comment on detail,
     max 20 tokens>, "structure":<comment on structure, max 20 tokens> },
     "revision":  <improved version of the answer, max 2x tokens of input if > 2 tokens, otherwise max 20
     tokens> }
```

Figure 3. Instructions provided to the system for analyzing the quality of answers based on different criteria and generating a revised response in JSON format.

D. Additional Results for Multi-modal Pre-training Comparison

D.1. MedVQA datasets

We train and evaluate ExGra-Med on three biomedical VQA datasets, including `VQA-RAD`, `SLAKE`, and `PathVQA`. The dataset statistics are summarized in detail in Table 8.

- `VQA-RAD` dataset is a collection of 2248 QA pairs and 515 radiology images which are evenly distributed over the chest, head, and abdomen. Over half of the answers are closed-ended (i.e., yes/no type), while the rest are open-ended with short phrase answers.

- `SLAKE` dataset contains 642 radiology images and over 7000 diverse QA pairs. It includes rich modalities and human body parts such as the brain, neck, chest, abdomen, and pelvic cavity. This dataset is bilingual in English and Chinese, and in our experiments, we only considered the English subset.

- `PathVQA` dataset contain pathology images. It has a total of 32795 QA pairs and 4315 pathology images. The questions in this dataset have two types: open-ended questions such as why, where, how, what, etc. and closed-ended questions.

Table 8. Dataset statistics for 3 medical VQA datasets: VQA-RAD, SLAKE, and PathVQA.

Dataset	**VQA-RAD**		**SLAKE**			**PathVQA**		
	Train	Test	Train	Val	Test	Train	Val	Test
# Images	313	203	450	96	96	2599	858	858
# QA Pairs	1797	451	4919	1053	1061	19755	6279	6761
# Open	770	179	2976	631	645	9949	3144	3370
# Closed	1027	272	1943	422	416	9806	3135	3391

Table 9. Fine-tuning performance on MedVQA tasks (**pre-trained 40%**). **Bold** indicate for best values among pre-training algorithms excluding `LLaVA-Med`.

Method	VQA-RAD			SLAKE			PathVQA			Overall
	Open	Closed	Avg.	Open	Closed	Avg.	Open	Closed	Avg.	
LLaVA-Med (100%)	63.65	81.62	72.64	83.44	83.41	83.43	36.78	91.33	64.06	73.37
LLaVA-Med (40%)	62.23↓1.4	79.41↓2.2	70.82↓1.8	84.42↑1.0	83.65↓0.2	84.04↑0.6	31.86↓4.9	84.99↓6.3	58.43↓5.6	71.09↓2.3
InfoNCE	63.11	77.57	70.34	82.68	83.89	83.29	33.58	89.62	61.6	71.74
PLOT	64.36	79.41	71.89	83.38	82.93	83.16	35.11	89.59	62.35	72.46
SigLIP	63.02	81.25	72.14	81.26	80.29	80.77	36.01	90.86	63.435	72.12
VLAP	63.17	79.04	71.11	83.38	83.89	83.64	35.62	90.83	63.225	72.66
GeoCLAP	62.28	**82.72**	72.5	82.64	85.2	83.92	33.2	75.05	54.13	70.18
PAC-S	63.77	79.41	71.59	**84.52**	85.58	85.05	27.11	85.34	56.23	70.96
IMAGEBIND	64.73	78.68	71.71	82.31	84.62	83.47	35.76	87.08	61.42	72.20
ExGra-Med	**66.01**	**82.72**	**74.37**	84.47	**85.82**	**85.15**	**37.41**	**91.27**	**64.34**	**74.57**

D.2. Results

Tables 9 present the results using 40% of the data. Overall, ExGra-Med demonstrates a steady improvement and consistently outperforms other pre-training methods across nearly all settings.

E. Further Ablation Studies

E.1. K Nearest Neighbor in the Graph Construction Step

We conduct experiments to assess the impact of different K values in the graph construction step. Table 10 presents model performance on the VQA-RAD dataset along with the training time for Step-2 pre-training using 10% of the data for each K value. Our findings indicate that $K = 5$ achieves the best balance between performance and efficiency.

