Comparative Analysis of Fine-tuned Multimodal Models in Radiology Image Captioning

Mahmudul Hoque
Computer Science Department
Morgan State University
Baltimore, USA
mahoq1@morgan.edu

Md Rakibul Hasan
Management Information Systems Dept.
University of Dhaka
Dhaka, Bangladesh
hasan.mr@du.ac.bd
mdhas1@morgan.edu

Md. Ismail Siddiqi Emon
Computer Science Department
Morgan State University
Baltimore, USA
mdemo1@morgan.edu

Ejiga Peter Ojonugwa Oluwafemi
Computer Science Department
Morgan State University
Baltimore, USA
ojeji1@morgan.edu

Md Mahmudur Rahman
Computer Science Department
Morgan State University
Baltimore, USA
md.rahman@morgan.edu

Fahmi Khalifa
Electrical & Computer Engineering Dept.
Morgan State University
Baltimore, USA
fahmi.khalifa@morgan.edu

2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI) | 979-8-3315-0913-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMI65310.2025.11141312

***Abstract*—Multimodal artificial intelligence (AI) models offer a promising approach for automatically generating captions from radiology images, bridging visual and textual data to enhance medical image interpretation. This study compares a selectively fine-tuned Large Multimodal Model (LMM) with a fully fine-tuned multimodal encoder–decoder architecture on a subset of the Radiology Objects in Context version 2 (ROCOv2) dataset. Specifically, Vision Generative Pre-trained Transformer 2 (VisionGPT-2) is evaluated against Large Language and Vision Assistant (LLaVA) version 1.6, a Mistral-7B-based LMM adapted using Low-Rank Adaptation (LoRA) on selected projection matrices. Experimental results show that LLaVA-1.6 outperforms VisionGPT-2 while significantly reducing the computational cost of full fine-tuning through selective adaptation with LoRA. Performance gains are observed across multiple metrics, including MedBERTScore (0.633), ClinicalBLEURT (0.456), ROUGE (0.251), BLEU-1 (0.209), BLEURT (0.317), METEOR (0.092), CIDEr (0.245), CLIPScore (0.816), RefCLIPScore (0.815), and BERTScore (0.628). These findings underscore the potential of selectively fine-tuned LMMs for efficient, semantically rich caption generation in medical imaging tasks, particularly in contexts constrained by sensitive data and limited computational resources.**

***Index Terms*—Large Multimodal Model, Transformer, Large Language and Vision Assistant, Mistral-7B, Medical Image Analysis, Low-Rank Adaptation, Muitlimodal Data, Radiology Image Captioning, Vision Generative Pre-Trained Transformer-2**

I. Introduction

The rapid progress of artificial intelligence (AI) has had a substantial impact on diverse domains, including medical image analysis. An essential application in medical domain is the automated generation of captions for radiology images. Recently, Large Multimodal Models (LMMs), which are extensions of Large Language Models (LLMs) [1], have shown considerable promise by incorporating a wide range of data modalities (e.g., text, images, video, and signal.) By addressing the limitations of text-only approaches, LMMs enable the interpretation of complex and heterogeneous medical data [2]. In LMMs, visual modalities are translated into text with semantic information, identifying objects, actions, and features [3]. These capabilities improve domain-specific content generation, making LMMs suitable for tasks like medical image captioning and radiology report generation [4]. Traditional encoder-decoder models (e.g., CLIP, BERT) also play a significant role in image captioning, alongside LMMs.

Medical image captioning studies integrate radiology and deep learning knowledge [5]. Early solutions, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), performed well for specific diseases and imaging modalities [6]. Later, Transformer-based architectures like Show-Attend-Tell and GPT-3 gained prominence [7]. The CDGPT2 model exemplifies radiology report generation from chest X-rays, fine-tuning ChexNet, and GPT-2 models with the IU-Xray dataset [8].

