

# Pretraining Vision-Language Model for Difference Visual Question Answering in Longitudinal Chest X-rays

Yeongjae Cho<sup>\*†1</sup>

Taehee Kim<sup>\*2</sup>

Heejun Shin<sup>2</sup>

Sungzoon Cho<sup>‡1</sup>

Dongmyung Shin<sup>‡2</sup>

YJCHO@BDAL.SNU.AC.KR

TAEHEE@RADISENTECH.COM

SHJ4901@RADISENTECH.COM

ZOON@SNU.AC.KR

SHINSAE11@RADISENTECH.COM

<sup>1</sup> *Seoul National University*

<sup>2</sup> *Artificial Intelligence Engineering Division, Radisen Co. Ltd.*

## Abstract

Difference visual question answering (diff-VQA) is a challenging task that requires answering complex questions based on differences between a pair of images. This task is particularly important in reading chest X-ray images because radiologists often compare multiple images of the same patient taken at different times to track disease progression and changes in its severity in their clinical practice. However, previous works focused on designing specific network architectures for the diff-VQA task, missing opportunities to enhance the model’s performance using a pretrained vision-language model (VLM). Here, we introduce a novel VLM called PLURAL, which is pretrained on natural and longitudinal chest X-ray data for the diff-VQA task. The model is developed using a step-by-step approach, starting with being pretrained on natural images and texts, followed by being trained using longitudinal chest X-ray data. The longitudinal data consist of pairs of X-ray images, along with question-answer sets and radiologist’s reports that describe the changes in lung abnormalities and diseases over time. Our experimental results show that the PLURAL model outperforms state-of-the-art methods not only in diff-VQA for longitudinal X-rays but also in conventional VQA for a single X-ray image. Through extensive experiments, we demonstrate the effectiveness of the proposed VLM architecture and pretraining method in improving the model’s performance. Our code is available at: <https://github.com/yjch00/PLURAL>

**Keywords:** Vision-Language Model, Visual Question Answering, Chest X-ray, Longitudinal Data, Pretraining

## 1. Introduction

Assessing longitudinal medical images is an essential daily task for radiologists. It involves reading and comparing multiple images of the same patient taken at different times. The purpose of this assessment is to track disease progression and changes in its severity. For example, radiologists review longitudinal CT images to determine whether a patient has any malignant nodules with increased sizes, which is a common indication of lung cancer (Callister et al., 2015). In the case of tuberculosis, they regularly compare longitudinal

---

\* Contributed equally

† Work done while interning at Radisen Co. Ltd.

