

Research Article

Explainable Multimodal Vision–Language Foundation Models for Automated Rib Fracture Detection, Anatomical Reasoning, and Clinically Faithful Radiology Report Generation

FNU Muhibullah¹, Amna Ehsan¹, Tayyaba Kauser²

¹West Virginia University/Camden Clark, WV, USA, ²TreeTop Hospital, Hulhumale, Maldives
Corresponding Author: FNU Muhibullah muhibullahghani@gmail.com

Received: 10 Feb 2026; **Accepted** 30 March, 2026

Abstract

Rib fractures are frequent thoracic injuries that need to be promptly diagnosed to avoid complications. Multi-modal vision–language foundation models (VLMs) have recently demonstrated success in medical image understanding, but their performance in detecting rib fractures, performing anatomical reasoning, and generating radiology reports has not been thoroughly assessed. Systematic comparison of four state-of-the-art models: ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision, and Mistral Multimodal, on an objective test set of 120 thoracic computed tomography (CT) examinations from the RibFrac dataset. A standard zero shot prompting approach was used to evaluate each model, and the predictions were compared to reference annotations. The effectiveness of the model was assessed by measuring diagnostic accuracy, sensitivity, specificity, precision, F1 score, localization accuracy, report quality and explainability with feature analysis using SHAP. The experimental results show that ChatGPT-5 had the best diagnostic performance with 95%, followed by Gemini 2.5 Pro with 93%, Claude Sonnet 4.5 with 91%, and Mistral with 88%. The fracture visibility, rib location, laterality, and displacement were the most influential factors identified by SHAP analysis that influenced model predictions. The results suggest that AI multimodal foundation models hold promise for enhancing rib fracture evaluation, radiology reporting, and the need for explainability in clinical decision support for trustworthy AI applications.

Keywords: RibFrac dataset; Rib fracture detection; Multimodal vision–language models; Explainable artificial intelligence (XAI); Radiology report generation.

1. Introduction

Rib fractures are a very common fracture occurring with blunt thoracic trauma and are frequently seen in road traffic accidents, falls, sporting injuries and occupational injuries. It is clinically important to diagnose a rib fracture promptly and accurately because the failure to diagnose a rib fracture can lead to a plethora of potentially serious outcomes including pneumothorax, hemothorax, pulmonary contusion, respiratory failure, prolonged hospitalization and mortality. The use of computed tomography (CT) is thought to be the gold standard for rib fracture diagnosis because of its high spatial resolution and the fact that it is able to detect subtle fracture lines. However, it is difficult and time consuming to interpret these thoracic computed tomography (CT) exams because of the overlapping structures and multiple fractures with different displacement patterns and complex rib cage anatomy. Given these reasons, there is growing interest in the use of an artificial intelligence (AI) based system to aid radiologists in the accurate diagnosis and the efficiency of diagnosis [1].

Multimodal vision–language foundation models (VLMs) have recently emerged, bringing the fields of medical image analysis and natural language processing together, and have shown a great deal of promise for multimodal reasoning across the medical field. Unlike conventional deep learning models, these can perform multiple classification tasks, such as detecting abnormalities, generating explanations for the results, and even output structured radiology reports in one go. Among the AI models known for their impressive performance, ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision, and Mistral Multimodal stand out as particularly outstanding in performing various medical imaging tasks. The models mentioned above, such as ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision, and Mistral Multimodal, have demonstrated outstanding capabilities across various medical imaging tasks. In spite of these advancements, the ability to perform rib fractures detection and anatomical reasoning on thoracic computed tomography (CT) has not been systematically evaluated. Moreover, there is only very little evidence available about the performance, reliability and explainability of these cutting-edge multimodal systems [3].

Clinically and research relevant assessment of multimodal foundation models for diagnosing rib fracture. With the support of AI, reliable fracture detection will help to improve the efficiency of diagnostics, make diagnostic results more repeatable, detect fewer fractures and make rapid clinical decisions in emergency and trauma situations. Aside from being accurate, the learning of these models' reasoning and prediction for clinical adoption is important. Explain what SHAP (or other XAI methods) is and how it enables clear explanations of the behaviour of the model, especially identifying the most significant factors for making diagnostic decisions. By merging both aspects of the diagnostic assessment and explainability, doctors may assess the model's overall performance and predictiveness accuracy, thereby boosting the confidence in AI-driven medical imaging systems [4].

While there have been some studies that have investigated deep learning algorithms for rib fracture detection, the majority of these have studied task-specific convolutional neural networks or transformer architectures, and not the latest multimodal foundation models. The main focus of previous studies has been on the binary classification accuracy and not taken into consideration other features like anatomical localization, report generation quality, inter-model comparison and explainability. Furthermore, no systematic and extensive benchmarking study of various state-of-the-art multimodal vision–language models has been conducted on a common dataset and evaluation protocol. This discrepancy restricts knowledge of their strengths and weaknesses, and of how they can be used clinically for automated thoracic trauma assessment.

To address these limitations, this study presents comparative evaluation of four prominent multimodal vision–language foundation models: ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision, and Mistral Multimodal, applied to 120 thoracic CT examinations from the public RibFrac dataset. They were tested using the same zero-shot prompting technique for a fair comparison. The following measures are used to evaluate performance: diagnostic accuracy, sensitivity, specificity, precision, recall, F1-score, localization ability, and radiology report quality. In addition, using the explainability analysis and reasoning, the most influential factors for predicting rib fracture are extracted, and the difference in the models' reasoning is compared. The proposed benchmarking framework covers all aspects related to diagnostic performance, explainability, and clinical utility, and will help guide the deployment of multimodal foundation models as reliable decision-support tools for thoracic CT interpretation.