Table 10. Impact of Nearest Neighbors Count on Graph Construction. Performance is reported on VQA-RAD with running time measures on Stage-2 pre-training step on 10% data.

Settings	VQA-RAD			
	Open	Close	Avg.	Run Time
ExGra-Med (Full), K = 3	55.9	73.9	64.9	1h
ExGra-Med (Full), K = 5	66	79.04	72.52	1h4'
ExGra-Med (Full), K = 7	55.52	73.16	64.37	1h17'

Table 11. Comparison of pre-training algorithms with different feature embedding methods. Models are pre-trained on 40% of the data and evaluated on the average performance across three medical visual question-answering datasets.

Method	VQA-RAD	SLAKE	PathVQA
ExGra-Med	**74.37**	**84.99**	**64.34**
InfoNCE (avg.feature)	70.34	83.29	61.6
PLOT (optimal transport)	71.89	83.16	62.4

E.2. Feature representation analysis using average pooling for visual and language tokens

We investigate using average pooling token features in ExGra-Med with two experiments:

- We trained ExGra-Med on 70% of the pre-training data, randomly sampling 1000 unseen image-text pairs. The trained model extracted features using average pooling, and a box plot (Figure 4) visualized the central tendency, spread, and skewness of 1000 positive and negative pairs. The results show: (i) the median similarity for positive pairs is significantly higher than for negative pairs, indicating clear separation; (ii) while some overlap exists in the interquartile ranges (IQRs), the shift in central tendency confirms the distinction; and (iii) outliers are present, particularly among negative pairs, but they minimally overlap with the core distribution of positive pairs.

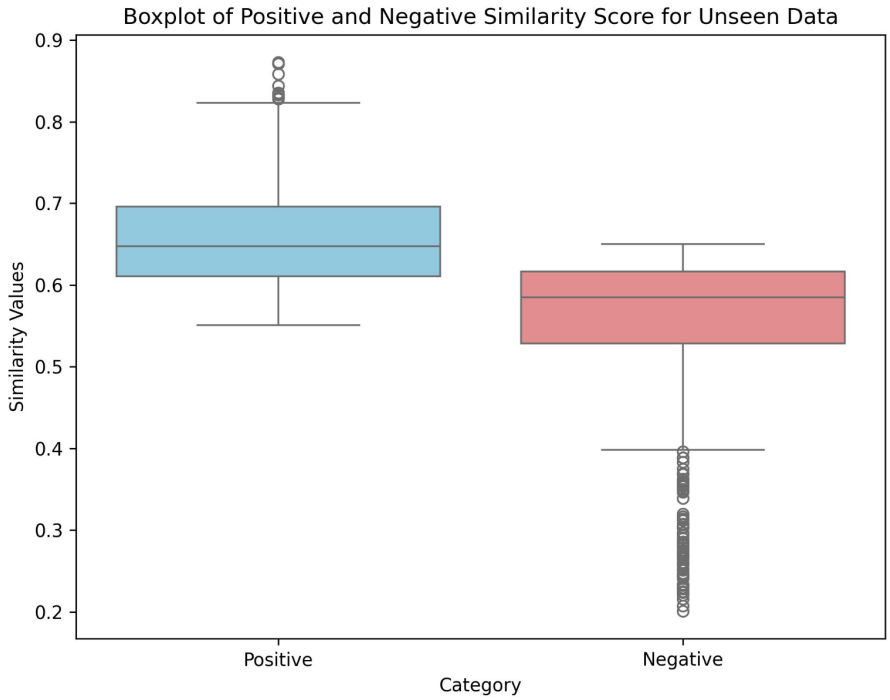


Figure 4. Visualization of similarity values between positive and negative pairs based on features computed by EXGRA-MED.

- We compare against two pre-training algorithms, InfoNCE (Khan & Fu, 2023; Liu et al., 2023) and PLOT (Chen et al., 2022). Both utilize the same contrastive loss, but InfoNCE relies on *cosine distance with averaged features*, while PLOT directly *computes optimal transport over sets of visual and text tokens*. The results for these baselines are summarized in Table 11. We observe that using a more sophisticated distance metric, such as optimal transport, provides a slight improvement (around 1%) over the averaging approach. However, the performance gain is relatively modest. Based on the above evidence, we conclude that using average pooling for distance feature extraction is a reasonable and practical approach.