Recent benchmarks have tested multimodal models on two key tasks: Medical Visual Question Answering (MedVQA) and Medical Report Generation (MRG) [9]. MedVQA models show limitations in handling open-ended questions in both classification-based and generative tasks. Generative models like GPT-4v and Med-Flamingo offer broader application [10]. MRG models such as R2Gen and ITHN enhance medical report generation [11], demonstrating LMMs' superiority over standard language models in handling diverse data types.

LaCLIP employs LLaMA to refine captions but faces challenges with hallucinations arising from low-quality data [12]. Methods like filtering and combining raw and synthetic captions improve CLIP performance [13] [14] [15]. VIVO pre-training acquires unified visual-textual representations through image-tag pairs rather than image-caption pairs [16]. During

M. Hoque and M. R. Hasan have contributed equally to this work.

This work was supported by the National Science Foundation (NSF) grant (ID. 2131307) "CISE-MSI: DP: IIS: III: Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support."

Authorized licensed use limited to: Morgan State University. Downloaded on December 15,2025 at 18:43:28 UTC from IEEE Xplore. Restrictions apply.

pre-training, regions are annotated with labels; fine-tuning links images to phrases by identified objects, enabling zero-shot generalization. While vision-language pre-training excels in comprehension tasks, it often underperforms with generative tasks. XGPT addresses this by employing a cross-modal encoder-decoder for image captioning [17]. Recent research work leverage LLMs to enhance LMMs, incorporating pre-trained visual encoders. A common approach targets predicting text tokens [18], while some LMMs predict both image and text tokens, improving text and image production without compromising image-to-text performance. This capability stems from GPT-style auto-regressive training, enabling task-appropriate completions [19].

This study evaluates the performance of selectively fine-tuned LMMs (e.g., LLaVA-1.6, Mistral-7B) compared to fully fine-tuned encoder-decoder models (e.g., VisionGPT-2) using the LoRA technique on LLaVA's projection matrices. The results show that LMMs not only surpass the performance of traditional encoder-decoder models like VisionGPT-2 when selectively fine-tuned using LoRA, but also significantly reduce computational overhead compared to full model fine tuning.

II. Methodology

The primary focus of this work is to develop a framework for predicting and generating medical captions for radiology images utilizing multimodal models. Particularly, we explored and compared two distinct methodologies. The first approach uses the LLaVA-1.6-Mistral-7B model [20], fine-tuned using LoRA [21]. Focusing on a subset of parameters allows LoRA to efficiently update pre-trained models, decreasing computing cost, and improving caption accuracy alongside contextual relevance. In contrast, the second methodology utilizes the VisionGPT2 [22] model, a traditional multimodal encoder-decoder architecture designed to combine visual and textual data. VisionGPT2 serves as a benchmark to assess how traditional models perform in comparison to selectively fine-tuned LMMs, especially when dealing with complex medical image captioning tasks.

A. Large Language and Vision Assistant (LLaVA)

LLaVA [23]–[26] integrates a visual encoder and an LLM for joint visual and linguistic comprehension. The system is trained on many pairs of images, instructions, and extensively annotated captions. LLaVA uses ChatGPT/GPT-4 [27] to process image-caption pairings into instruction-following formats, generating visual questions ($\mathbf{X_q}$) from the input image and caption. This triggers context-aware image content descriptions from the language model.

In Figure 1, the LLaVA model connects a pre-trained CLIP-ViT-L/14 vision encoder to the Mistral-7B language model ($\mathbf{f}_\phi$). The encoder creates visual feature representations $\mathbf{Z_v} = \mathbf{g}(\mathbf{X_v})$ from input images, and a trainable multimodal projection matrix $\mathbf{W}$ aligns these with the LLM's language embedding space. The projection matrix transforms visual feature $\mathbf{Z_v}$ into a compatible format $\mathbf{H_v}$, matching the dimensionality of language model word embeddings. This alignment