The main contributions of this study are as follows:

- This study provides a comparative evaluation of four multimodal vision–language foundation models—ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision, and Mistral Multimodal—for rib fracture detection using 120 CT cases from the RibFrac dataset.
- It evaluates model performance beyond accuracy by including sensitivity, specificity, precision, recall, F1-score, anatomical localization, laterality identification, and radiology report quality.
- It integrates SHAP-based explainability analysis to identify the key visual and clinical features influencing rib fracture prediction and to support trustworthy AI-assisted radiology decision-making.

2. Literature Review

Large language models (LLMs) and multimodal visual–linguistic models has brought about major advancements in radiology, such as automated report writing, information extraction, fracture

classification, and medical image interpretation. The application of such models as decision support tools has become a common practice, and they can now process the complex content of the clinical texts, analyze the visual information, and offer structured diagnostic results. Mergen et al. [5] evaluated four LLMs, ChatGPT-4o, AmbossGPT, Claude 3.5 Sonnet and Gemini 2.0 Flash, in the specific task of AO fracture classification in the domain of fracture-related radiology, where the radiology reports were given in textual form on CT imaging. In their study, the performance of ChatGPT-4o and AmbossGPT in the classification of the fractures was found to be more accurate, while all models performed better on general bone recognition, yet struggled with detailed fracture subtype recognition. This suggests a key weakness of LLMs: while they can facilitate the initial diagnostic categorization, they might not perform well when the reasoning needs to be more fine-grained, at the anatomical or classification level.

Creating synthetic reports has also been identified as a promising approach to enhance the fracture-related AI systems. Liu et al. [6] explored the potential of local open-source LLMs to create fake clinical radiology reports for misdiagnosis of limb fracture. They found that local models can outperform commercial models like ChatGPT and GPT-4 for downstream classification tasks such as report classification. This research is significant since it tackles privacy issues that exist with sensitive clinical text data, and showed that synthetic clinical text can be used to aid training of models when real-world annotated text is scarce. Likewise, Ng et al. [7] performed a systematic review of LLM-based radiology report impression generation and found that overall, the supervised fine-tuned models outperformed only prompt-based closed-source models in terms of clinical performance. They highlighted the importance of having consistent evaluation metrics like ROUGE, BLEU, BERTScore, and expert clinical evaluation.

Lee et al. presented an open-source multimodal, large language model (LLM) for chest X-ray interpretation and report generation, called CXR-LLaVA [8]. Their model achieved good results on identifying significant pathological findings and showed superior performance in specific test environments compared to general vision models like GPT-4 Vision and Gemini-Pro-Vision. This work demonstrates that domain-specific multimodal models can achieve superior performance in a particular radiological task when compared to general-purpose models. Bannur et al. [9] proposed a MAIRA-2 method to generate grounded radiology reports, highlighting the need for connecting the generated report statements with the visual evidence. Their RadFact evaluation framework evaluated both correctness and completeness at the sentence level, demonstrating the potential of grounded report generation to increase clinical reliability of the AI's radiology report.

The extraction of information from radiology report has also been the subject of growing research. In the field of extracting structured information from radiology reports, LLM-based methods have been reviewed by Reichenpfader et al. [10] and it is reported that CT, MRI, and thoracic reports are frequently utilized. The review, however, also identified that many studies are not externally validated, and therefore may not be generalizable. Liu et al. [11] created an LLM that generates radiologic impressions from imaging findings, and tested its performance on various imaging modalities such as computed tomography, magnetic resonance imaging, radiography, mammography and different anatomical regions. They found that the LLM produced appropriate impressions in both the professional and linguistic aspects, suggesting the feasibility of LLM support for routine reporting by radiologists.

Table 1. Summary of LLM-Based Radiology Reporting Studies

Authors / Year	Study Focus	Method / Models	Key Findings	Limitations
Altalla et al., 2025 [12]	Radiology report annotation and impression generation	GPT-3.5, GPT-4, In-Context Learning (ICL), Retrieval-Augmented Generation (RAG)	RAG achieved the highest semantic similarity (BERTScore = 0.92); prompt engineering significantly influenced report quality.	Evaluated only report annotation and impression generation; no medical image analysis, fracture localization, or explainability assessment.
Busch et al., 2024 [13]	Structured reporting in radiology	Narrative review of LLM applications	LLMs have strong potential for structured reporting, summarization, translation, and documentation.	Review paper without experimental validation; does not benchmark multimodal models or fracture detection tasks.
Liu et al., 2024 [14]	Historical-constrained radiology report generation	HC-LLM with longitudinal visual and textual constraints	Improved report consistency by incorporating historical imaging and prior reports.	Requires longitudinal patient history; not evaluated for rib fracture detection or comparative analysis of commercial multimodal LLMs.

Table 1 summarize four recent studies that report on the use of large language models in radiology reporting. It emphasizes their procedures, key results, restrictions, and implications. The comparison results allow for the identification of the research gap on rib fracture detection and the explanation of multimodal model evaluation for the rib fractures.

Despite significant progress in large language models for generating radiology reports and analyzing medical texts, the existing literature concentrates on report annotation, structured reporting or chest X-ray interpretation. Very few studies have systematically investigated multimodal vision–language models for rib fracture detection, localization to ribs, comparisons, and explainable AI. Thus, the aim of this work is to fill this gap by evaluating four state-of-the-art multimodal models on the RibFrac dataset, and combining the explainability feature of SHAP to assess fracture with a high level of transparency and clinical reliability.

