# LLaVA Needs More Knowledge: Retrieval Augmented Natural Language Generation with Knowledge Graph for Explaining Thoracic Pathologies

**Ameer Hamza**[1], **Abdullah**[1], **Yong Hyun Ahn**[2], **Sungyoung Lee**[1], **Seong Tae Kim**[1*]

[1]Department of Computer Science and Engineering, Kyung Hee University, Republic of Korea
[2]Department of Artificial Intelligence, Kyung Hee University, Republic of Korea
{ameer, abdullahijaz, yhahn, sylee0301, st.kim}@khu.ac.kr

## Abstract

Generating Natural Language Explanations (NLEs) for model predictions on medical images, particularly those depicting thoracic pathologies, remains a critical and challenging task. Existing methodologies often struggle due to general models' insufficient domain-specific medical knowledge and privacy concerns associated with retrieval-based augmentation techniques. To address these issues, we propose a novel Vision-Language framework augmented with a Knowledge Graph (KG)-based datastore, which enhances the model's understanding by incorporating additional domain-specific medical knowledge essential for generating accurate and informative NLEs. Our framework employs a KG-based retrieval mechanism that not only improves the precision of the generated explanations but also preserves data privacy by avoiding direct data retrieval. The KG datastore is designed as a plug-and-play module, allowing for seamless integration with various model architectures. We introduce and evaluate three distinct frameworks within this paradigm: KG-LLaVA, which integrates the pre-trained LLaVA model with KG-RAG; Med-XPT, a custom framework combining MedCLIP, a transformer-based projector, and GPT-2; and Bio-LLaVA, which adapts LLaVA by incorporating the Bio-ViT-L vision model. These frameworks are validated on the MIMIC-NLE dataset, where they achieve state-of-the-art results, underscoring the effectiveness of KG augmentation in generating high-quality NLEs for thoracic pathologies.

**Code** — https://github.com/ailab-kyunghee/KG-LLaVA
**Extended version** — https://arxiv.org/abs/2410.04749

## 1 Introduction

In recent years, natural language processing (NLP) has witnessed the development of numerous models trained on vast amounts of general domain data (Radford et al. 2019; Brown et al. 2020; Touvron et al. 2023; Jiang et al. 2023; Peng et al. 2023) While these models exhibit remarkable capabilities across various tasks, they often lack the specialized knowledge required for domain-specific applications, such as generating Natural Language Explanations (NLEs) for thoracic pathologies. This limitation is particularly pronounced in the medical domain, where accurate and contextually relevant explanations are crucial for diagnostic decision-making.

To bridge this gap, including Pre-Training, Fine-Tuning, and Retrieval-Augmented Generation (RAG) methods has been explored. One popular approach among these strategies is Pre-Training large models on medical data. However, this strategy necessitates extensive data and substantial computational resources (Saab et al. 2024; Li et al. 2024). In the medical domain, publicly available datasets are often limited, and concerns about data privacy, authenticity, and potential data leakage persist. Even when models are pre-trained on medical data, they frequently struggle with task-specific performance and may suffer from reduced factual accuracy. Moreover, pre-training these models is both expensive and time-consuming, and subsequent fine-tuning is often required to adapt them to specific downstream tasks.

Fine-tuning is another common strategy, where general domain models are adapted directly to medical tasks by training on specialized datasets. While this approach can be effective, it is hampered by the scarcity of high-quality medical datasets and the need to protect patient privacy and data security. Additionally, fine-tuned models can be prone to hallucination, generating explanations that lack factual correctness. Recent advancements in parameter-efficient fine-tuning, such as training low-rank adapters (Hu et al. 2022), have aimed to reduce computational costs while maintaining model performance, yet challenges remain, particularly in maintaining the model's generalization across diverse tasks.

The third strategy, RAG (Lewis et al. 2020), has gained traction as a method for enhancing general domain models with domain-specific knowledge dynamically. In this approach, models are fine-tuned on task-specific data while being supplemented with relevant information retrieved from a datastore. RAG methods have shown promising results in maintaining factual accuracy and reducing hallucination risks. However, the effectiveness of RAG is highly dependent on the quality of the retrieval mechanism. Also, in the medical domain, concerns about data privacy are amplified, as retrieved information might still be traceable to individual patients, even after de-identification, thus posing a risk of data leakage.

To overcome these challenges, we propose a novel approach that combines the strengths of vision-language models with a Knowledge Graph (KG)-based retrieval system.

[*]Corresponding author.

Our method, KG-based Retrieval-Augmented Generation (KG-RAG), addresses privacy risks by abstracting patient-specific details and providing models with more relevant and factual information tailored to individual cases. The KG-based datastore serves as a robust source of domain-specific knowledge, enabling the generation of accurate and contextually appropriate NLEs for thoracic pathologies.

