

Supplementary Materials: Efficiency in Focus: LayerNorm as a Catalyst for Fine-tuning Medical Visual Language Models

Anonymous Authors

1 METRICS

In this study, we utilized a suite of metrics designed to rigorously evaluate the performance of Medical Visual Language Models (Med-VLMs) across diverse tasks such as Medical Visual Question Answering (Med-VQA) and Medical Imaging Report Generation (Med-IRG). These metrics not only assess accuracy but also measure the semantic and structural quality of the generated texts.

Accuracy (ACC), serving as the gold standard for the Med-VQA task, quantifies the percentage of correct answers provided by the models. For small-scale VLMs, accuracy is assessed manually, ensuring precise validation. For large-scale models, accuracy evaluation is automated using sophisticated GPT-based algorithms, as illustrated in Figure 2 in the **Main Paper**. This method provides scalable and consistent assessment capabilities across multiple datasets:

$$\text{ACC} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\%$$

BERTScore utilizes the contextual embeddings from BERT-like models, comparing the cosine similarity between the tokens of the generated text and the reference text. This metric provides a measure of semantic accuracy that is particularly useful for contexts requiring a deep understanding:

$$\text{BERTScore} = \frac{1}{|C|} \sum_{x \in C} \max_{y \in R} \frac{x \cdot y}{\|x\| \|y\|}$$

where C is the set of tokens in the candidate sentence, and R is the set of tokens in the reference sentence.

METEOR Score, another crucial metric, assesses translation quality by considering exact, stem, synonym, and paraphrase matches between words in the generated text and the reference. It is computed as:

$$\text{METEOR} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

where P (precision) and R (recall) are calculated based on the alignments between the model’s output and the reference data, factoring in synonymy and paraphrasing.

Rouge-L focuses on the longest common subsequence between the generated text and the reference, suitable for evaluating the performance of models on Med-IRG tasks. It is particularly useful for assessing the fluency and order of the generated narratives:

$$\text{Rouge-L} = \frac{(1 + \beta^2) \cdot R_l \cdot P_l}{R_l + \beta^2 \cdot P_l}$$

where R_l is the recall, P_l is the precision, and β is typically set to favor recall (e.g., $\beta = 1.2$), reflecting the importance of capturing all necessary information.

Mean Output Token Length quantifies the average number of tokens produced per prompt, which is indicative of the model’s verbosity or succinctness in generating medical reports:

$$\text{Mean Token Length} = \frac{\text{Total Number of Tokens in all Outputs}}{\text{Number of Prompts}}$$

These metrics collectively provide a robust framework for evaluating the adaptability, accuracy, and utility of Med-VLMs across different medical tasks, ensuring that the models not only achieve high accuracy but also maintain consistency and reliability in their output.

2 MORE APPLICATION DETAILS

For the large-scale Med-VLM (LLaVA-Med), LoRA-tuning was applied with specific settings to optimize performance without extensive parameter increase. The settings included a rank (Lora-r) of 4, a scaling factor (LoRA-alpha) of 16, a dropout rate (LoRA-dropout) of 0.05, and no bias adjustment (LoRA-bias set to None). In contrast, when Prefix-tuning was employed, we initialized 10 virtual tokens (num-virtual-tokens = 10) and enabled prefix projection (prefix-projection = True) to modify the input sequence effectively.

Similarly, for the smaller-scale model MISS, LoRA-tuning was configured with a rank of 4 and an increased scaling factor of 32, alongside a higher dropout rate of 0.1 to account for its lesser parameter base, enhancing its adaptability with minimal structural changes. Prefix-tuning for MISS mirrored that of LLaVA-Med, with 10 virtual tokens and active prefix projection.

Furthermore, during the inference phase with LLaVA-Med, we adjusted the temperature parameter to 0.2, which is commonly used in LLM inference to control the randomness in the prediction process. We utilized a beam search strategy with a single beam (num-beams=1) and processed the input in one chunk (num-chunks=1), ensuring efficient and focused generation of medical language responses.

3 EVALUATION AND PROMPTS

As illustrated in Figure 2 in the **Main Paper**, for evaluating the accuracy of large-scale Med-VLMs on the Med-VQA task, we employ an automated evaluation using ChatGPT 3.5 and the prompt we used is:

"You will act as an intelligent evaluator of answers generated by Generative Medical Visual Language Model (Med-VLM). Please note that Med-VLM answers may be more varied than benchmark answers. If a Med-VLM answer is approximately correct about the image from a medical point of view, it should be regarded as correct even if there are some differences from the benchmark answer, so do not arbitrarily give an incorrect assessment. I'll give you the \$questions\$, \$Med-VLM answer\$, and \$ground-truth\$. You must output a word \$correct\$ or \$incorrect\$."

This method assesses the model’s ability to generate longer responses that are contextually accurate.

For small-scale Med-VLMs, which typically produce responses that are brief and closely align with the ground-truth—often consisting of only one or two words—we utilize manual assessment to determine accuracy. This approach ensures precision in evaluating

the direct alignment of model outputs with the expected responses in a controlled, medical-specific setting.

During the inference phase of the Med-VQA task, prompts are structured to simulate a dialogue between a curious user and an artificial intelligence healthcare assistant, capable of understanding and interpreting visual content through natural language. The typical prompt format used is:

"A chat between a curious user and an artificial intelligence healthcare assistant. The assistant is able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language. The visual content will be provided with the following format: <Image>visual content</Image>."

For the Med-IRG task, a uniform prompt format is adopted to guide the generation of detailed descriptive reports based on medical imaging. The standardized prompt for this task is:

"<Image>visual content</Image>. Describe the given chest x-ray image in detail."