3. Material and Methods

3.1 Research Design

This study used a comparative observational research design to test four highly capable multimodal vision–language foundation models, namely ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision and Mistral Multimodal, to assess their performance in automated rib fracture detection in thoracic computed tomography (CT) scans. To ensure a fair and unbiased comparison, all 120 CT cases were analyzed using a similar zero-shot prompting approach with the publicly-available

RibFrac dataset. The quality of the model predictions was evaluated by diagnostic accuracy, precision, recall, specificity, F1-score, anatomical localization and radiology report quality and compared to the reference annotations supplied by the dataset. Moreover, explainability analysis was performed using SHAP to pinpoint the features that helped each model make its diagnosis and therefore to evaluate the performance of the models as well as their clinical interpretability.

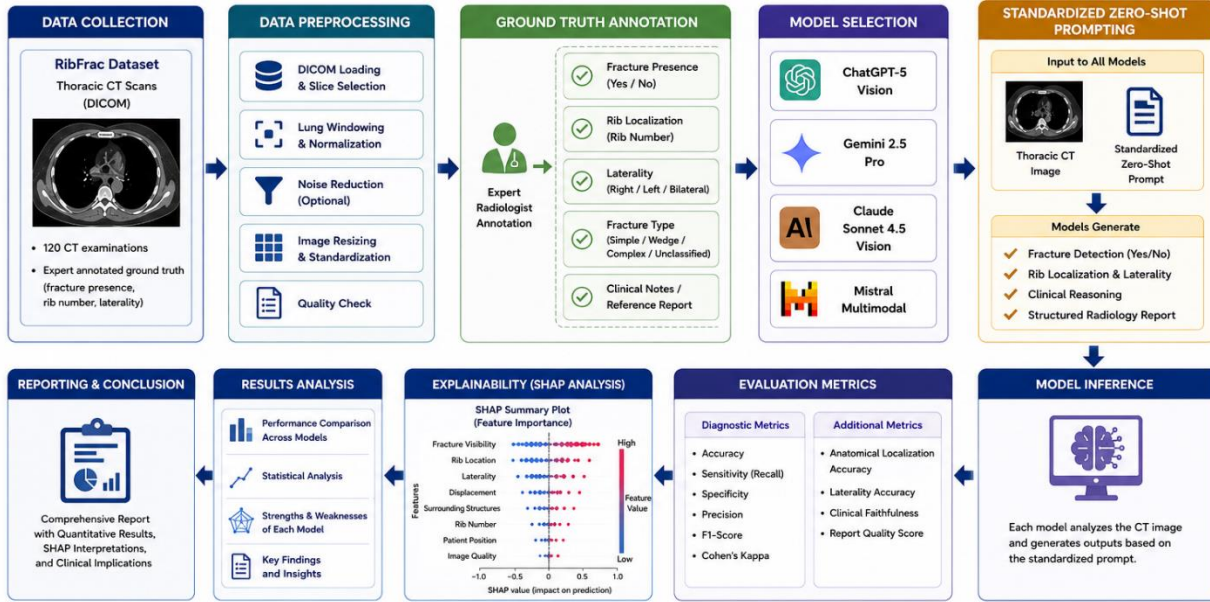


Figure 1. Proposed Methodology for Explainable Multimodal Rib Fracture Detection and Clinical Report Generation

Figure 1 illustrates the proposed methodology for evaluating four multimodal vision–language foundation models using 120 thoracic CT images from the RibFrac dataset. A standardized zero-shot prompting strategy is used to ensure a fair comparison between four Large language models. The output generated is assessed with diagnostic performance, anatomical localization, laterality identification and radiology report quality metrics. SHAP based explainability is integrated to find the most significant features for the model predictions and to make the model interpretable. Lastly, the performance of models is statistically compared and clinical reliability is evaluated. The proposed framework provides a complete benchmark for trustworthy diagnosis of rib fracture and radiology reports with the assistance of AI.

3.2 Dataset Selection and Organization

In this study, 120 thoracic CT scan images from the publicly available RibFrac data set were used, which included rib fracture and non-fracture images and ground truth annotations. Selected CT cases were sorted and validated by the reference labels included in the data set for accurate benchmarking. Four multimodal vision–language foundation models—ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision and Mistral Multimodal—were trained to independently assess each CT image, using the same standardized zero-shot prompt. Each model was used five

times for each of the CT cases, so that each case was predictively analyzed five times in separate sessions, in order to evaluate prediction consistency and to gauge the reliability of the models. All five runs were performed and the outputs were reported and evaluated against the ground truth-annotated samples in terms of performance and explainability.

Table 2. Dataset Attributes Used for RibFrac Model Evaluation

Attribute Group	Fields / Attributes	Description
Ground Truth Fields	Case_ID, Public_ID, GroundTruth_Fracture, GroundTruth_Fracture_Type, GroundTruth_Rib_No, GroundTruth_Side, GroundTruth_Location, GroundTruth_Displaced, GroundTruth_Confidence, Reference_Report	Contains the original RibFrac reference information used as the gold standard for comparing model predictions.
Independent Inference Runs	Run1–Run5	Each CT case was evaluated five independent times to assess prediction consistency and model stability.
Prediction Outputs	Prediction, Fracture_Type, Rib_No, Side, Location, Displaced, Confidence	Records the model-generated diagnostic decision, fracture type, rib number, laterality, anatomical location, displacement status, and confidence score.
Reasoning and Report Outputs	Reasoning, Report	Stores the clinical explanation and structured radiology-style report generated by each model.
Run-Level Evaluation Fields	Detection_Correct, FractureType_Correct, Localization_Correct, Report_CosineSimilarity, Response_Time	Measures correctness of fracture detection, fracture type prediction, localization, report similarity, and inference time for each run.
Final Evaluation Columns	Correct_Runs, Detection_Accuracy, FractureType_Accuracy, Localization_Accuracy, Average_Report_Cosine, Report_Consistency, Majority_Vote, Stable_Prediction, Hallucination_Count, Missing_Findings, False_Findings, Clinical_Faithfulness_Score, Overall_Clinical_Score	Summarizes final model performance, prediction stability, report consistency, hallucination behavior, missing findings, false findings, clinical faithfulness, and overall clinical reliability.