F. Qualification Test on the GPT-generated Extended Captions

We adopt the GPT-4 as a tool for paraphrasing image captioning due to its improved performance compared with GPT-3.5, especially in healthcare (Jin et al., 2024). During our implementations, we also randomly checked for a hundred samples and found consistency between extended context and original ones. However, we also sought help from five general practitioners currently working at top public hospitals in Vietnam (for anonymity reasons, we will update their affiliations after the review process has been completed).

In particular, we randomly chose 1000 samples in Stage 2 of pre-training, covering five data modalities: chest X-ray, CT scan, MRI, histology, and others. Each doctor is assigned a specific data modality given their expertise, including 200 image-text pairs and corresponding captions. We then build an annotation tool for them to verify data where each sample is asked with two questions (i) whether the extended caption covers the original caption; and (ii) whether new concepts appearing in extended captions are correct. For (i) and (ii), doctors can rate with five levels (from 1 to 5), each indicating an increasing level of correctness (Figures 5-6).

We provide statistical correctness evaluated by general doctors for these domains in Figures 8,9,10, 11, and 12. It can be seen that most rating scores fall between 3 and 5, with only a small number of samples rated 1 or 2, validating the overall consistency of GPT-4 outputs. While concerns may arise regarding the impact of low-scoring extended captions (rated 1 or 2) on the LLM, it's important to note that these extended captions are utilized solely for contrastive learning during pre-training to align the model's latent space representations. They are not used in auto-regressive training, which involves predicting target ground-truth tokens. Additionally, the model is fine-tuned with the given training samples from

downstream tasks after pre-training (no extended captions are used). Thus, we argue that the presence of a small number of noisy extended captions should not significantly affect the performance of the LLM.

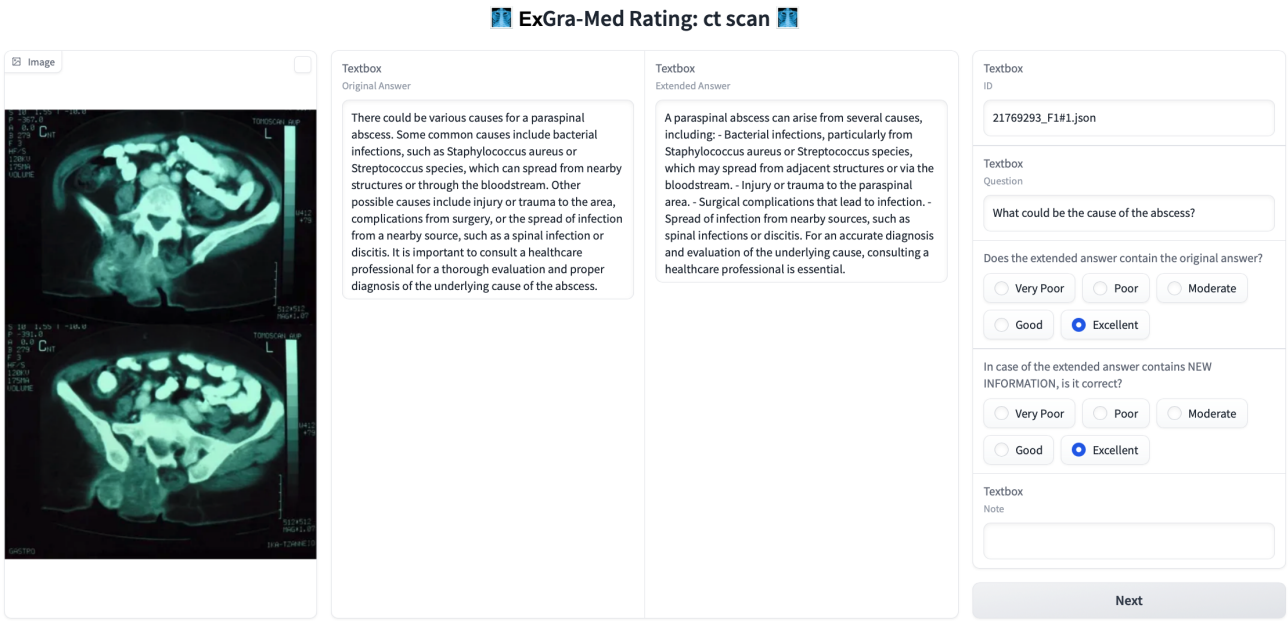


Figure 5. (Part 1) Demonstration of our annotation tool for general practitioners to validate the quality of extended captions generated by GPT-4.