‡ Corresponding author

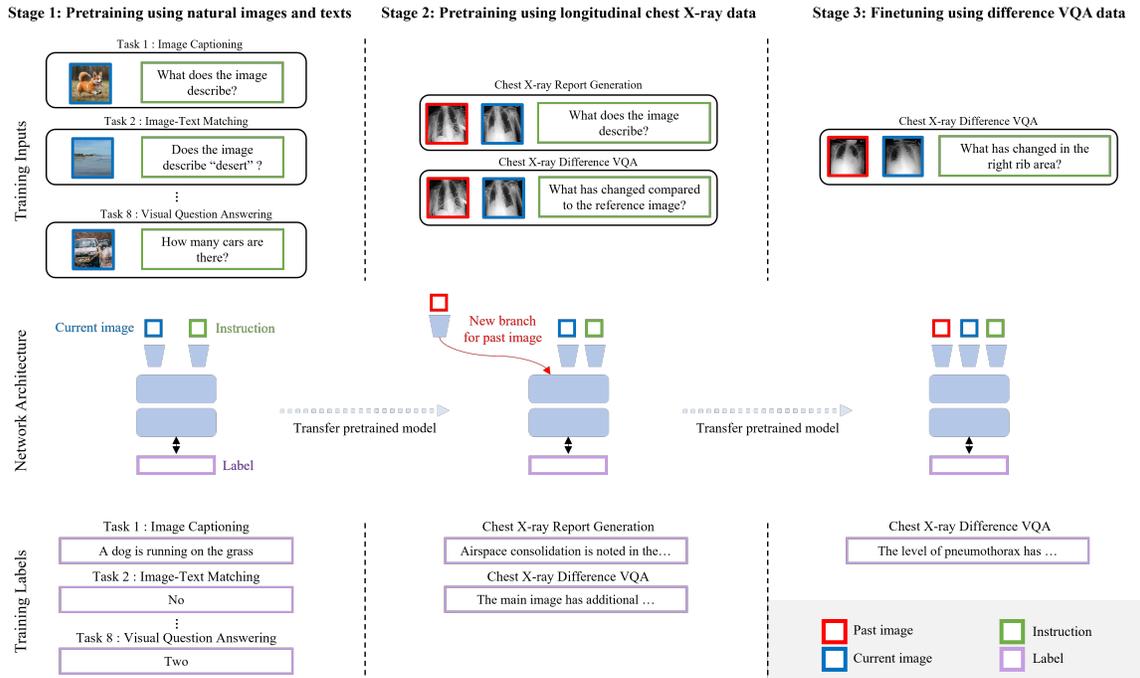


Figure 1: The training process of the PLURAL model. We adopted a Transformer-based network that takes a single image and an instruction as an input (see blue and green boxes) in the first stage. In the second and third stages, we added a new input branch for a past image (see red boxes) to utilize longitudinal chest X-ray data, including chest X-ray reports and difference VQA data.

chest X-ray images to monitor the effect of treatment and check for any improvement in pulmonary findings (Lee et al., 2020).

Visual question answering (VQA) is a challenging task that involves answering a set of complex questions about an image. With the increasing popularity of vision-language models (VLMs) (Alayrac et al., 2022; OpenAI, 2023; Liu et al., 2023), VQA is gaining attention as an effective method to interpret medical images (Lin et al., 2023), assist visually impaired individuals (Bigham et al., 2010), and so on. In particular, a recent study (Hu et al., 2023) has attempted to solve a VQA task based on longitudinal images, detecting the changes between two X-ray images taken at different time points. This task, called difference VQA (diff-VQA), is especially helpful in assisting radiologists in their clinical practice.

In previous works (Qiu et al., 2021; Yao et al., 2022; Hu et al., 2023), specific network architectures were designed to solve the diff-VQA task effectively because this task requires an AI model to differentiate between common and distinguished features in two images. However, customizing neural networks has caused them to miss valuable opportunities to enhance the model’s performance by utilizing VLMs pretrained on a vast number of texts and images, in general, (Chen et al., 2022; Wang et al., 2022; Bao et al., 2022) and/or medical domains (Zhang et al., 2023; Moor et al., 2023; Li et al., 2023). As a result, no studies have proposed an effective pretraining pipeline for diff-VQA or investigated the impact of pretraining settings, such as pretraining steps and data composition.

In this study, we introduce a novel VLM called PLURAL, which stands for pretrained vision-language model based on natural and longitudinal chest X-ray data for diff-VQA. The training process of this model involves using a Transformer-based network (Wang et al., 2022) that has been pretrained on natural images and texts (see Stage 1 in Figure 1) and then optimizing it using longitudinal chest X-ray data (see Stage 2 and 3 in Figure 1). The longitudinal data consists of pairs of longitudinal X-rays, QA sets, as well as X-ray reports that provide detailed descriptions of the changes in lung abnormalities and diseases over time. This step-by-step training approach allows the PLURAL model to benefit from the transferred knowledge of a pretrained VLM, while also taking advantage of the rich temporal information in the X-ray reports.