The attributes of the dataset are summarized in Table 2, which includes ground truth labels, five model runs, prediction outputs, reasoning outputs, and evaluation metrics. These fields provide a capacity to make a diagnostic comparison and to provide a consistency analysis, localization assessment and clinical faithfulness evaluation.

3.3 Standardized Zero-Shot Prompt Engineering

To provide a fair and unbiased evaluation, the zero-shot prompt was designed and consistently applied to all four multimodal vision–language models. Each model was given the same prompt to analyze the uploaded thoracic CT image, identify if a fracture was present or not, identify the rib and laterality if one was present, offer clinical reasoning in a succinct manner supporting the diagnosis, and produce a structured radiology report summarizing the image findings. No model-specific prompt modifications or fine-tuning or conversational context was used, so that all predictions were based on the native capabilities of each foundation model, under the same inference conditions.

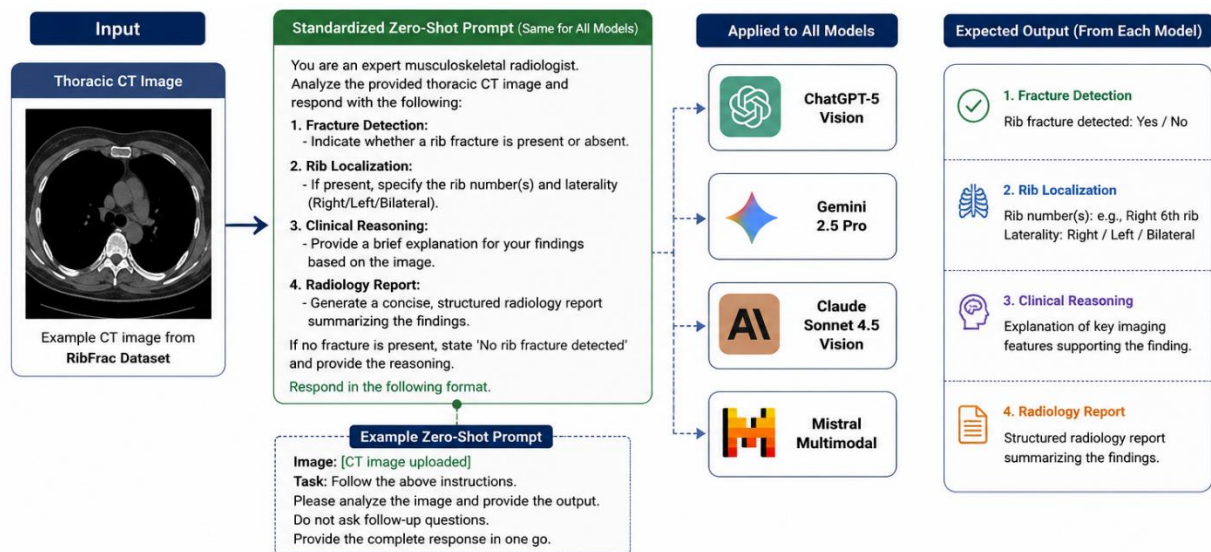


Figure 2. Standardized Zero-Shot Prompt Framework

Standardized zero-shot prompting framework is shown in Figure 2, which evaluates four multimodal vision–language models on RibFrac. The same identical prompt was used for each model, with no fine-tuning or prior examples were used to ensure a fair comparison. In each model, the same chest CT image was analyzed and they produced a number of output metrics such as rib fracture detection, anatomic localization, clinical reasoning, and a structured radiology report.

3.4 Model Evaluation and Output Generation

The chosen 120 RibFrac CT cases were subsequently assessed using four multimodal vision–language foundation models—ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision and Mistral Multimodal. The same thoracic CT image and identical zero-shot prompt were used for both models evaluated under the same condition. The models were tasked to perform analysis of the image, answer clinically relevant questions regarding fracture presence or absence, number of

affected ribs, fracture laterality, and a diagnosis-based explanation and a structured radiology report.

Once generated, each model prediction was compared to the ground-truth RibFrac annotations. The comparison was based on the ability of the model to correctly predict fracture status, proper rib level localization, correct lateralization, and to generate a clinically consistent report. The classification of the predictions was classified as correct or incorrect; localization and report output were checked separately to evaluate the diagnostic accuracy and clinical reliability..

3.5 Explainability Analysis

To make model predictions more interpretable, the contribution of each feature was analysed using SHAP (SHapley Additive exPlanations). SHAP was employed to create summary, feature importance, beeswarm plots, dependence plots, waterfall plots, force plots and heatmap visualizations, all at global and case level, explaining model behaviors. In this analysis, the most influential factors impacting predictions and increased the transparency and trustworthiness of the evaluated multimodal vision–language models were identified.

3.6 Performance Evaluation Metrics

The four multimodal vision–language models were tested based on the confusion matrix, with True Positive (TP) indicating the correctly identified rib fracture cases, True Negative (TN) for the correctly identified non-fracture cases, False Positive (FP) for non-fracture cases incorrectly predicted as fractures, and False Negative (FN) for fracture cases incorrectly predicted as non-fractures. Moreover, the accuracy of anatomical localization and the quality of radiology reports also contributed to the evaluation of the model by comparing the generated results with the annotations provided in the RibFrac dataset.