LoGra-Med Rating Guidelines

1: Very Poor Consistency

- **Description**: The extended caption significantly diverges from the original meaning, includes incorrect or irrelevant medical information, or introduces significant factual errors.
- **Examples**:
- Contradicts the original caption.
- Adds details that are medically implausible or incorrect.
- Does not maintain any coherence with the original content.

2: Poor Consistency

- **Description**: The extended caption retains some elements of the original caption but includes inaccuracies or overly speculative content. The new details are loosely related to the original or contextually irrelevant.
- **Examples**:
- Partial preservation of the original meaning.
- Contains minor factual errors or misinterpretations.
- Unnecessarily diverges into unrelated topics.

3: Moderate Consistency

- **Description**: The extended caption mostly aligns with the original caption but includes minor inaccuracies, redundant information, or slightly irrelevant expansions. The medical context remains intact but could be improved.
- **Examples**:
- Retains the main idea but adds unnecessary or repetitive details.
- Expansion is overly generic and lacks depth in medical relevance.

4: Good Consistency

- **Description**: The extended caption is well-aligned with the original caption, provides accurate and relevant additional details, and enhances the context without deviating from the medical focus.
- **Examples**:
- Adds valuable, medically relevant information that complements the original.
- Maintains high factual accuracy and stays within the context.

5: Excellent Consistency

- **Description**: The extended caption perfectly aligns with the original, enriching the content with precise, relevant, and insightful details. It enhances understanding without introducing errors or deviating from the topic.
- **Examples**:
- Seamlessly extends the original caption with meaningful medical context.
- Fully accurate and maintains clarity and relevance.

Figure 6. (Part 2) A detailed guideline for scoring, ranging from 1 to 5, is provided.

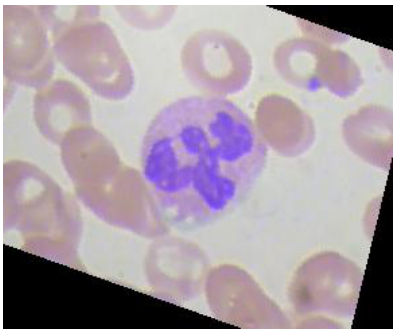


Figure 6. Q: What are the types of cells depicted in this image?
A: Neutrophils
B: Melanocytes
C: Lymphocytes
D: Hepatocytes

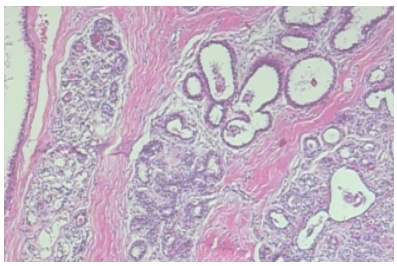


Figure 6. Q: What is the diagnosis of the histopathology in this image?
A: Breast hyperplasia without atypia histopathology
B: Normal breast histopathology
C: Benign breast histopathology
D: Fibrocystic breast histopathology

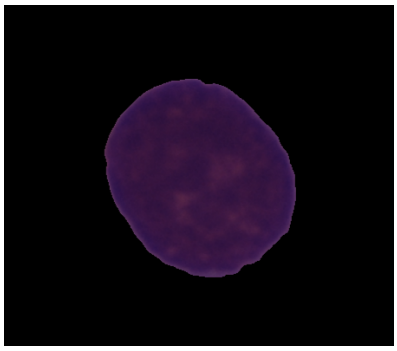


Figure 6. Q: What is the probable diagnosis depicted in this image?