We have demonstrated the effectiveness of the PLURAL method by comparing its performance in the diff-VQA data with that of previous state-of-the-art (SOTA) methods. We conducted several ablation studies to confirm the impact of each pretraining component, such as natural data pretraining, longitudinal data pretraining, inputs of past images, and some sections in the X-ray reports (i.e., *Findings* and *Impression*). Through extensive experiments, we showed that the PLURAL model outperformed the other SOTA methods not only in diff-VQA for longitudinal X-rays but also in conventional VQA for a single X-ray image.

## 2. Method

The training process of the PLURAL model includes a series of three training stages as follows (see Figure 1): pretraining using natural images and texts, pretraining using longitudinal chest X-ray data, and finetuning using diff-VQA data. To train the network using single or longitudinal images according to the different stages, we adopted a Transformer-based network (Wang et al., 2022) in the first stage and modified its inputs in the second and third stages (see Stage 2 and 3 in Figure 1).

### 2.1. Vision-Language Model Architecture

The VLM architecture we used in the second and third stages is shown in Figure 2. Based on a Transformer encoder-decoder architecture proposed by Wang et al. (2022), we designed our own architecture by attaching a new input branch for a past image (i.e., input branch for  $i_{past}$  in Figure 2). This design allowed the model to take two longitudinal images as input simultaneously.

Both past ( $i_{past} \in R^{H \times W}$  where  $H = 384$  and  $W = 384$ ) and current ( $i_{cur} \in R^{H \times W}$ ) images are encoded using the same image encoder ( $E_{img}(i_{past}) \in R^{\hat{H} \times \hat{W} \times E}$  where  $\hat{H} = 24$ ,  $\hat{W} = 24$ , and  $E = 1024$ ), which is a ResNet101(He et al., 2016), and were flattened ( $v_{past} = F(E_{img}(i_{past})) \in R^{N \times E}$  where  $N = \hat{H} \cdot \hat{W} = 576$ ). Then, the flattened feature vector for each image is linearly projected ( $P_{img}(v_{past}) \in R^{N \times D}$  where  $D = 768$ ), followed by the addition of the positional ( $p_{enc}^{img} \in R^{N \times D}$ ) encoding. After that, to differentiate the time points of the two input images, we separately added the time encoding for each image ( $t_{enc}^{past} \in R^{N \times D}$  for past;  $t_{enc}^{cur}$  for current) as trainable parameters:

$$v_{trans}^{past} = P_{img}(v_{past}) + p_{enc}^{img} + t_{enc}^{past} \quad (1)$$