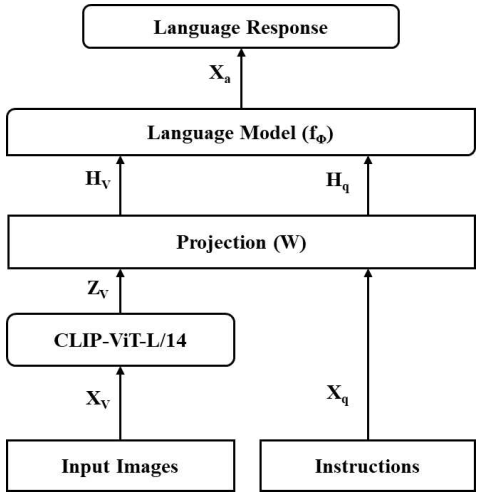

Fig. 1: Thematic view of the LLaVA model architecture.

allows the model to combine visual and textual data for meaningful captions or responses.

The final output, $\mathbf{X_a}$, represents the generated captions corresponding to the input images. This architecture enables LLaVA to leverage both visual and linguistic data to predict correct contextual captions. The projection matrix plays a pivotal role in facilitating the communication between the visual encoder and the LLM, ensuring that the model can perform multimodal tasks effectively.

LLaVA-1.5 [23] builds on this structure with modifications such as the CLIP-ViT-L-336px visual encoder and Multi-layer Perceptron (MLP) projection layers. It also introduces academic task-oriented Visual Question Answering (VQA) data featuring streamlined response formatting prompts. LLaVA-1.6 [24]–[26] further enhances this model by introducing capabilities for advanced reasoning, Optical Character Recognition (OCR), and domain-specific knowledge integration. This version includes key technical innovations that improve visual conversation capabilities.

B. Mistral-7B

Mistral-7B [28], [29] is a transformer based LLM built to deliver high performance with computational efficiency. Grouped-query attention (GQA) speeds up inference while sliding window attention (SWA) supports lengthier input sequences with low processing overhead. Together, GQA and SWA enable scalable processing across diverse text lengths without significantly increasing resource demands. The structural configuration of Mistral-7B is presented in Figure 2, highlighting its attention based module.

C. Low-Rank Adaptation

LoRA is a finetuning technique developed to efficiently adapt large models for specific tasks and datasets [21], [30]. LoRA addresses the challenge of computationally expensive fine-tuning by freezing the original model weights and introducing a smaller set of trainable parameters, that are added to the model through low-rank matrices. This process enable

Authorized licensed use limited to: Morgan State University. Downloaded on December 15,2025 at 18:43:28 UTC from IEEE Xplore. Restrictions apply.

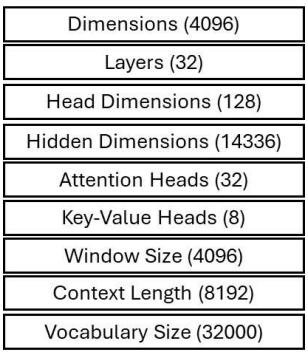

Fig. 2: Layer configuration of the Mistral-7B model.

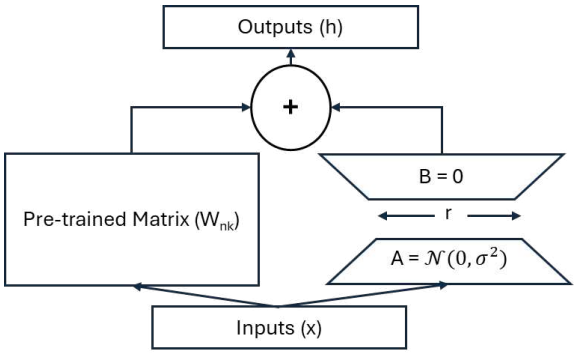

Fig. 3: LoRA method.

faster adaptation with fewer resources while preserving the original model's general capabilities. LoRA is particularly effective for adapting large models like LMMs for tasks such as medical image captioning, where only a small subset of features needs to be re-learned or emphasized [21].