KG-RAG emulates the cognitive process of radiologists, who rely on extensive experience and domain-specific knowledge to formulate diagnostic explanations. By leveraging a KG-based datastore tailored to each patient case, our approach significantly enhances the model's performance in generating precise and informative explanations. To demonstrate the versatility and effectiveness of our method, we integrated KG-RAG into three distinct frameworks: KG-LLaVA, Med-XPT, and Bio-LLaVA.

KG-LLaVA integrates the pre-trained LLaVA model with our KG-RAG module, fine-tuning it on our dataset to enrich its ability to generate detailed and accurate explanations by leveraging the CLIP ViT-L vision model. Med-XPT is a custom-built framework combining MedCLIP as the vision encoder, a transformer-based projector, and GPT-2 as the language model, trained from scratch on the MIMIC-NLE dataset to fully exploit the domain-specific knowledge provided by the KG-RAG module. Lastly, Bio-LLaVA adapts the LLaVA model by replacing the vision encoder with Bio-ViT-L, a model tailored for biomedical tasks, and modifying the projection layer to accommodate the unique feature dimensions of Bio-ViT-L. This framework was trained exclusively on the MIMIC-NLE dataset, showcasing its ability to generate precise NLEs without relying on pre-trained projector weights.

This integration of advanced vision-language models with a domain-specific KG not only provides transparent and comprehensible reasoning for detected abnormalities but also elevates the model's diagnostic accuracy. Our approach underscores the potential of combining state-of-the-art machine learning techniques with domain-specific knowledge to achieve expert-level reasoning, thereby improving the interpretability and accuracy of diagnostic outcomes in chest X-ray images.

Our main contributions can be summarized as follows:

- We propose the first KG retrieval-augmented Vision-Language Model (VLM) framework specifically designed for generating NLEs for thoracic pathologies. This approach integrates domain-specific medical knowledge into the explanation generation process, enhancing the accuracy and relevance of the outputs.
- Our method addresses critical privacy concerns associated with medical data by abstracting patient-specific details through the use of a KG-based datastore. Furthermore, the proposed method is designed as a plug-and-play module, making it easily adaptable to existing radiology tasks and compatible with previous methods.
- We validate the effectiveness of our approach by achieving state-of-the-art results on a benchmark dataset, MIMIC-NLE. Our method outperforms previous models, demonstrating the robustness and applicability of the KG-augmented framework in the medical domain.

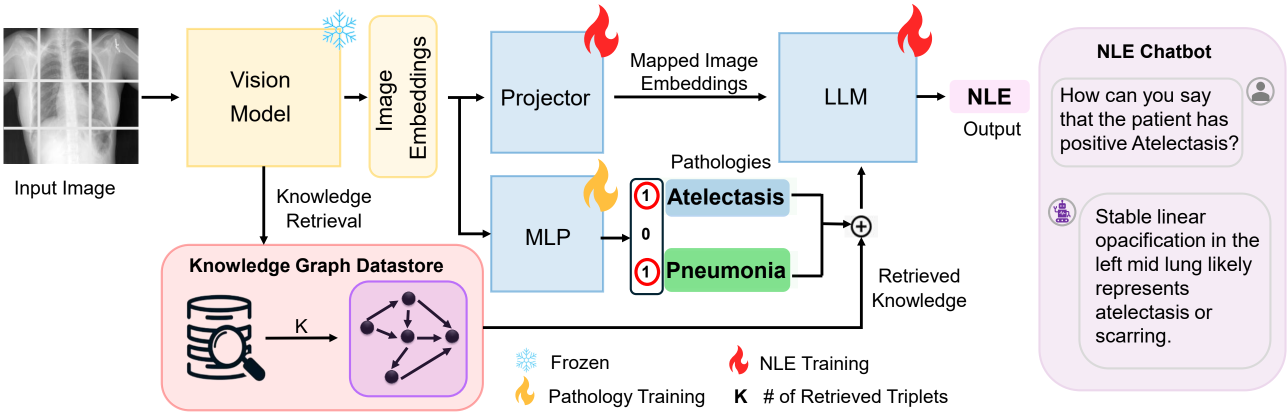
## 2 Related Work

**Natural Language Explanation.** NLEs provide textual interpretations of deep learning model predictions, aiming to offer accessible and comprehensible insights for users, particularly in complex domains like medical diagnostics. Hendricks et al. was the first to introduce the NLE task. This task was later extended to encompass the vision-language domain (Kayser et al. 2021; Li et al. 2018; Marasović et al. 2020; Park et al. 2018; Wu and Mooney 2019). Kayser et al. introduced the MIMIC-NLE dataset, derived from the MIMIC-CXR dataset (Johnson et al. 2019), to advance interpretability and accessibility in the context of chest X-ray analysis. This dataset is currently the only publicly available resource specifically designed for generating NLEs related to chest X-rays.