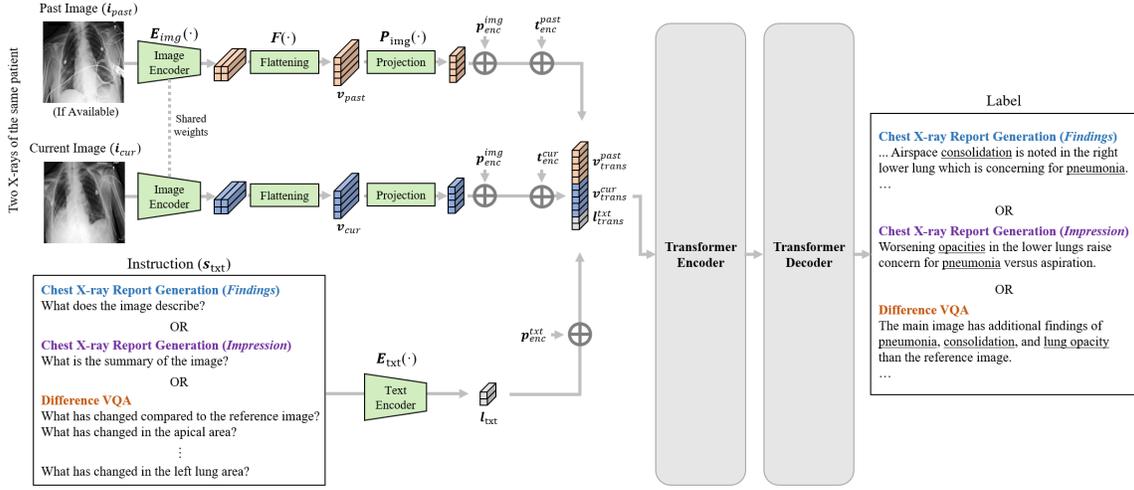


Figure 2: Proposed VLM architecture used in the second and third training stages of PLURAL. We modified a basic Transformer architecture by adding a new input branch for a past image. This enabled the model to take concurrently two longitudinal chest X-ray images as an input.

$$v_{trans}^{cur} = P_{img}(v_{cur}) + p_{enc}^{img} + t_{enc}^{cur} \quad (2)$$

The text instruction ( $s_{txt}$ ) was encoded using the text encoder ( $l_{txt} = E_{txt}(s_{txt}) \in R^{N_t \times D}$  where  $N_t$  is the length of an instruction sequence; byte-pair encoding (Sennrich et al., 2015) and word embedding used), followed by the addition of the positional encoding ( $p_{enc}^{txt} \in R^{N_t \times D}$ ,  $l_{trans}^{txt} = l_{txt} + p_{enc}^{txt}$ ). Finally, all the processed features of the input images and instruction were aggregated, producing the concatenated input feature ( $i_{trans} \in R^{(2 \cdot N + N_t) \times D}$ ) for the Transformer encoder:

$$i_{trans} = \text{concat}(v_{trans}^{past}, v_{trans}^{cur}, l_{trans}^{txt}) \quad (3)$$

The Transformer encoder (6 layers) subsequently encoded the combined feature ( $i_{trans}$ ), and the Transformer decoder (6 layers) disentangled the output of the Transformer encoder to generate an answer to the instruction. The loss function was a cross-entropy loss as follows:  $\mathcal{L} = -\sum_{j=1}^{|y|} \log P_{\theta}(y_j | y_{<j}, i_{past}, i_{cur}, s_{txt})$ , where  $\mathcal{L}$  is a cross-entropy loss,  $y$  is an output of the Transformer decoder, and  $\theta$  is trainable model parameters.

## 2.2. Stage 1: pretraining using natural images and texts

In the first stage of the training process of PLURAL, the network was pretrained using various tasks (eight tasks including image captioning, image-text matching, VQA, etc.) and datasets of natural images and texts (e.g., CC12M dataset (Changpinyo et al., 2021), SBU dataset (Ordonez et al., 2011), COCO dataset (Lin et al., 2014), etc.) as described in the work of Wang et al. (2022). Since all the tasks in the first stage consider a single image (or no image) as an input, in this stage, the network architecture included neither an input branch nor time encoding for a past image (i.e., no input branch for  $i_{past}$  in Figure 2). The loss function was defined only to consider a single input image as follows:  $\mathcal{L} = -\sum_{j=1}^{|y|} \log P_{\theta}(y_j | y_{<j}, i_{cur}, s_{txt})$ , which is different from the loss function used in the

second and third stages. In our implementation, we utilized a pretrained model weight available on the web<sup>1</sup> (OFA<sub>base</sub> model with 184M parameters) instead of training the model from scratch by ourselves. Detailed training parameters (e.g., a learning rate, an optimizer, and a dropout ratio) can be found in Wang et al. (2022).

### 2.3. Stage 2: pretraining using longitudinal chest X-ray data

In the second stage, we continued to pretrain the model from the first stage using longitudinal chest X-ray data. To do that, we added a new input branch for a past image (see Stage 2 in Figure 1), modifying the network architecture as shown in Figure 2.