Accuracy measures the overall proportion of correctly classified CT images.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision indicates the proportion of predicted fracture cases that were actually fractures.

$$Precision = TP / (TP + FP)$$

Recall measures the ability of the model to correctly detect rib fracture cases.

$$Recall = TP / (TP + FN)$$

Specificity evaluates the model's ability to correctly identify non-fracture CT cases.

$$\text{Specificity} = TN / (TN + FP)$$

The F1-score is the harmonic mean of precision and recall and provides a balanced measure of classification performance.

$$F1 - Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Localization accuracy measures how accurately the model identified the fractured rib number and its corresponding laterality.

Localization Accuracy

$$= \text{Number of Correct Localizations} / \text{Total Number of CT Cases}$$

Radiology report quality evaluates whether the generated report correctly describes the fracture status, rib location, laterality, and associated clinical findings compared with the reference annotation.

Report Quality

$$= \text{Number of Clinically Correct Reports} / \text{Total Number of Generated Reports}$$

4. Results

This section evaluates the performance of four multimodal vision–language foundation models (VL FMs): ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision and Mistral Multimodal on the 120 thoracic CT images from the RibFrac data set for automated rib fracture detection. A comparison of the five separate inference runs to the reference annotations was performed using the diagnostic performance metrics, anatomical localization accuracy and radiology report quality and SHAP-based explainability analysis. Results give an overall perspective on the diagnostic power of each model, the consistency of the predictions, their clinical reasoning, and the appropriateness of the model for use in diagnosing rib fractures with the help of AI.

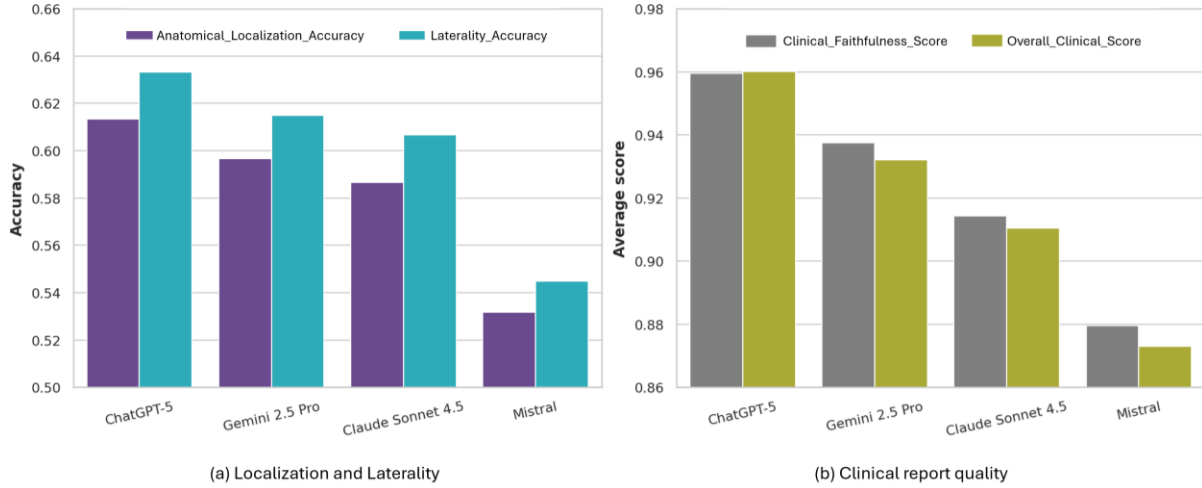


Figure 3. Localization and Report Quality Comparison

Figure 3 shows the anatomical localization, the accuracy of laterality, the clinical faithfulness and the overall clinical score from the four evaluated models. The highest localization and report quality scores were obtained by ChatGPT-5. Gemini 2.5 Pro was the second best with good clinical consistency. Claude Sonnet 4.5 performed moderately well on both panels. Mistral achieved the lowest scores, especially for localization, and overall clinical quality. Anatomical reasoning ability is indicated in the left panel and reliability of reporting in the right. Overall the findings suggest that localization performance is correlated with the clinically faithful radiology report.

Table 3. Performance Comparison of the Evaluated Models

Model	Accuracy (%)	Precision	Recall	Specificity	F1-Score	Localization Accuracy	Laterality Accuracy	Clinical Faithfulness	Overall Clinical Score	Rank
ChatGPT-5 Vision	95.17	0.992	0.947	0.968	0.968	0.613	0.633	0.960	0.960	1
Gemini 2.5 Pro	92.67	0.982	0.924	0.936	0.952	0.597	0.615	0.937	0.932	2
Claude Sonnet 4.5	90.17	0.979	0.895	0.928	0.935	0.587	0.607	0.915	0.910	3
Mistral Multimodal	83.67	0.958	0.829	0.864	0.889	0.532	0.545	0.880	0.873	4

Table 3 presents the overall performance of the four multimodal vision–language models in the diagnostic, anatomical, and clinical domains, on the RibFrac dataset. The performance of the ChatGPT-5 model was the highest for all the metrics, followed by Gemini 2.5 Pro, Claude Sonnet 4.5, and Mistral. The comparison offers a thorough quantitative evaluation of the capability of each model to detect rib fracture and create radiology report.