At the core of LoRA is the assumption that the weight updates required during fine-tuning exhibit a low intrinsic rank. This means that the changes in weights, ΔW_{nk}, can be effectively modeled using a low-rank decomposition. Specifically, if W_{nk} represents the weights of a single layer and ΔW_{nk} represents the weight changes during adaptation, then ΔW_{nk} is approximated as a low-rank matrix [21], [30]. To achieve this, LoRA introduces two small trainable matrices, A and B, into each fine-tuned layer. These matrices first project input features to a lower-dimensional space and then map them back to the original space. Formally, ΔW_{nk} is the product of two low-rank matrices: $\Delta W_{nk} \approx B_{nr} A_{rk}, \quad \text{where } r \ll \min(n, k)$. The layer's forward pass is adjusted to $Wx + BAx$, where W is the original weight matrix and B and A are the low-rank matrices. Initially, B is set to zero and A is randomly initialized, ensuring that the adaptation process starts without altering the original pre-trained weights. The update matrix BA is scaled by a factor of α/r, controlling the magnitude of the updates during fine-tuning [21], [30]. Figure 3 demonstrates the techniques adopted in LoRA.

D. LLaVA-1.6-Mistral-7B Fine-Tuning with LoRA

The LLaVA-1.6-Mistral-7B model integrates the CLIP-ViT-L/14 vision encoder, which is a 14-layer Vision Transformer, with the Mistral-7B language model for text generation. In this study, LoRA was employed to fine-tune selected layers of the combined architecture. In the proposed analysis pipeline, LoRA was applied to the following layer-specific projection matrices:

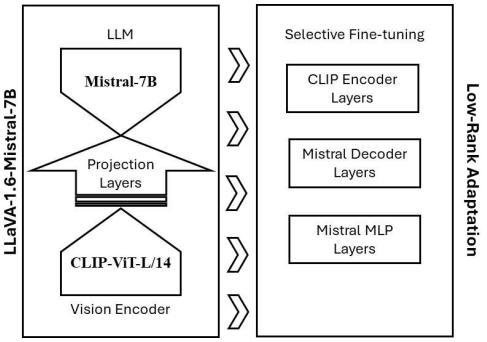

Fig. 4: Thematic view of the LLaVA–1.6-Mistral-7B approach.

- CLIP encoder: Key, value, and query projection layers.
- Mistral decoder: Query, key, and value projection layers.
- Mistral MLP: Gate, up, and down projection layers.

The final output represents the generated captions for the input images, where the integration of visual features from the CLIP encoder and language features from Mistral-7B ensures accurate and semantically rich descriptions. Figure 4 illustrates the architecture of the LLaVA-1.6-Mistral-7B model, along with the implementation of LoRA in various layers.

E. VisionGPT-2

VisionGPT-2 adopts a sequential encoder-decoder framework in which a pre-trained ViT [31], [32] extracts visual representations from input images. The features are then forwarded to a GPT-2 language model [33]–[35], which serves as the decoder for caption generation. Unlike LLaVA, which uses a projection matrix to align multimodal embeddings, VisionGPT-2 performs direct feature handoff from the visual encoder to the text decoder, facilitating a modular training structure. Training was conducted using the Hugging Face `Seq2SeqTrainer` API [36], which enabled progressive end-to-end fine-tuning of the entire architecture. The decoder incorporates a multi-head cross-attention mechanism to incorporate visual features into the captioning process. Figure 5 illustrates the structure of the encoder and decoder blocks used in this implementation.

III. Results and Discussions

The dataset used in this research consisted of radiology image-text pairs from ROCOv2 [37], [38], an updated version of the ROCO dataset [39], collected from biomedical articles in PMC OpenAccess. It includes six imaging modalities (X-ray, CT, MRI, ultrasound, PET, and combinations) and relevant anatomical regions. The training set contains 70,108 images and 220,859 concept occurrences across 1,945 unique concepts, while validation and test sets have 9,972 and 17,237 images with 32,060 and 48,563 occurrences, respectively. Average image dimensions are 600×600, requiring resizing and padding for smaller images. Captions are mostly under

Authorized licensed use limited to: Morgan State University. Downloaded on December 15,2025 at 18:43:28 UTC from IEEE Xplore. Restrictions apply.