We utilized two sets of longitudinal data for pretraining, consisting of longitudinal chest X-rays with QA pairs (MIMIC-Diff-VQA dataset (Hu et al., 2023); see Section 2.5 for details) and radiologist’s reports (MIMIC-CXR dataset (Johnson et al., 2019)). Two main sections, *Findings* and *Impression*, in the chest X-ray reports were used for pretraining. We introduced separate instructions for each section, “What does the image describe?” for the *Findings* section and “What is the summary of the image?” for the *Impression* section. These sections include valuable information about the longitudinal changes in chest X-ray images, which can serve as important clues to solve the diff-VQA task effectively. For example, as shown in the ‘Label’ box in Figure 2, they imply many indications for the additional findings, such as consolidation, opacity, and pneumonia (see underlying words in Figure 2) that are the key answers to the difference VQA question (e.g., “What has changed compared to the reference image?”).

### 2.4. Stage 3: finetuning using difference VQA data

In the last training stage of PLURAL, the model from the second stage was finetuned using only the diff-VQA data without any modification of the network architecture (see Stage 3 in Figure 1). Details of training configurations are summarized in Appendix C.

### 2.5. Longitudinal Chest X-ray Datasets

**MIMIC-CXR** (Johnson et al., 2019) is a large-scale chest X-ray dataset comprising 377,110 chest X-ray images and 227,827 reports from 63,478 patients. Each report contains two main sections: *Findings*, which contains a detailed description of the image, and *Impression*, which provides a concise summary of the findings. We utilized the training and validation data based on the official split to train the PLURAL model in the second stage. We extracted the *Findings* sections from the reports and performed preprocessing (e.g., eliminating special characters, typos, and scarce words), following the work of Chen et al. (2020). To extract the *Impression* sections, we followed the method of Endo et al. (2021). We only used chest X-ray images of posterior-anterior or anterior-posterior views, discarding those of other views such as lateral. For each pair of an X-ray image and a report (i.e., *Findings* or *Impression*), a chest X-ray of the same patient taken at a prior visit was selected and allocated to construct a longitudinal image dataset. We also utilized a single X-ray image to pretrain the PLURAL model in cases where there were no prior visits. We excluded all the images (and corresponding reports) included in the test sets of the VQA

---

1. <https://github.com/OFA-Sys/OFA>

Table 1: Comparison of AI performance between previous SOTA methods and PLURAL in diff-VQA.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MCCFormers (2021)	0.214	0.190	0.170	0.153	0.319	0.340	0
IDCPCL (2022)	0.614	0.541	0.474	0.414	0.303	0.582	0.703
EKAID (2023)	0.628	0.553	0.491	0.434	0.339	0.577	1.027
PLURAL	<b>0.704</b>	<b>0.633</b>	<b>0.575</b>	<b>0.520</b>	<b>0.381</b>	<b>0.653</b>	<b>1.832</b>

(a)	(b)	(c)			
Past Image	Current Image	Past Image	Current Image	Past Image	Current Image
<b>Q:</b> What has changed compared to the reference image?		<b>Q:</b> What has changed compared to the reference image?		<b>Q:</b> What has changed in the right lung area?	
<b>GT:</b> The main image is missing the findings of pneumonia, and consolidation than the reference image.		<b>GT:</b> The main image has additional findings of lung opacity, and pneumonia than the reference image. The main image is missing the finding of cardiomegaly than the reference image.		<b>GT:</b> The level of pleural effusion has changed from small to moderate.	
<b>EKAID:</b> The main image has additional findings of lung opacity, and atelectasis than the reference image.		<b>EKAID:</b> The main image has additional findings of atelectasis, and lung opacity than the reference image.		<b>EKAID:</b> The level of pneumothorax has changed from small to moderate.	
<b>PLURAL:</b> The main image is missing the findings of pneumonia, and consolidation than the reference image.		<b>PLURAL:</b> The main image has additional findings of pneumonia, and lung opacity than the reference image. The main image is missing the finding of cardiomegaly than the reference image.		<b>PLURAL:</b> The level of pleural effusion has changed from small to moderate.	