In their work, Kayser et al. also introduced benchmark methods for generating explanations, such as DPT (DenseNet-121 (Huang et al. 2017) combined with GPT-2 (Radford et al. 2019)) and RATCHET (Hou et al. 2021). While GPT-2 has demonstrated effective performance in general domains (Kayser et al. 2021), it shows limitations when applied to medical NLE generation due to its reliance on commonsense knowledge, which is often insufficient for specialized medical contexts. The DPT model, in particular, struggled with generating accurate explanations for chest X-ray images due to its dependency on non-specialized knowledge sources. The MIMIC-NLE dataset provides explanations for predicted pathologies, making it a comprehensive resource for evaluating the quality of NLEs in medical imaging. Our approach demonstrates a substantial improvement over previous methods, underscoring the value of integrating domain-specific knowledge through KG augmentation in the generation of NLEs. Also, Rio-Torto, Cardoso, and Teixeira have investigated parameter-efficient training techniques for the NLE task.

**Vision-Language Models.** The emergence of advanced Large Language Models (LLMs) like LLaMA (Touvron et al. 2023) and GPT-4 (Achiam et al. 2023) has showcased significant improvements in text generation capabilities. Building on these developments, researchers have increasingly focused on extending these models to handle multimodal inputs, such as visual data. Despite these efforts, fully integrating visual and textual modalities remains a challenging endeavor, particularly in areas such as understanding spatial relationships, mathematical reasoning, and counting. Bordes et al. categorize VLMs into four primary categories, contrastive training, masking, pre-trained backbones, and generative vision-language models.

VLMs in the pre-trained backbone category often utilize open-source LLMs, such as LLaMA/Viccuna, to learn mappings between a pre-trained image encoder and the LLM. This approach is computationally efficient, as it avoids training both text and image encoders from scratch, instead focusing on aligning the representations generated by these pre-trained models (Li et al. 2024, 2023; Zhu et al. 2024).



How can you say that the patient has positive Atelectasis?

Stable linear opacification in the left mid lung likely represents atelectasis or scarring.

Figure 1: Overview of the KG-LLaVA framework with integrated Knowledge Graph Retrieval Augmented Generation (KG-RAG) module. The framework combines a pre-trained LLaVA model with a CLIP ViT-L vision encoder to extract visual features, which are then projected into the language model's embedding space. The KGR module uses MedCLIP to map input images to a shared latent space and retrieve relevant KG triplets via the FAISS library. These triplets provide domain-specific context that enhances the generation of accurate and informative NLEs for thoracic pathologies. The modular design allows for seamless integration with other architectures, such as Med-XPT and Bio-LLaVA, ensuring flexibility and adaptability across different vision-language tasks.

Our work falls within the category of VLMs based on pre-trained backbones. Specifically, we employ an LLaVA (Li et al. 2024) model which is based on pre-trained LLM alongside a pre-trained vision model, connected via a Projector that learns to map the visual information to the language model space. This approach leverages the strengths of pre-trained vision and language models.

Visual Instruction Tuning has been effectively employed in models like LLAVA (Li et al. 2024), where a pre-trained LLM is trained on visual inputs provided by a vision model using carefully curated instruction datasets. This process typically involves mapping image features from pre-trained models like CLIP into the LLM's embedding space. While more sophisticated projection methods, such as lightweight transformers in CLIPCap (Mokady, Hertz, and Bermano 2021) and the Bert-based Q-Former in BLIP-2 (Li et al. 2023), have been explored, the linear layer approach remains a popular choice due to its simplicity and efficiency.

**Knowledge Graph.** A KG represents relationships between a set of entities, offering a structured approach to capturing and utilizing interconnected information. In addition to KG-TREAT (Liu, Wu, and Zhang 2024) and MKG-FENN (Wu et al. 2024), recent works such as ImgFact (Liu et al. 2024) and VCTP (Chen et al. 2024) have pushed the boundaries of KG integration. ImgFact introduces a large-scale multimodal KG with triplet fact grounding to improve visual reasoning tasks, while VCTP applies visual chain-of-thought prompting to facilitate knowledge-based visual reasoning. In the radiology domain, the RadGraph (Jain et al. 2021) method introduces a pioneering approach to constructing KGs from medical reports, providing a systematic means of extracting and representing clinically relevant information. RadGraph not only defines the methodology for creating these graphs but also supplies a publicly available dataset, which significantly aids research and development in this area.

Unlike these methods, our privacy-preserving KG-RAG framework focuses on medical NLE tasks. By leveraging clinically relevant relationships (e.g., 'suggestive_of') and ensuring data privacy through de-identified triplets for NLE generation.

**Retrieval Augmented Generation.** RAG (Lewis et al. 2020) represents a significant advancement in enhancing language models beyond the capabilities achieved through traditional supervised fine-tuning. In RAG, external knowledge is dynamically retrieved and integrated into the language model, thereby augmenting its generative capabilities with new information that the model has not been explicitly trained on. This approach is particularly valuable when addressing dynamic or domain-specific knowledge, such as medical information, where up-to-date and specialized content is crucial. However, as identified by (Zeng et al. 2024), the use of RAG in sensitive domains like healthcare raises substantial security concerns. Specifically, there is a risk that malicious actors could exploit RAG systems by prompting the model to retrieve and expose sensitive information from the retrieval datastore, potentially leading to privacy breaches. This vulnerability underscores the necessity for more secure implementations of RAG, particularly in the context of handling confidential medical data.