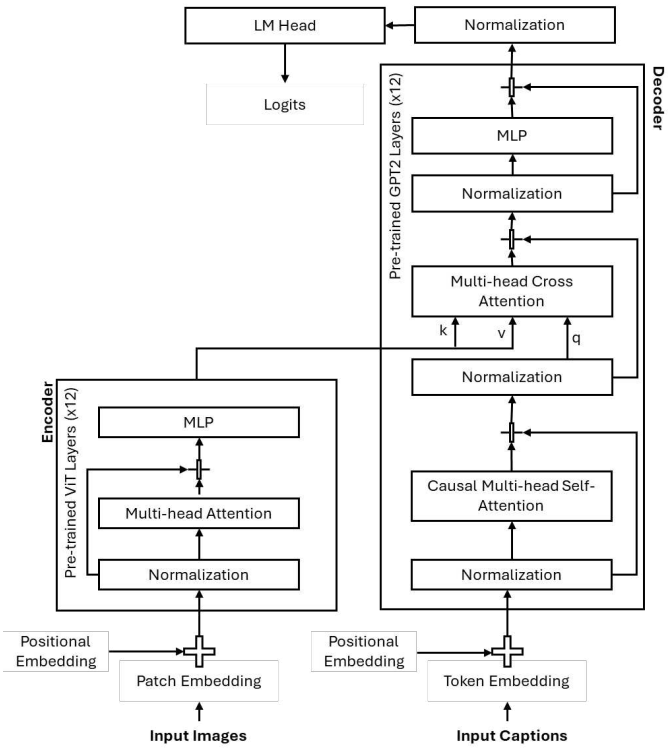

Fig. 5: Thematic view of the VisionGPT-2 approach.

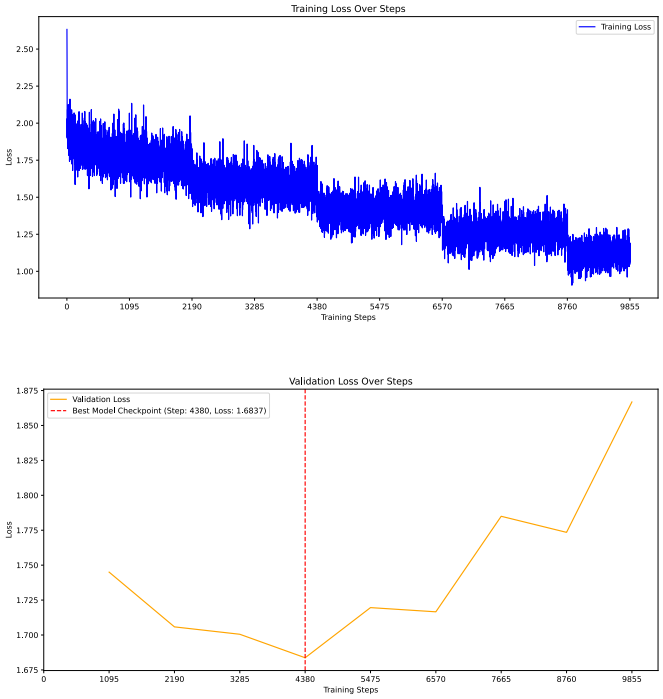

Fig. 6: Training & validation loss of LLaVA-1.6-Mistral-7B.

100 words. Models were trained on a Google Cloud VM with 2xNVIDIA-A100-40GB GPUs, 24 vCPUs, and 170 GB RAM. The reproducible codes of the corresponding models and their implementations can be found in [40].