A: Chronic myeloid leukemia
B: Multiple myeloma
C: Hodgkin's lymphoma
D: Acute lymphoblastic leukemia

Figure 6. Q: What is the diagnosis of the cancer seen in this image?
A: Adenocarcinoma of the right hilum, T3 N1 M0, Stage IIb
B: Mesothelioma of the right hilum, T2 N1 M0, Stage IIb
C: Large cell carcinoma of the left hilum, T2 N2 M0, Stage IIIa
D: Non-small cell carcinoma of the left hilum, T2 N0 M0, Stage I

Figure 6. Q: Is COVID-19 apparent in this CT scan image?
A: No
B: Yes

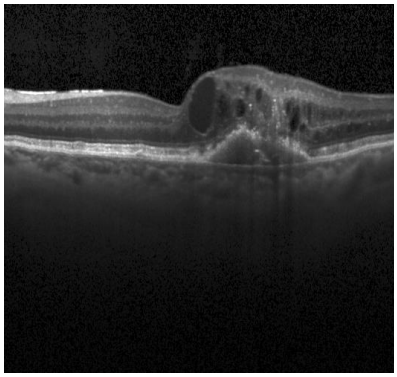


Figure 6. Q: Which imaging technique was utilized to obtain this image?
A: Ultrasound
B: Optical Coherence Tomography
C: Magnetic Resonance Imaging (MRI)
D: Thermography

Figure 7. Examples from the OmniMedVQA dataset: microscopy (top) and CT images (bottom) with corresponding questions and options, with the correct answers highlighted in blue.

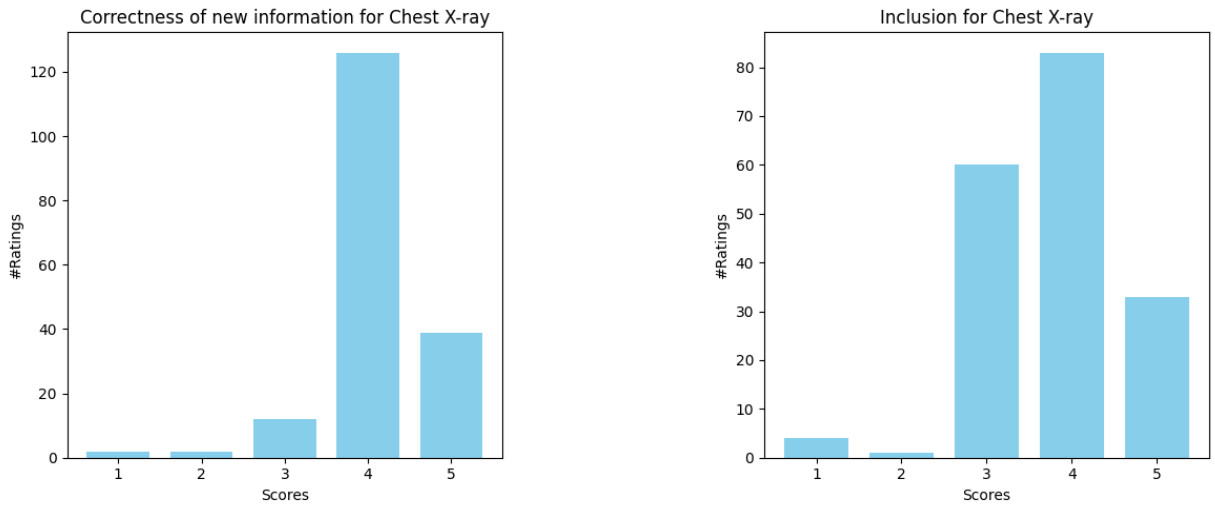


Figure 8. Statistical correctness of extended captions generated by GPT-4 on Chest X-rays.