Figure 3: Selected test cases to qualitatively compare the outputs of EKAID and PLURAL in diff-VQA. We found that PLURAL was better at capturing the longitudinal change of multiple lung findings (a), (b), and change of the severity level of abnormalities, such as worsening of pleural effusion in the right low part of the lung (c).

dataset (i.e., MIMIC-Diff-VQA). In other words, the final test set for diff-VQA was totally independent of this dataset.

**MIMIC-Diff-VQA** (Hu et al., 2023) is a VQA dataset containing 700,703 QA pairs about 164,324 sets of longitudinal chest X-ray images. The images are a subset of those of the MIMIC-CXR dataset. QAs were created using an automatic algorithm based on radiologist’s reports. Those QA pairs are categorized as difference, presence, abnormality, view, location, level, and type based on the forms of questions. We used 131,563 and 16,372 QA pairs of ‘difference’ form to train and validate the PLURAL model during the second and third stages (see Figure 1) and 16,389 pairs as an independent test set to measure the performance of the model, based on the official split of the dataset.

Table 2: Results of the ablation study when we removed each pretraining stage in PLURAL.

Model	Pretraining		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
	Natural images and texts	Longitudinal chest X-ray data							
(a)			0.651	0.575	0.513	0.455	0.341	0.588	1.088
(b)	✓		0.682	0.611	0.552	0.496	<b>0.381</b>	0.640	1.733
(c)		✓	0.680	0.612	0.554	0.501	0.370	0.647	1.796
PLURAL	✓	✓	<b>0.704</b>	<b>0.633</b>	<b>0.575</b>	<b>0.520</b>	<b>0.381</b>	<b>0.653</b>	<b>1.832</b>

Table 3: Results of the ablation experiments to investigate how the structure of inputs in the pretraining step (Stage 2 in Figure 1) affects the PLURAL model’s performance.

Model	Past Image	Chest X-ray Report		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
		<i>Findings</i>	<i>Impression</i>							
(a)				0.682	0.611	0.552	0.496	0.381	0.640	1.733
(b)		✓	✓	0.695	0.623	0.562	0.504	0.370	0.635	1.728
(c)	✓		✓	0.678	0.611	0.557	0.505	<b>0.391</b>	0.650	1.731
(d)	✓	✓	✓	0.696	0.627	0.571	0.517	0.388	<b>0.656</b>	<b>1.847</b>
PLURAL	✓	✓	✓	<b>0.704</b>	<b>0.633</b>	<b>0.575</b>	<b>0.520</b>	0.381	0.653	1.832

### 3. Experiments

#### 3.1. Difference Visual Question Answering

As shown in Table 1, PLURAL achieved the highest scores in all the natural language generation (NLG) metrics, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), compared to previous state-of-the-arts methods, such as MCCFormers (Qiu et al., 2021), IDCPCl (Yao et al., 2022), and EKAID (Hu et al., 2023). The differences between the scores were significant (e.g., CIDEr: 1.832 for PLURAL vs. 1.027 for EKAID). Additionally, Figure 3 showcases selected examples to qualitatively compare the AI outputs of EKAID and PLURAL. We discovered that, on average, PLURAL was better at capturing the longitudinal change of multiple lung findings (see Figure 3(a) and Figure 3(b)) and the severity level of abnormalities (see Figure 3(c)).

#### 3.2. Ablation Study

We performed ablation studies to assess the impact of pretraining steps in PLURAL. We removed each pretraining stage individually to check how it affected performance. When we removed all pretraining stages and only performed the finetuning stage (Stage 3 in Figure 1), the model reported the lowest performance (see (a) in Table 2). However, incorporating each pretraining step significantly improved model performance. For instance, adding the pretraining stage using longitudinal chest X-ray data led to a significant increase in performance (e.g., CIDEr: 1.088 for model (a) vs. 1.796 for model (c) in Table 2). The best performance was achieved when both pretraining stages were combined (see the PLURAL model in Table 2).