LoRA was applied during training with a rank (r) of 16, scaling factor of 32, and a 0.05 dropout rate. This configuration led to 40,108,032 trainable parameters, accounting for only 0.52% of the total number of parameters. Key hyperparameters included a learning rate of 1e-4, mixed precision using *bfloat16*, the `adamw_torch` optimizer, and a batch size of 4. The model was trained for 10 epochs across 21,900 steps, incorporating early stopping with a patience of 5 steps. Training was halted at the 9855th step, while the model checkpoint corresponding to the lowest validation loss was saved at the 4380th step. Figure 6 presents the progression of training and validation losses. Additionally, Flash Attention-2 [41] was applied to optimize attention mechanism computations further. Experimental evident based on literature clearly indicates that FlashAttention-2 can achieve up to a 2× speedup compared to FlashAttention-1 and up to a 9× speedup over PyTorch's standard attention implementation [41]. Therefore, MistralFlashAttention2 is used within the 32 Mistral decoder layers in this model. Finally, the approach of applying bfloat16 precision along with MistralFlashAttention2 ensures the multimodal capability of the model in handling both visual and textual features.

In the VisionGPT-2 training phase, the pre-trained layers were initially kept frozen to focus on updating the cross-attention layers. After completing the tuning of GPT-2, the ViT encoder was incrementally unfrozen and trained during the later epochs. The training process utilized the Adam optimizer [42] along with a OneCycleLR scheduler for learning rate adjustment. Mixed-precision computation (*fp16*) was applied using PyTorch's autocast and gradient scaler to improve memory usage and computational speed. Cross-entropy loss served as the evaluation metric, and the model checkpoint with the lowest validation perplexity was retained for downstream caption generation. The VisionGPT-2 configuration consisted of 12 transformer layers, 12 attention heads, 768-dimensional embeddings, variable dropout settings, and an MLP ratio of 4. Training was executed on a CUDA-enabled GPU using a batch size of 8 over five epochs. The evolution of training and validation losses is shown in Figure 7.

The trained models were evaluated based on the BERTScore [43], ROUGE [44], BLEU-1 [45], BLEURT [46], METEOR [47], CIDEr [48], CLIPScore [49], RefCLIPScore [49], ClinicalBLEURT [50], and MedBERTScore [50] performance measurement metrics. The results section provides a detailed comparison of these metrics for the LLaVA-1.6-Mistral-7B and VisionGPT-2 models. These metrics are critical in evaluating the effectiveness of models in generating accurate and meaningful captions for medical images. Detailed analysis of the performance metrics are summarized in Table I. This table highlights the consistent superiority of the LLaVA-1.6-Mistral-7B over VisionGPT-2 across most evaluation metrics. This superiority suggests that LLaVA-1.6-Mistral-7B is more adept at generating high-quality, semantically accurate captions from the radiology medical images.

Figure 8 shows a comparative scenario of the predictive per-

Authorized licensed use limited to: Morgan State University. Downloaded on December 15,2025 at 18:43:28 UTC from IEEE Xplore. Restrictions apply.

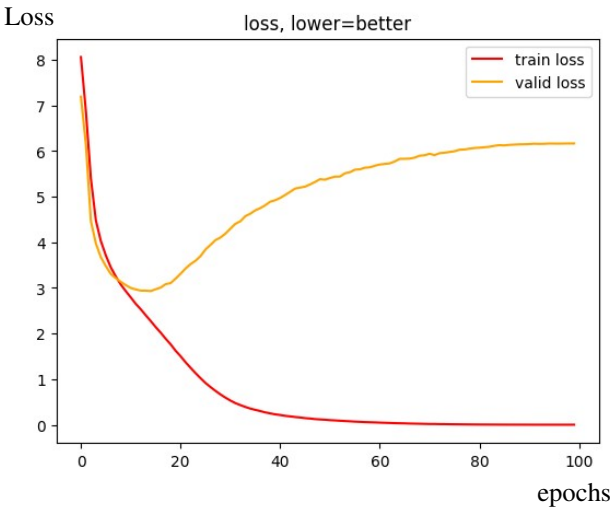

Fig. 7: Training & validation loss of VisionGPT2.