## 3 Methodology

In this paper, we present a novel approach for generating NLEs via KG-RAG for thoracic pathologies. Our methodology utilizes medical vision models in conjunction with large language models, resulting in three distinct frameworks: KG-LLaVA, Med-XPT, and Bio-LLaVA, as illustrated in Figure 1. Each framework is designed to leverage the strengths of KG-RAG in enhancing the accuracy and contextual relevance of NLEs in the medical domain.

### Pathology Classification

For pathology classification, the input X-ray image $X \in \mathbb{R}^{H\times W\times C}$ is processed by a visual encoder, such as MedCLIP or ViT-L/14, to generate visual feature embeddings:

$$Z_v = V(X), \tag{1}$$

where $V(X)$ represents the visual encoder that maps $X$ to a high-dimensional feature vector $Z_v \in \mathbb{R}^{d_v}$.

The extracted visual features $Z_v$ are then passed through a multi-layer perceptron (MLP) to classify pathologies:

$$\hat{y} = \text{MLP}(Z_v), \tag{2}$$

where $\hat{y} \in \mathbb{R}^p$ denotes the predicted probabilities for $p$ pathology classes.

In addition to predicting the associated certainty levels along with the identified pathologies in X-ray images, we adopted a methodology based on the approach outlined by (Kayser et al. 2022). This method involves the prediction of 10 distinct pathologies, each categorized into three certainty levels—negative, uncertain, or positive—using the UMultiClass strategy introduced by (Irvin et al. 2019).

The UMultiClass strategy applies a classification function $f$ to map each pathology prediction $\hat{y}_i$ to a certainty level $c_i \in \{\text{negative, uncertain, positive}\}$ based on predefined thresholds:

$$c_i = f(\hat{y}_i) = \begin{cases} \text{negative} & \text{if } \hat{y}_i < \theta_{\text{neg}}, \\ \text{uncertain} & \text{if } \theta_{\text{neg}} \le \hat{y}_i < \theta_{\text{pos}}, \\ \text{positive} & \text{if } \hat{y}_i \ge \theta_{\text{pos}}, \end{cases} \tag{3}$$

where $\theta_{\text{neg}}$ and $\theta_{\text{pos}}$ are the thresholds for classifying certainty levels.

We process the visual features $Z_v$ extracted from the medical vision model through the MLP to enhance the model's capability to interpret and classify the visual features of the X-ray images. This process improves the accuracy of both the pathology predictions $\hat{y}$ and their corresponding certainty levels $c_i$.

### Knowledge Graph Retrieval

The Knowledge Graph (KG) is represented as a collection of triplets $(e_i, r_{ij}, e_j)$, where $e_i$ and $e_j$ are medical entities, and $r_{ij}$ is the relationship between them extracted from clinical reports:

$$\text{KG} = \{(e_i, r_{ij}, e_j) \mid e_i, e_j \in \mathcal{E}, r_{ij} \in \mathcal{R}\}. \tag{4}$$

Each triplet is embedded into a high-dimensional feature space using the MedCLIP text encoder $f_{KG}$:

$$Z_k = f_{KG}(e_i, r_{ij}, e_j), \tag{5}$$

where $Z_k \in \mathbb{R}^{d_k}$ represents the embedding of the triplet $(e_i, r_{ij}, e_j)$.

To retrieve the most relevant triplets for a given query image $X$, the image is first processed by the visual encoder to obtain the image embedding (refer to Equation 1).

The retrieval process is conducted by computing the cosine similarity between the image embedding $Z_v$ and each triplet embedding $Z_k$ in the datastore:

$$\text{sim}(Z_v, Z_k) = \frac{Z_v \cdot Z_k}{\|Z_v\|\|Z_k\|}. \tag{6}$$

The top-$k$ most relevant triplets $T_{\text{retrieved}}$ are then selected based on the highest similarity scores:

$$T_{\text{retrieved}} = \arg\max_{Z_k \in \text{KG}} \text{sim}(Z_v, Z_k). \tag{7}$$

To address the privacy risks highlighted in Section 2, we propose a KG-based RAG approach as a secure alternative. Unlike traditional RAG systems, which may retrieve information traceable to specific patients, our approach employs a KG composed of general medical terms, entities, and relationships. This structured representation of knowledge graph abstracts away direct patient-specific details, significantly reducing the risk of inadvertently exposing sensitive information.

In our framework, we augment the model with knowledge instances retrieved from a constructed datastore. This method enhances the model's performance by providing relevant contextual knowledge while protecting sensitive medical data. To enable effective knowledge retrieval, we constructed a datastore comprising KG triplets derived exclusively from the MIMIC-CXR training set (Johnson et al. 2019) using the RadGraph model (Jain et al. 2021). These triplets are aligned with the MIMIC-NLE (Kayser et al. 2022) training set, ensuring no overlap with the dev and test sets, thereby preventing data leakage. We focus specifically on triplets with the 'suggestive_of' relationship (e.g., "opacity suggestive_of pneumonia") as these triplets are more directly relevant to explaining the presence of pathologies. These triplet embeddings were generated using a medical CLIP model and are exclusively stored in the datastore, deliberately excluding any image features. This design facilitates a cross-modal retrieval process, where images can be used to query and retrieve relevant KG triplets.