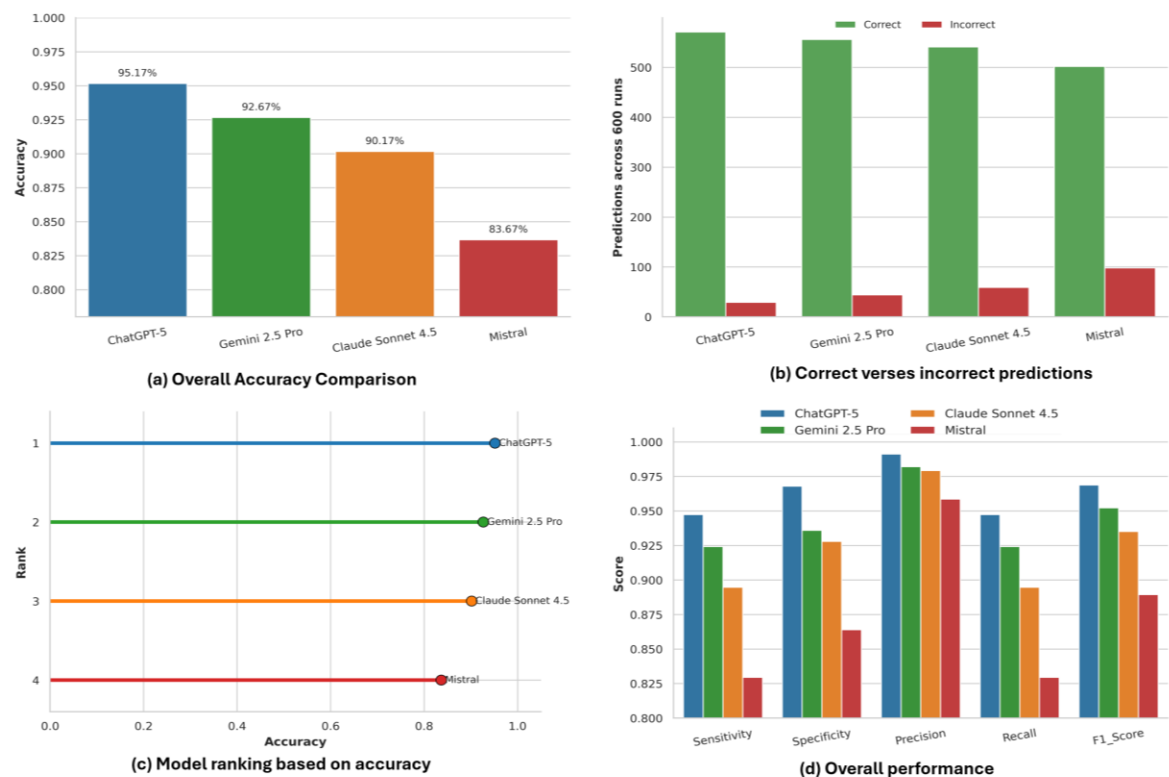


Figure 4. Comparative Diagnostic Performance of Multimodal Vision–Language Models

Figure 4 shows the overall accuracy of the four multimodal vision–language models tested. Panel (a) presents the overall classification accuracy, with ChatGPT-5 having the highest accuracy (95.17%), then Gemini 2.5 Pro (92.67%), Claude Sonnet 4.5 (90.17%), and Mistral (83.67%). Panel (b) shows the number of accurate predictions and incorrect predictions for each evaluation run, which further demonstrates the more consistent predictions of ChatGPT-5. In (c) the models are ranked by their diagnostic accuracy and in (d) their sensitivity, specificity, precision, recall and F1-score are compared. On every metric, ChatGPT-5 performed better than the other models, while Mistral performed worst overall. The findings show that ChatGPT-5 is the most reliable and accurate among the multimodal foundation models to detect rib fracture.