TABLE I: Performance scores and relative gains of LLaVA-1.6-Mistral-7B over VisionGPT-2.

Performance Metrics	LLaVA-1.6-Mistral-7B	VisionGPT-2
CLIPScore	**0.821262** (+0.136)	0.685257
MedBERTScore	**0.632664** (+0.076)	0.556308
BERTScore	**0.628059** (+0.082)	0.545773
RefCLIPScore	**0.815534** (+0.123)	0.692432
ClinicalBLEURT	**0.455942** (+0.006)	0.450116
BLEURT	**0.317385** (+0.073)	0.244725
CIDEr	**0.245029** (+0.216)	0.028805
ROUGE	**0.250801** (+0.132)	0.118446
BLEU-1	**0.209298** (+0.107)	0.102386
METEOR	**0.092682** (+0.058)	0.035134

formance of the two models using a representative radiology image from the ImageCLEFmedical'24 test set [38]. The figure displays both the actual and predicted captions, highlighting the stronger performance of the LLaVA-1.6-Mistral-7B model with LoRA adaptation compared to VisionGPT-2.

IV. Conclusions and Future Research

This work explores the application of an LMM in medical image caption generation task and evaluated its performance in comparison to a pre-trained Transformer model using the ImageCLEF2024 caption dataset. The experimental results of our study reveal that LMMs (LLaVA-1.6- Mistral-7B model) are more effective than traditional Transformer models (VisionGPT-2) in generating clinically relevant image captions. These findings indicate a promising avenue for future research, highlighting the potential of open-source LMMs in medical image interpretation tasks. Morover, the models like Med-Flamingo [51], MedBLIP [52], and HuatuoGPT-Vision [53] are the recent advancements made in analyzing medical image-text analysis. In our incremental work, we are also exploring the potential of these models in improving predictive accuracy. It is believed that the dataset's size and diverse image modalities are adequate for comparative analysis in this

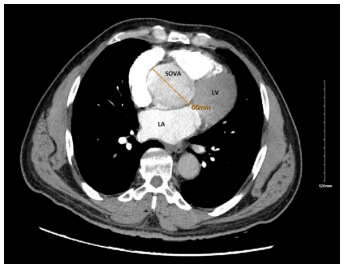

Fig. 8: Comparison of model generated captions on a representative chest CT image *(CC BY [Muacevic et al. (2023)])* [38]. **Predicted Caption (LLaVA):** *"Computed tomography angiography of the chest showing a large saccular aneurysm of the left ventricle measuring 0.8 cm × 1.5 cm."* **Predicted Caption (VisionGPT-2):** *"CT scan of the chest showing a right upper lobe nodule (red arrow) with a moderate sized right upper lobe nodule (red arrow)."* **Ground Truth:** *"CT chest revealing 6 cm dilated sinus of Valsalva aneurysm. LA: Left Atrium; LV: Left Ventricle; SOVA: Sinus of Valsalva Aneurysm."*

study, but larger datasets could improve results. Therefore, the initiative for exploring additional data sources in the future will remain active. Furthermore, more rigorous image denoising, enhancement, and edge detection techniques will be explored in search for any scope of improving the model performance.

In future work, we plan to involve clinicians, particularly in evaluating the clinical relevance of the generated captions. We believe that incorporating clinician feedback will provide valuable insights into how the models can be refined for practical clinical use.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[2] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ digital medicine*, vol. 5, no. 1, p. 149, 2022.

[3] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4467–4480, 2019.

[4] M. Cui, C. Li, and Y. Yang, "Explicit image caption reasoning: Generating accurate and informative captions for complex scenes with lmm," *Sensors*, vol. 24, no. 12, p. 3820, 2024.

[5] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artificial Intelligence in Medicine*, vol. 106, p. 101878, 2020.