KG-RAG retrieves contextually relevant knowledge by focusing on semantically similar information rather than exact matches. Even when minor deviations occur, the Vision-Language Model (VLM) leverages visual cues to maintain accurate pathology identification.

## 4 Experiment

### Dataset

For our study, we utilized the MIMIC-NLE dataset (Kayser et al. 2022), which is derived from MIMIC-CXR chest X-ray dataset (Johnson et al. 2019). This is currently the only publically available dataset for chest X-ray NLEs. The MIMIC-NLE dataset includes diagnoses, evidence labels, and corresponding NLEs for those diagnoses. For a detailed description of the dataset, please refer to the comprehensive overview provided in (Kayser et al. 2022). The dataset consists of 38,003 NLEs and is divided into training, validation, and testing subsets, containing 37,016, 273, and 714 entries, respectively.

The visual instruction tuning approach is employed in both KG-LLaVA and Bio-LLaVA, where it enhances the model's instruction-following abilities and generalization performance across various medical imaging tasks. However, instruction tuning was not applied to the Med-XPT framework, which focuses instead on leveraging its custom architecture for generating NLEs.

For a comprehensive description of the instruction-format dataset and the tailored prompts for pathology explanations, please refer to Appendix A in the supplementary material.

## Implementation Details

**Projector.** The GPT-2 based model uses a transformer-based projector, while the LLaMA/Vicuna-based model uses an MLP-based projector. The projector aligns the image embeddings with the features required by the Language Model (LM). Specifically, the visual features $Z_v$ are projected into the language model's embedding space using a trainable linear layer:

$$H_v = W_p \cdot Z_v + b_p, \tag{8}$$

where $W_p \in \mathbb{R}^{d_l \times d_v}$ is the weight matrix, $b_p \in \mathbb{R}^{d_l}$ is the bias term, and $H_v \in \mathbb{R}^{d_l}$ represents the projected visual features aligned with the language model's embedding space.

**Vision Models.** In our proposed framework, we employ two distinct vision models tailored for different components of the system. First, we utilize the MedCLIP model (Wang et al. 2022), which serves dual purposes. It is integrated with the DPT (Kayser et al. 2022) framework for NLE generation and is also instrumental in constructing the KGR datastore. MedCLIP's robust capabilities enable effective retrieval of relevant information based on image features, ensuring that the retrieved knowledge aligns with the visual content of medical images.

For the KGR process within our LLaVA-based framework, we again utilize MedCLIP to perform the retrieval itself based on the image features. However, for the extraction of visual features and their subsequent projection into the language embedding space, we employ the ViT-L/14 (Radford et al. 2021) CLIP model as the vision encoder. The ViT-L CLIP model extracts visual features from input images (refer to Equation 1).

These visual features are then projected into the language model's embedding space through a trainable projection matrix (refer to Equation 8). This simple linear layer facilitates seamless integration with the language model. This dual-model approach allows us to leverage the specific strengths of both MedCLIP and ViT-L/14 within the framework, optimizing the generation of NLEs and ensuring accurate and contextually relevant retrieval of knowledge.

**Language Models.** For the decoding mechanism, we integrated language models such as GPT-2 (Radford et al. 2019) and LLaMA/Viccuna (Touvron et al. 2023; Peng et al. 2023), known for their effectiveness in natural language generation (NLG) tasks. The language model $LM$ generates a probability distribution over target tokens $Y = \{y_1, y_2, \ldots, y_n\}$ conditioned on the input sequence $S$, which includes projected visual features, pathology labels, and retrieved knowledge triplets:

$$P(Y \mid S) = \prod_{i=1}^{n} P(y_i \mid S, y_1, y_2, \ldots, y_{i-1}). \tag{9}$$

This structured prompt ensures that the generated NLEs are conditioned on relevant image features, pathology information, and retrieved knowledge. This process effectively combines visual data and textual information to produce accurate and contextually relevant NLEs.

## Training

Our proposed frameworks, KG-LLaVA, Med-XPT, and Bio-LLaVA, each incorporate a KGR module to enhance their performance in generating NLEs for thoracic pathologies. KG-LLaVA builds upon the pre-trained LLaVA model, integrating the KGR module to leverage domain-specific knowledge derived from the input image. Med-XPT is trained from scratch, combining MedCLIP as the vision encoder with a transformer-based projector and GPT-2 as the language model, while incorporating the KG module to enrich the generation process. Bio-LLaVA adapts the LLaVA model by replacing the vision encoder with Bio-ViT-L and modifying the projection layer to accommodate the unique feature dimensions, also integrating the KG module to improve model performance. These frameworks were trained on the MIMIC-NLE dataset, allowing us to evaluate the effectiveness of the KGR module across different architectures.