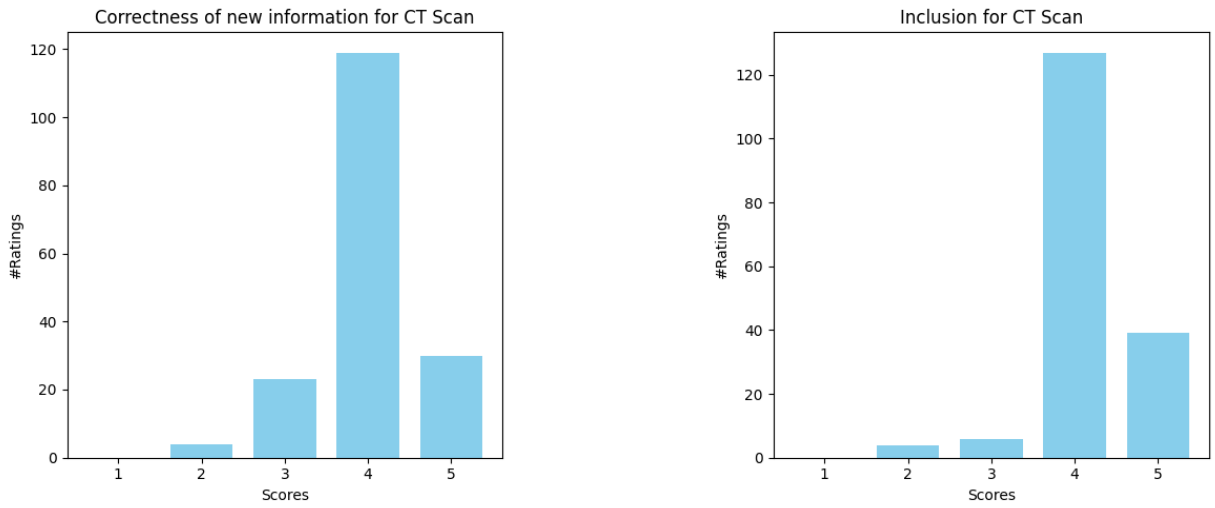


Figure 9. Statistical correctness of extended captions generated by GPT-4 on CT scans.

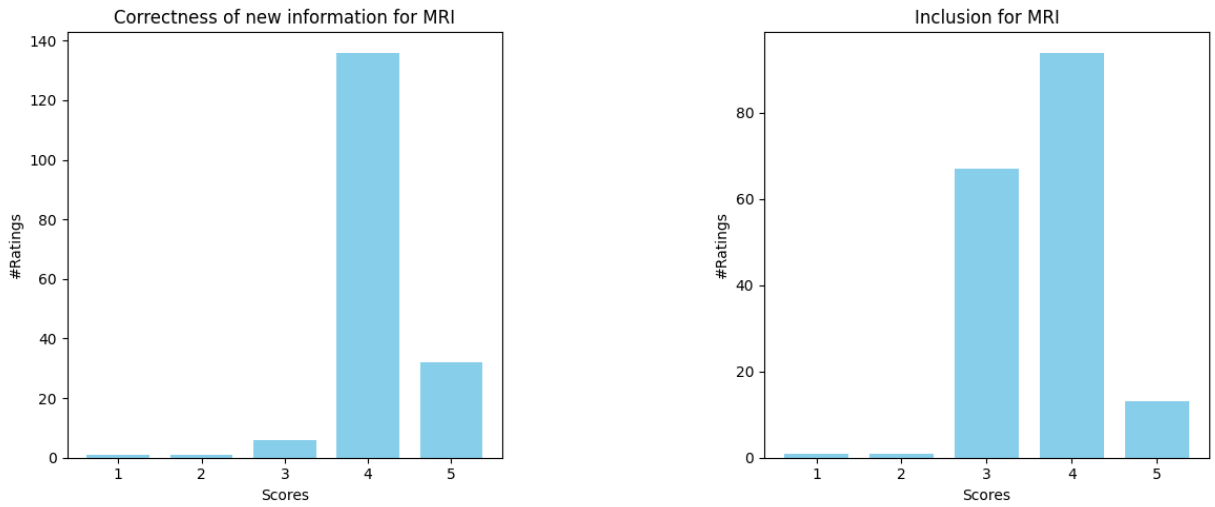


Figure 10. Statistical correctness of extended captions generated by GPT-4 on MRI.