We also conducted ablation experiments to investigate how the structure of inputs in the pretraining stage affects the PLURAL model’s performance. Specifically, we removed past images, *Findings*, or *Impression* sections of the X-ray reports during the model’s pretraining in stage 2 (Figure 1). As shown in Table 3, the best performance in terms of BLEU scores was achieved when we used past images and all report sections together as inputs. We

Table 4: Comparison of AI performance (exact-match accuracy) between the PLURAL and two other SOTA methods (MMQ and EKAID) on non-difference VQA.

Model	Open Question (%)	Closed Question (%)	All Question (%)
MMQ (2021)	11.5	10.8	11.5
EKAID (2023)	26.4	79.9	52.5
PLURAL <i>w/o Pretraining</i>	46.2	65.0	55.3
PLURAL	<b>51.2</b>	<b>87.3</b>	<b>68.8</b>

observed a significant improvement in performance compared to the baseline model (see (a) in Table 3) across all NLG metrics (e.g., BLEU-4: 0.496 for baseline vs. 0.520 for PLURAL). This indicates that temporal information in past images and reports was indeed useful in solving the diff-VQA task.

### 3.3. Non-difference Visual Question Answering

We conducted further training and the evaluation of the PLURAL model with two other SOTA methods, MMQ (Do et al., 2021) and EKAID (Hu et al., 2023), on six categories of questions other than the ‘difference’ category (i.e., ‘location’, ‘views’, etc.; see Section 2.5). These questions did not require the consideration of longitudinal differences between two X-ray images (i.e., non-difference VQA). The questions were of two types: open and closed. For open questions, answers were in free form, while for closed questions, answers were either ‘yes’ or ‘no’. We calculated exact-match accuracy between the model outputs and ground-truths in both cases.

Table 4 summarizes the results of the evaluation of non-difference VQA. The PLURAL model achieved the highest accuracy compared to PLURAL without pretraining (e.g., 65.0% vs. 87.3% in closed questions), and the two SOTA methods (e.g., 26.4% for EKAID vs. 51.2% for PLURAL in open questions). These results indicate that the PLURAL model is effective not only in answering questions about the changes in longitudinal images but also in answering other types of questions, such as the ones related to the location of abnormalities or X-ray view (see Appendix B for examples).

## 4. Conclusion

In this study, we present a Transformer-based VLM called PLURAL that is pretrained on natural images and texts, and longitudinal chest X-ray data to solve the difference VQA problem. Our approach allows the PLURAL model to benefit from both pretrained VLM knowledge and temporal information in the X-ray reports that describe the changes in lung abnormalities and diseases over time. Throughout the experiments, we have demonstrated that the PLURAL method outperformed previous SOTA methods in both diff-VQA for longitudinal X-rays and conventional VQA for a single X-ray image. We believe that this new model and training pipeline would be beneficial in promoting the application of the difference VQA task to assist radiologists in reading longitudinal chest X-rays.

## References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.
- Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.
- MEJ Callister, DR Baldwin, AR Akram, S Barnard, P Cane, J Draffan, K Franks, F Gleeson, R Graham, P Malhotra, et al. British thoracic society guidelines for the investigation and management of pulmonary nodules: accredited by nice. *Thorax*, 70(Suppl 2):ii1–ii54, 2015.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020.
- Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer, 2021.
- Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165, 2023.
- Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
- Jang Ho Lee, Ock-Hwa Kim, Yeon Joo Kim, Tae Sun Shim, and Kyung-Wook Jo. Changes in chest x-ray findings in 1-and 2-month group after treatment initiation for suspected pulmonary tuberculosis. *The Korean Journal of Internal Medicine*, 35(5):1145, 2020.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Ming-guang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.
- OpenAI. Gpt-4 technical report. 2023.

- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980, 2021.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- Linli Yao, Weiyang Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3108–3116, 2022.
- Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.