Refer to Appendix B in the supplementary material for detailed implementation and training configurations of KG-LLaVA, Med-XPT, and Bio-LLaVA.

## Evaluation Metrics

In line with (Kayser et al. 2022), we evaluate NLEs only for correctly predicted labels. Prior research has indicated that standard automated natural language generation (NLG) metrics often fall short in assessing NLE quality due to the variability in expressing similar meanings using different syntactic structures and semantic interpretations (Kayser et al. 2021). For our evaluation, we report the widely used NLG metrics: BLEU, ROUGE, CIDEr, and METEOR (Kayser et al. 2021).

Among these, CIDEr is particularly relevant as it measures the n-gram overlap with ground-truth explanations, placing emphasis on contextual relevance. By retrieving pathology-specific triplets through KG-RAG, our approach increases the likelihood of incorporating relevant medical terms into the generated NLEs. This contributes to the higher CIDEr scores observed in our results, highlighting the effectiveness of KG-RAG in enhancing the contextual accuracy of explanations.

| Method | AUC | B4 | MET. | R.L. | CIDEr |
|---|---|---|---|---|---|
| RATCHET (Hou et al. 2021) | 66.4 | 4.7 | 14.1 | 22.2 | 37.9 |
| TieNet (Wang et al. 2018) | 64.6 | 3.5 | 12.4 | 19.4 | 33.9 |
| DPT (Kayser et al. 2022) | 62.5 | 2.4 | 11.3 | 15.4 | 17.4 |
| LoRA AE (Rio-Torto, Cardoso, and Teixeira 2024) | 63.9 | 4.0 | **15.3** | 20.6 | 24.4 |
| Prompt AE (Rio-Torto, Cardoso, and Teixeira 2024) | 61.3 | 3.7 | 14.4 | 19.7 | 23.4 |
| Prefix AE (Rio-Torto, Cardoso, and Teixeira 2024) | 65.2 | 3.7 | 14.7 | 19.7 | 21.5 |
| LLaMA-Adapt AE (Rio-Torto, Cardoso, and Teixeira 2024) | 63.9 | 4.3 | 14.6 | 21.4 | 29.7 |
| + multi-modal (Rio-Torto, Cardoso, and Teixeira 2024) | 64.9 | 3.0 | 14.1 | 18.6 | 18.4 |
| + MSE loss (Rio-Torto, Cardoso, and Teixeira 2024) | 62.3 | 2.0 | 10.0 | 14.2 | 14.2 |
| KG-LLaVA | **83.0** | **7.2** | 15.1 | **25.0** | **62.2** |

Table 1: Comparison of our KG-LLaVA with other baselines on the MIMIC-NLE test set, focusing on NLG metrics. The metrics include Area Under the Curve (AUC), BLEU-4 (B4), METEOR (MET.), ROUGE-L (R.L.), and CIDEr scores.

| Method | B4 | MET. | R.L. | CIDEr |
|---|---|---|---|---|
| Bio-LLaVA | 5.7 | 14.3 | 23.0 | 46.7 |
| Med-XPT | 7.0 | 11.0 | 22.9 | **62.7** |
| KG-LLaVA | **7.2** | **15.1** | **25.0** | 62.2 |

Table 2: Performance comparison of our proposed frameworks—Bio-LLaVA, Med-XPT, and KG-LLaVA—on the MIMIC-NLE test set, focusing on NLG metrics. All frameworks incorporate KG-RAG module. Evaluation metrics include BLEU-4 (B4), METEOR (MET.), ROUGE-L (R.L.), and CIDEr, scores.

## 5 Results and Discussions

### Comparison with Other Methods

In this study, we evaluated the performance of our proposed KG-LLaVA framework against several well-established methods, including RATCHET (Hou et al. 2021), TieNet (Wang et al. 2018), and DPT (Kayser et al. 2022), using the MIMIC-NLE (Kayser et al. 2022) dataset. The results, as summarized in Table 1, clearly demonstrate that KG-LLaVA outperforms the previous methods across a range of evaluation metrics.

KG-LLaVA achieves an AUC of 83.0, which is significantly higher than the AUC scores reported for RATCHET (66.4), TieNet (64.6), and DPT (62.5). This substantial improvement underscores the effectiveness of our approach in accurately classifying and generating relevant explanations for thoracic pathologies. Additionally, KG-LLaVA excels in key NLG metrics, including BLEU-4 (7.2), ROUGE-L (25.0) and CIDEr (62.2), highlighting its ability to generate high-quality, contextually accurate explanations.