These prompts are meticulously designed to elicit detailed and contextually relevant responses from Med-VLMs, facilitating a comprehensive evaluation of their practical utility and performance in real-world medical applications.

4 MORE RESULTS

Figure 1 further demonstrates the out-of-distribution (OOD) performance of LLaVA-Med on the VQA-RAD dataset, after fine-tuning with various PEFT methods on the Slake dataset. In the first two open-ended questions, models fine-tuned with LN-tuning provided correct answers, while those tuned with attention mechanisms and FFN-tuning failed to do so. Notably, although FFN-tuning exhibited superior performance on the Slake test dataset, it lagged behind LN-tuning in both length and accuracy on the VQA-RAD dataset. Compared to the two extrinsic fine-tuning methods, LN-tuning also demonstrated better output diversity and accuracy.

Figure 2 supplements this by showcasing the zero-shot performance of LLaVA-Med on the MIMIC-test dataset, trained on the Slake dataset. In the visualization of the second medical report generation task, the model fine-tuned with LN-tuning was the only one that did not mistakenly diagnose a normal chest x-ray as abnormal. This further illustrates the robust stability and generalizability of LN-tuning, affirming its effectiveness in handling complex medical imaging tasks.

Question: Is there airspace consolidation on the left side?



Assistant



LayerNorm: Yes, the chest X-ray shows airspace consolidation on the left side.

Attention: No, the chest X-ray does not show any airspace consolidation on the left side.

FFN: No.

LoRA: Yes, the chest X-ray shows airspace consolidation on the left side.

Prefix: Yes, the chest X-ray shows airspace consolidation on the left side.

Question: Is the cardiac silhouette less than half the lateral diameter of the chest wall?



Assistant



LayerNorm: Yes, the cardiac silhouette in the image is less than half the lateral diameter of the chest wall, which is a sign of cardiomegaly or an enlarged heart.

Attention:No, the cardiac silhouette is greater than half the lateral diameter of the chest wall.

FFN: No.

LoRA: No, the cardiac silhouette is greater than half the lateral diameter of the chest wall.

Prefix: Yes, the cardiac silhouette in the image is less than half the lateral diameter of the chest wall, which is a sign of cardiomegaly or an enlarged heart.

Question: What is the abnormality by the right hemidiaphragm?



Assistant



LayerNorm: The chest X-ray shows a soft tissue density by the right hemidiaphragm. This could indicate an abnormality or mass in that area, such as a tumor or other growth. Further evaluation and tests would be needed to determine the exact cause and nature of the abnormality.

Attention: The chest X-ray shows a homogenous opacity by the right hemidiaphragm. This means that there is an area of increased density in that region, which could be due to various reasons such as fluid accumulation, infection, or a mass. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment.

FFN: The chest X-ray shows a homogenous opacity by the right hemidiaphragm. This means that there is an area of increased density in that region, which could be due to various reasons such as fluid accumulation, infection, or a mass. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment.

LoRA: The chest X-ray shows an abnormality by the right hemidiaphragm, which is consistent with a diaphragmatic hernia. A diaphragmatic hernia is a condition where there is an opening or defect in the diaphragm, allowing abdominal organs to protrude into the chest cavity.

Prefix: The chest X-ray shows a soft tissue density by the right hemidiaphragm. This could indicate an abnormality or mass in that area, such as a tumor or other growth. Further evaluation and tests would be needed to determine the exact cause and nature of the abnormality.

Question: How is the patient oriented?



Assistant



LayerNorm: The patient appears to be oriented, meaning they are able to respond to stimuli and communicate their needs.

Attention: The patient is oriented, meaning they are able to respond to stimuli and orient themselves in their environment. This is an important aspect of a patient's mental and emotional well-being, as it indicates that they are able to perceive their surroundings and react to stimuli appropriately.

FFN: The patient appears to be oriented, meaning they are able to respond to stimuli and communicate their needs.

LoRA: The patient appears to be oriented, meaning they are able to respond to stimuli and communicate their needs.

Prefix: The patient appears to be oriented, meaning they are able to respond to stimuli and communicate their needs.

Figure 1: More results of LLaVA-Med out of distribution performance on the VQA-RAD dataset

Instruction: Describe the given chest x-ray image in detail.



Assistant 

LayerNorm: The chest x-ray image shows a patient with a right-sided pneumothorax. A pneumothorax is a condition where air accumulates in the pleural space, which is the area between the lung and the chest wall. This can cause the lung to collapse, leading to difficulty breathing and chest pain. The image likely displays the presence of air in the pleural space, which is a key feature of a pneumothorax.

Attention: The chest x-ray image shows a right-sided pleural effusion, which is an abnormal accumulation of fluid in the pleural space surrounding the lungs. Additionally, there is a left-sided pleural effusion, which is smaller in size compared to the right one. The image also reveals a right-sided pneumothorax, which is the presence of air in the pleural space causing the lung to collapse. Lastly, there is a left-sided pneumothorax that is less prominent than the right one.

FFN: The chest x-ray image shows a left-sided pleural effusion, which is an abnormal accumulation of fluid in the pleural space surrounding the lung. Additionally, there is a right-sided pleural effusion, which means there is fluid accumulation on the right side as well. The image also reveals a left-sided pneumothorax, which is the presence of air in the pleural space causing the lung to collapse. Lastly, there is a right-sided pneumothorax, indicating air in the pleural space on the right.

LoRA: The chest x-ray image shows a left-sided pleural effusion, which is an abnormal accumulation of fluid in the pleural space surrounding the lung. Additionally, there is a right-sided pleural effusion, which means there is fluid accumulation on the right side as well. The image also reveals a left-sided pneumothorax, which is the presence of air in the pleural space causing the lung to collapse. Lastly, there is a right-sided pneumothorax, indicating air in the pleural space on the right.