## Appendix A. Error Analysis and Radiologist Evaluation

To perform an error analysis, we randomly selected five cases where PLURAL failed to provide the exact-matched texts with ground truths. Then, we asked a radiologist (14 years of experience) to score the correctness (1-5) of the ground truths and the answers of PLURAL and explain the reasons for the scoring (see sample cases Figure 4). In three of five cases, the scores of PLURAL were the same as those of the GTs. In one case, PLURAL produced an output text that described the change of abnormality in time-reverse order. Surprisingly, in the last remaining case, the score of PLURAL was higher than that of the GT, implying the possibility that GTs might not be the optimal description for longitudinal chest X-rays.

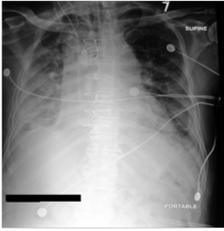
(a)	(b)
<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p><b>Past Image</b></p>  </div> <div style="text-align: center;"> <p><b>Current Image</b></p>  </div> </div>	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p><b>Past Image</b></p>  </div> <div style="text-align: center;"> <p><b>Current Image</b></p>  </div> </div>
<p><b>Q:</b> What has changed compared to the reference image?</p> <p><b>GT:</b> The main image has an additional finding of atelectasis than the reference image.</p> <p><b>PLURAL:</b> The main image has an additional finding of edema than the reference image. The main image is missing the finding of atelectasis than the reference image.</p> <p><b>Correctness of GT: 2</b>  <b>Reason for the score:</b> Consolidation in the left lung is noted on the main image, which could be pneumonia.</p> <p><b>Correctness of PLURAL: 2</b>  <b>Reason for the score:</b> Consolidation in the left lung is noted on the main image, which could be pneumonia.</p>	<p><b>Q:</b> What has changed in the left lung area?</p> <p><b>GT:</b> The level of pleural effusion has changed from small to moderate.</p> <p><b>PLURAL:</b> The level of pleural effusion has changed from moderate to small.</p> <p><b>Correctness of GT: 5</b>  <b>Reason for the score:</b> Correct.</p> <p><b>Correctness of PLURAL: 1</b>  <b>Reason for the score:</b> Reading order appears to have been reversed.</p>

Figure 4: Radiologist evaluation of error cases: (a) GT and PLURAL with equal correctness scores; (b) Lower PLURAL score due to reversed severity order.

## Appendix B. Qualitative Analysis on Non-difference Visual Question Answering

(a)		(b)		(c)	
Past Image	Current Image	Past Image	Current Image	Past Image	Current Image
					
<b>Q:</b> What type is the pneumonia?		<b>Q:</b> What level is the atelectasis?		<b>Q:</b> Where in the image is the atelectasis located?	
<b>GT:</b> Multifocal		<b>GT:</b> Minimal		<b>GT:</b> Right upper lung area	
<b>EKAID:</b> Interstitial		<b>EKAID:</b> Mild		<b>EKAID:</b> Left lower lung area	
<b>PLURAL:</b> Multifocal		<b>PLURAL:</b> Minimal		<b>PLURAL:</b> Right upper lung area	

Figure 5: Selected test cases to qualitatively compare the outputs of EKAID and PLURAL in non-difference VQA. Non-difference VQA contains various categories of QA sets such as type (a), level (b), and location (c).

## Appendix C. Implementation Details

We used an AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate =  $1e-4$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and a batch size = 16 for pretraining and finetuning stages. The other parameters for the Transformer include a dropout ratio = 0.1, a weight decay = 0.01, and a stochastic depth = 0.1. In both stages 2 and 3, we stopped the training earlier when the validation loss was saturated and not improved for 5K steps. The maximum input and output text sequence length is set to 100. The hardware specifications we used were CPU = AMD Ryzen Threadripper PRO 3955WX 16-Cores, GPU = three NVIDIA A6000 GPUs with 48GB memory.