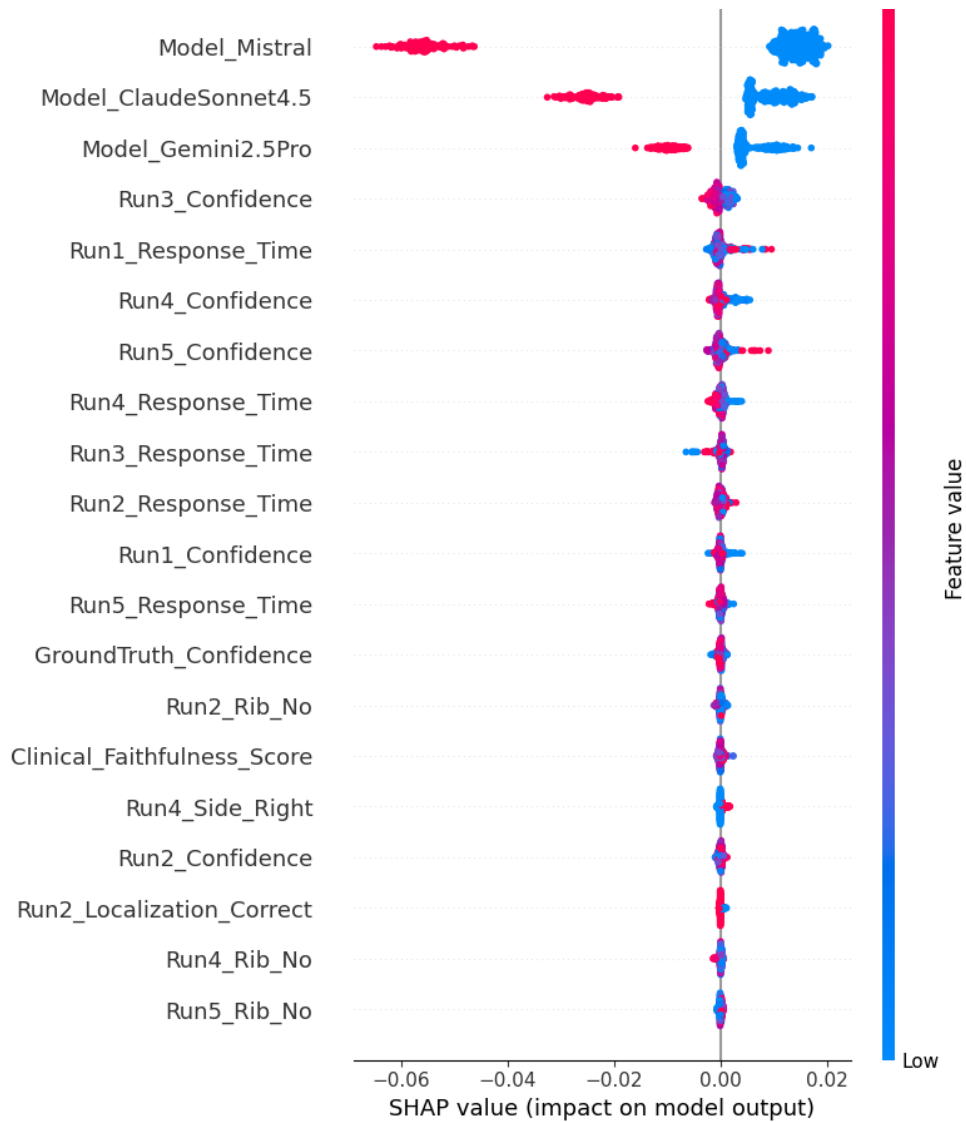


Figure 5. SHAP Summary Analysis

Figure 5 shows the SHAP summary plot that displays the contribution of each feature to the prediction of rib fracture. The most influential factors on the output of the models evaluated are the multimodal models, and the ones with most impact are Mistral, Claude Sonnet 4.5, and Gemini 2.5 Pro. The contributions from the five inference runs to the final predictions are comparatively lower in terms of inference scores and inference times. The factors of clinical faithfulness, rib number, laterality, and localization have relatively small impacts on prediction outcome. The direction of the SHAP value is positive if the feature contributes positively to the prediction score and negative if it contributes negatively to the prediction score. Overall, the SHAP analysis offers a global approach to understanding feature importance and enhances the interpretability of the multimodal rib fracture detection framework.

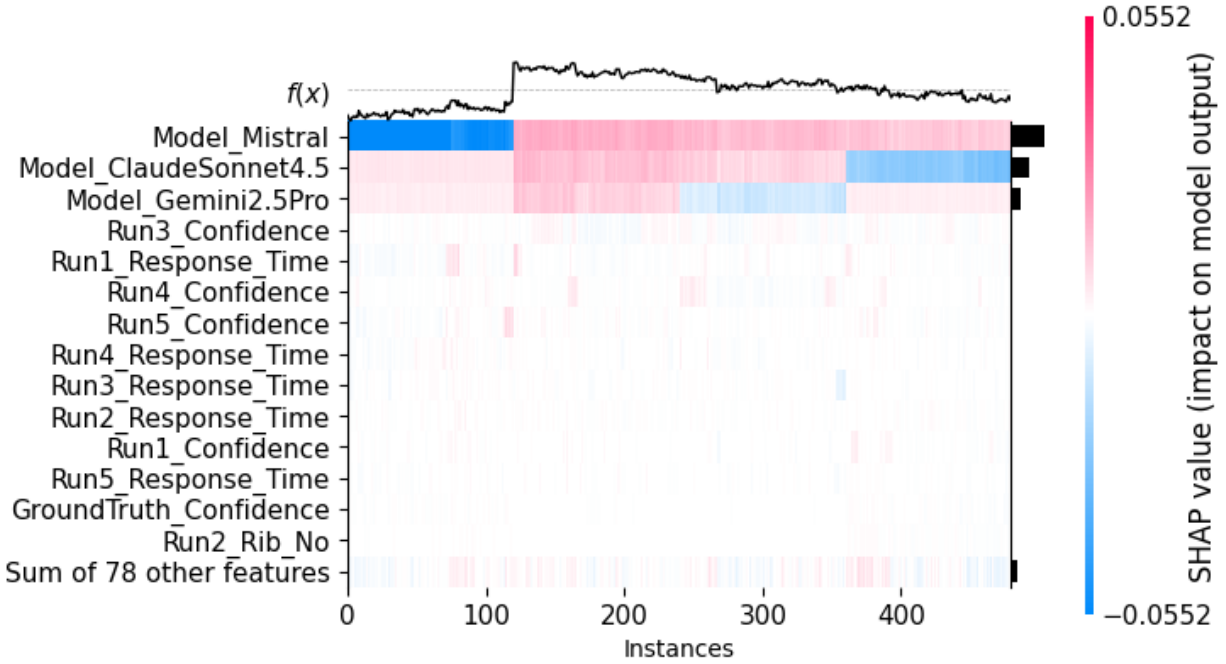


Figure 6. SHAP Heatmap Analysis

The SHAP heatmap in Figure 6 shows the contribution of features for all RibFrac CT cases. Confidence scores, response times and other clinical features have relatively limited impact on the prediction outcome, whereas the evaluated multimodal models have relatively larger impact. The difference in the values of the SHAPs across individual cases shows differences in the decisions made by the model. The red areas show features that have a positive contribution towards the prediction while the blue areas show the negative contribution. The entire heatmap gives an overall picture of the feature importance and the prediction behavior of the whole dataset on a global level.

Additionally, the experimental results show that ChatGPT-5 Vision outperforms all other models investigated across all the diagnostic accuracy, classification metrics, anatomical localization, laterality identification, clinical report quality, and SHAP-based explainability measures. The other top-performing models were Gemini 2.5 Pro and Claude Sonnet 4.5, with Mistral Multimodal performing comparatively poorly. These results demonstrate that the potential of an accurate and clinically reliable rib fracture detection from thoracic CT images using advanced multimodal vision–language models is promising.