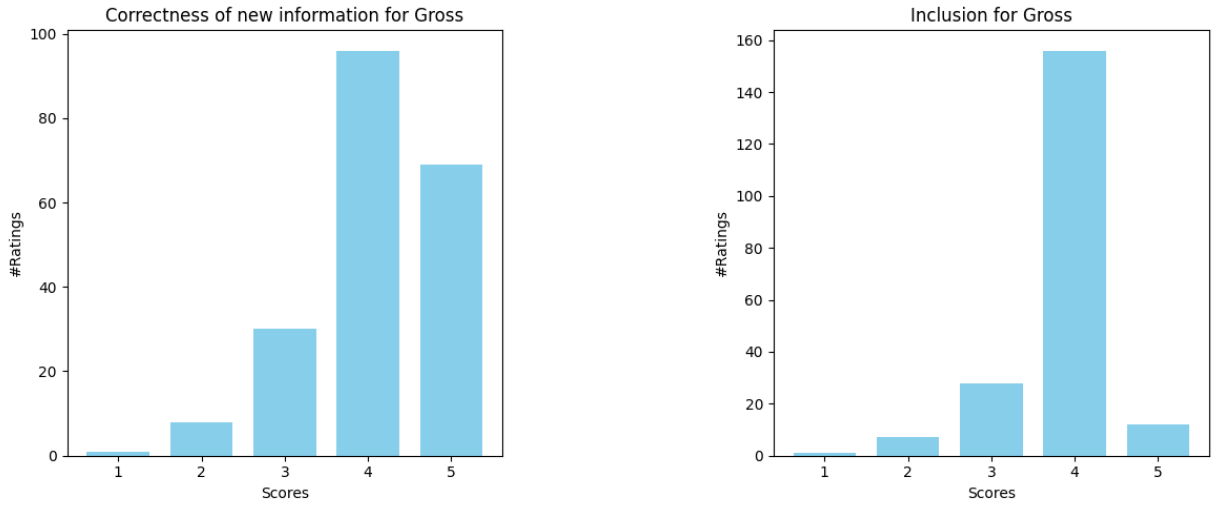


Figure 11. Statistical correctness of extended captions generated by GPT-4 on mixed domains.

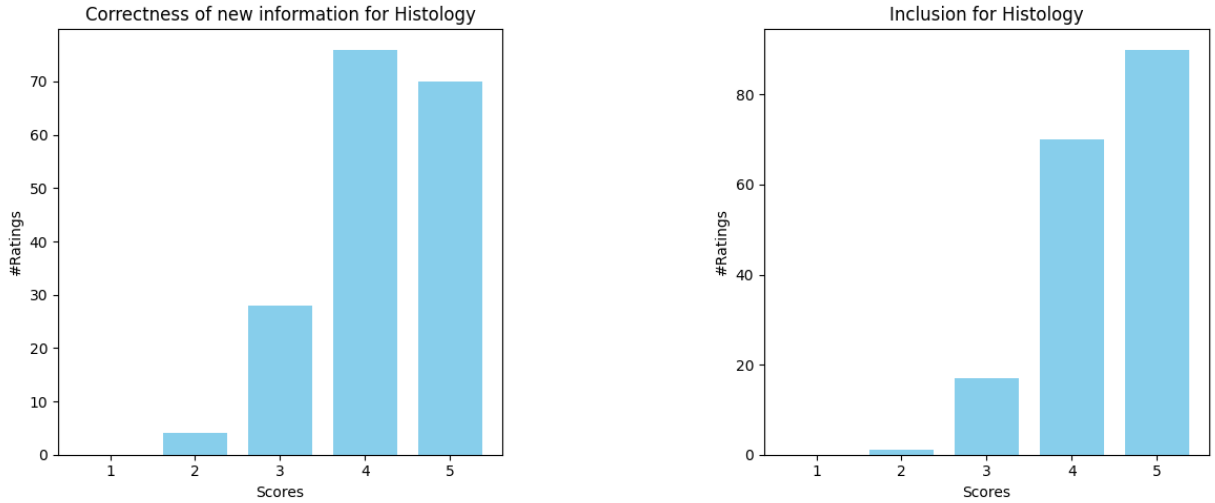


Figure 12. Statistical correctness of extended captions generated by GPT-4 on histology samples.