Notably, while KG-LLaVA slightly outperforms RATCHET (Hou et al. 2021) in METEOR (15.1 vs. 14.1), the overall superior performance of KG-LLaVA across the other metrics underscores the strength of incorporating KG-RAG module into a vision-language framework. These results suggest that KG-LLaVA has the potential to set a new benchmark for generating NLEs in medical imaging tasks, particularly for diagnosing thoracic pathologies.

| Methods | RAG | B4 | METEOR | R-L | CIDEr |
|---|---|---|---|---|---|
| Med-XPT | - | 2.0 | 7.8 | 12.8 | 17.4 |
| KG-LLaVA | - | 7.0 | 15.0 | 24.4 | 60.1 |
| Med-XPT | NLE | 6.7 | 13.5 | 22.2 | 59.3 |
| KG-LLaVA | NLE | 6.8 | 15.0 | 24.6 | 58.8 |
| Med-XPT | KG | 7.0 | 11.0 | 22.9 | **62.7** |
| KG-LLaVA | KG | **7.2** | **15.1** | **25.0** | 62.2 |

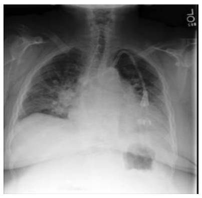
Table 3: Comparative analysis of the performance of Med-XPT, and KG-LLaVA across different RAG methods and without any RAG. The table includes results for NLG metrics such as BLEU-4 (B4), METEOR, ROUGE-L (R-L), and CIDEr. The ”-” row shows results without any RAG integration, the ”NLE” row represents results with natural language explanation-based RAG, and the ”KG” row reflects the performance when the knowledge graph retrieval module is used.

While (Rio-Torto, Cardoso, and Teixeira 2024) methods focused on optimizing model parameters, our approach leveraged the KG-RAG module and effective use of LoRA for fine-tuning, achieving superior performance without compromising the model's complexity or parameter efficiency. This makes KG-LLaVA not only the best-performing model but also a robust and scalable solution for medical NLE generation.

### Comparison of Different LLMs

We further assessed the performance of our three proposed frameworks—Bio-LLaVA, Med-XPT, and KG-LLaVA all of which incorporate the KG-RAG module. The results, detailed in Table 2, provide insights into the comparative strengths of each framework in generating NLEs for thoracic pathologies.

KG-LLaVA demonstrates the highest overall performance, leading in BLEU-4 (7.2), METEOR (15.1), and ROUGE-L (25.0). These results reflect its superior ability to generate accurate and contextually rich explanations. Med-XPT, on the other hand, performs exceptionally well in CIDEr (62.7), indicating its effectiveness in capturing



**Ground Truth (GT):** An underlying infectious infiltrate cannot be excluded.

**KG-LLaVA:** An underlying infectious infiltrate cannot be excluded.

**Med-XPT:** Right lower lobe opacity is concerning for consolidation.

**Bio-LLaVA:** There is a new right lower lobe opacity which could be due to aspiration or pneumonia.

Figure 2: Comparison of NLEs generated by different models—KG-LLaVA, Med-XPT, and Bio-LLaVA—against the ground truth (GT) for a specific thoracic pathology case. The image depicts a chest X-ray used as input, with the corresponding NLEs. KG-LLaVA accurately matches the GT by identifying the underlying abnormalities, while Bio-LLaVA and Med-XPT offer alternative interpretations, reflecting the models' varying strengths and limitations in clinical reasoning.

the diversity and richness of language necessary for high-quality NLEs. Bio-LLaVA, while slightly behind in some metrics, still shows strong performance in METEOR (14.3) and ROUGE-L (23.0). These findings underscore the flexibility and efficacy of the KG-RAG module across different architectures. The variability in performance across the frameworks suggests that the choice of architecture can be optimized based on specific aspects of the NLE task, such as accuracy, linguistic richness, or diversity in the generated explanations.

### Impact of Different RAG Methods

We conducted a detailed comparison of the two frameworks—Med-XPT and KG-LLaVA—across various configurations: without any Retrieval Augmented Generation (RAG), with standard NLE, and with our proposed KG Retrieval module. The results, as shown in Table 3, illustrate the impact of different RAG methods on the performance of these frameworks in generating accurate and contextually rich NLEs for thoracic pathologies.

In the NLE configuration, where standard NLEs are generated without KG enhancement, both Med-XPT and KG-LLaVA exhibit strong performance, with KG-LLaVA slightly leading in most metrics. This indicates that while both frameworks leverage their respective architectures effectively, the pre-training knowledge embedded in KG-LLaVA likely contributes to its superior performance.

Additional discussion is provided in Appendix C of the supplementary material.

### Role of Instructional Prompts

An ablation study on the role of instructional prompts is available in Appendix D.

### Qualitative Results

The qualitative analysis of the generated NLEs from our proposed frameworks—KG-LLaVA, Bio-LLaVA, and Med-XPT—highlights distinct differences in their alignment with the ground truth (GT) as shown in Figure 2. KG-LLaVA accurately replicates the GT by identifying the underlying infectious infiltrate, showcasing its strong alignment with expert annotations. In contrast, Bio-LLaVA introduces an alternative diagnosis, suggesting a new right lower lobe opacity possibly due to aspiration or pneumonia, which, while clinically plausible, diverges from the GT. Med-XPT incorrectly focuses on a right lower lobe opacity concerning consolidation, indicating challenges in precise localization and consistency. These findings underscore KG-LLaVA's effectiveness in generating accurate NLEs, while also illustrating the flexibility and limitations of Bio-LLaVA and Med-XPT in clinical interpretation.