5. Conclusion

This study aimed to perform a comparative study of four multimodal vision–language foundation models (GFM): ChatGPT-5 Vision, Gemini 2.5 Pro, Claude Sonnet 4.5 Vision, and Mistral Multimodal Vision to evaluate their performance in automated detection of rib fracture in 120 thoracic computed tomography (CT) images from the RibFrac dataset. The results indicated that

ChatGPT-5 Vision achieved the highest diagnostic accuracy, anatomical localization performance, clinical report quality and overall clinical score, while the SHAP analysis provided clear information on what factors influenced the model's predictions. The results highlight the promise of multimodal foundation models for supporting rib fracture diagnosis and radiology report generation and show that performance varies widely between the models' architectures. To sum up, the suggested assessment framework offers a comprehensive and powerful standard to evaluate the efficiency of diagnostic models, clinical reliability, and explainability, aiding in the incorporation of trustworthy AI systems into the future of thoracic trauma imaging.

References

- [1] N. Wu, J. Zhang, W. Zhao, J. Wang, C. Yu, Z. Xiu, D. Dai, Z. Xu, et al., “OrthoInsight: Rib fracture diagnosis and report generation based on multi-modal large models,” arXiv preprint arXiv:2507.13993, 2025, doi: 10.48550/arXiv.2507.13993.
- [2] R. C. Lee, R. Hadidchi, M. Coard, Y. Rubinov, T. T. Alamuri, A. Liaw, R. Chandrupatla, and T. Q. Duong, “Use of large language models on radiology reports: A scoping review,” *Journal of the American College of Radiology*, 2025, doi: 10.1016/j.jacr.2025.10.005.
- [3] Y. Artsi, E. Klang, J. Collins, B. Glicksberg, P. Korfiatis, G. Nadkarni, and V. Sorin, “Large language models in radiology reporting: A systematic review of performance, limitations, and clinical implications,” *medRxiv*, 2025, doi: 10.1101/2025.03.18.25324193.
- [4] B. Khosravi, T. Dapamede, F. Li, Z. M. Chisango, A. Bikmal, S. Garg, B. Owosela, A. Khosravi, et al., “Role of model size and prompting strategies in extracting labels from free-text radiology reports with open-source large language models,” *Journal of Imaging Informatics in Medicine*, 2025, doi: 10.1007/s10278-025-01505-7.
- [5] M. Mergen, D. Spitzl, C. Ketzer, M. Strenzke, A. Marka, M. R. Makowski, K. Bressemer, L. C. Adams, et al., “Leveraging large language models for accurate AO fracture classification from CT text reports,” *Journal of Imaging Informatics in Medicine*, 2025, doi: 10.1007/s10278-025-01603-6.
- [6] J. Liu, B. Koopman, N. J. Brown, K. Chu, and A. Nguyen, “Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports,” *Artificial Intelligence in Medicine*, vol. 158, Art. no. 103027, 2024, doi: 10.1016/j.artmed.2024.103027.
- [7] C. K. C. Ng, Z. Sun, and I. K. H. Te, “Performance of large language models for radiology report impression generation: A systematic review,” *Technologies*, vol. 14, no. 2, Art. no. 99, 2026, doi: 10.3390/technologies14020099.

- [8] S. Lee, J. Youn, H. Kim, M. Kim, and S. Yoon, “CXR-LLaVA: A multimodal large language model for interpreting chest X-ray images,” *European Radiology*, vol. 35, 2025, doi: 10.1007/s00330-024-11339-6.
- [9] S. Bannur, K. Bouzid, D. C. Castro, A. Schwaighofer, S. Bond-Taylor, M. Ilse, F. Pérez-García, V. Salvatelli, et al., “MAIRA-2: Grounded radiology report generation,” *arXiv preprint arXiv:2406.04449*, 2024, doi: 10.48550/arXiv.2406.04449.
- [10] D. Reichenpfader, H. Müller, and K. Denecke, “A scoping review of large language model based approaches for information extraction from radiology reports,” *npj Digital Medicine*, vol. 7, no. 1, Art. no. 224, 2024, doi: 10.1038/s41746-024-01219-0.
- [11] L. Zhang, M. Liu, L. Wang, Y. Zhang, X. Xu, Z. Pan, Y. Feng, J. Zhao, et al., “Constructing a large language model to generate impressions from findings in radiology reports,” *Radiology*, vol. 313, no. 3, Art. no. e240885, 2024, doi: 10.1148/radiol.240885.
- [12] B. Altalla, A. Ahmad, L. Bitar, M. O. Al-Bssol, A. A. Omari, and I. Sultan, “Radiology report annotation using generative large language models: Comparative analysis,” *International Journal of Biomedical Imaging*, vol. 2025, Art. no. 5019035, 2025, doi: 10.1155/ijbi/5019035.
- [13] F. Busch, L. Hoffmann, D. Pinto dos Santos, M. R. Makowski, L. Saba, P. Prucker, M. Hadamitzky, N. Navab, et al., “Large language models for structured reporting in radiology: Past, present, and future,” *European Radiology*, vol. 35, 2025, doi: 10.1007/s00330-024-11107-6.
- [14] T. Liu, J. Wang, Y. Hu, M. Li, J. Yi, X. Chang, J. Gao, and B. Yin, “HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation,” *arXiv preprint arXiv:2412.11070*, 2024, doi: 10.48550/arXiv.2412.11070.
- [15] Jiancheng Yang, Liang Jin, Bingbing Niand Ming Li, “RibFrac Dataset: A Benchmark for Rib Fracture Detection, Segmentation and Classification (Training Set Part 2)”. Zenodo, Jun. 15, 2020. doi: 10.5281/zenodo.3893498.