Overall, these results highlight the significant impact of the KG-RAG module on improving model performance across different architectures. KG-LLaVA consistently shows strong results across all configurations, underscoring its potential as a leading framework for generating precise and contextually relevant NLEs in the medical imaging domain. The findings also suggest that the choice of RAG method plays a crucial role in determining the quality of NLEs, with KG-RAG offering the most substantial benefits including data security.

### Analysis of Failure Cases

A detailed discussion of failure cases is provided in Appendix E.

### Limitation

Our framework requires a KG model like RadGraph for constructing triplets. For new modalities, practitioners may need to adapt or train alternative graph-processing models. While our design is intended for research purposes, further clinical validation is necessary before real-world deployment.

## 6 Conclusion

In this paper, we introduced a novel approach for generating NLEs for thoracic pathologies by integrating the KG-RAG module into vision-language models. Our KG-RAG effectively enhances the accuracy and contextual relevance of NLEs by incorporating domain-specific knowledge. Evaluated across three distinct frameworks—KG-LLaVA, Med-XPT, and Bio-LLaVA—our method consistently outperformed established models like RATCHET, TieNet, and DPT on the MIMIC-NLE dataset, highlighting the robustness and versatility of the KG-RAG approach.

Moreover, the inclusion of the KG-RAG module addresses critical privacy concerns by abstracting patient-specific details, thereby safeguarding data security and preventing data leakage. These findings underscore the critical role of integrating domain-specific knowledge in advancing vision-language models for medical imaging while ensuring the security and privacy of sensitive medical data. This approach sets a new benchmark for AI-driven diagnostics, paving the way for more transparent, accurate, and trustworthy healthcare systems.

## Acknowledgments

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) under Grant 2022-0-00078 (Explainable Logical Reasoning for Medical Knowledge Generation), Grant IITP-2024-RS-2023-00258649, Grant RS-2022-00155911, Grant 2021-0-02068 and by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant RS-2024-00334321.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bordes, F.; Pang, R. Y.; Ajay, A.; Li, A. C.; Bardes, A.; Petryk, S.; Mañas, O.; Lin, Z.; Mahmoud, A.; Jayaraman, B.; et al. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Sun, Z.; Gutfreund, D.; and Gan, C. 2024. Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2): 1254–1262.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 3–19. Springer.

Hou, B.; Kaissis, G.; Summers, R. M.; and Kainz, B. 2021. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *MICCAI*, 293–303.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. Lora: Low-rank adaptation of large language models. *ICLR*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, 590–597.

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

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv:1901.07042*.

Kayser, M.; Camburu, O.-M.; Salewski, L.; Emde, C.; Do, V.; Akata, Z.; and Lukasiewicz, T. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1244–1254.

Kayser, M.; Emde, C.; Camburu, O.-M.; Parsons, G.; Papiez, B.; and Lukasiewicz, T. 2022. Explaining chest x-ray pathologies in natural language. In *MICCAI*, 701–713.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

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

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

Li, Q.; Tao, Q.; Joty, S.; Cai, J.; and Luo, J. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 552–567.

Liu, J.; Zhang, M.; Li, W.; Wang, C.; Li, S.; Jiang, H.; Jiang, S.; Xiao, Y.; and Chen, Y. 2024. Beyond Entities: A Large-Scale Multi-Modal Knowledge Graph with Triplet Fact Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 18653–18661.

Liu, R.; Wu, L.; and Zhang, P. 2024. KG-TREAT: Pre-training for Treatment Effect Estimation by Synergizing Patient Data with Knowledge Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8805–8814.

Marasović, A.; Bhagavatula, C.; Park, J. S.; Bras, R. L.; Smith, N. A.; and Choi, Y. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. *Findings of EMNLP*.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8779–8788.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rio-Torto, I.; Cardoso, J. S.; and Teixeira, L. F. 2024. Parameter-Efficient Generation of Natural Language Explanations for Chest X-ray Classification. In *Medical Imaging with Deep Learning*.

Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

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

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; and Summers, R. M. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, 9049–9058.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. *EMNLP*.

Wu, D.; Sun, W.; He, Y.; Chen, Z.; and Luo, X. 2024. MKG-FENN: A Multimodal Knowledge Graph Fused End-to-End Neural Network for Accurate Drug–Drug Interaction Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9): 10216–10224.

Wu, J.; and Mooney, R. J. 2019. Faithful multimodal explanation for visual question answering. *ACL BlackboxNLP workshop*.

Zeng, S.; Zhang, J.; He, P.; Xing, Y.; Liu, Y.; Xu, H.; Ren, J.; Wang, S.; Yin, D.; Chang, Y.; et al. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ICLR*.