[6] T. do Carmo Nogueira, C. D. N. Vinhal, G. da Cruz Júnior, and M. R. D. Ullmann, "Reference-based model using multimodal gated recurrent units for image captioning," *Multimedia Tools and Applications*, vol. 79, pp. 30 615–30 635, 2020.

[7] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, "Medical image captioning via generative pretrained transformers," *Scientific Reports*, vol. 13, no. 1, p. 4171, 2023.

[8] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021.

Authorized licensed use limited to: Morgan State University. Downloaded on December 15,2025 at 18:43:28 UTC from IEEE Xplore. Restrictions apply.

[9] G. Liu, J. He, P. Li, G. He, Z. Chen, and S. Zhong, "Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging," 2024.

[10] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[11] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *arXiv preprint arXiv:2010.16056*, 2020.

[12] L. Fan, D. Krishnan, P. Isola, D. Katabi, and Y. Tian, "Improving clip training with language rewrites," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] T. Nguyen, S. Y. Gadre, G. Ilharco, S. Oh, and L. Schmidt, "Improving multimodal datasets with image captioning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

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

[16] X. Hu, X. Yin, K. Lin, L. Zhang, J. Gao, L. Wang, and Z. Liu, "Vivo: Visual vocabulary pre-training for novel object captioning," in *proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1575–1583.

[17] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou, "Xgpt: Cross-modal generative pre-training for image captioning," in *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*. Springer, 2021, pp. 786–797.

[18] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra *et al.*, "Language is not all you need: Aligning perception with language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] A. Radford, "Improving language understanding by generative pre-training," 2018.

[20] H. Liu, "Llava-v1.6-mistral-7b," https://huggingface.co/liuhaotian/llava-v1.6-mistral-7b, 2023, open-source chatbot trained by fine-tuning LLM on multimodal instruction-following data.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," 2021.

[22] S. Dan, "Visiongpt2 image captioning — pytorch," https://www.kaggle.com/code/shreydan/visiongpt2-image-captioning-pytorch, 2023, kaggle notebook demonstrating image captioning using VisionGPT2 and PyTorch.

[23] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[24] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[25] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[26] LLaVA Team, "Llava 1.6: Enhanced reasoning, ocr, and world knowledge," https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024, accessed: 2024-06-08.

[27] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, and et al., "Gpt-4 technical report," 2024.

[28] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[29] M. AI, "Mistral inference," https://github.com/mistralai/mistral-inference, 2024, accessed: 2024-06-08.

[30] Y. Zeng and K. Lee, "The Expressive Power of Low-Rank Adaptation," 2024.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[32] H. Face, "Vision transformer (vit) documentation," https://huggingface.co/docs/transformers/en/model_doc/vit, 2024, accessed: 2024-06-12.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[34] H. Face, "Gpt-2 model card," https://huggingface.co/openai-community/gpt2, 2024, accessed: 2024-06-12.

[35] ——, "Gpt-2 documentation," https://huggingface.co/docs/transformers/en/model_doc/gpt2, 2024, accessed: 2024-06-12.

[36] ——, "Trainer documentation," https://huggingface.co/docs/transformers/main_classes/trainer, 2024, accessed: 2024-06-12.

[37] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, and C. M. Friedrich, "ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset," 2024. [Online]. Available: https://arxiv.org/abs/2405.10004v1

[38] B. Ionescu *et al.*, "Overview of ImageCLEF 2024: Multimedia retrieval in medical applications," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ser. Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024). Grenoble, France: Springer Lecture Notes in Computer Science LNCS, September 9-12 2024.

[39] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (roco): a multimodal image dataset," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer, 2018, pp. 180–189.

[40] M. Hoque and M. R. Hasan, "Medical image interpretation with large multimodal models," https://github.com/HoqueMahmudul/Medical-Image-Interpretation-with-Large-Multimodal-Models, 2023, accessed: 2024-06-19.

[41] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations (ICLR)*, 2024.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," 2020.

[44] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[46] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," 2020.

[47] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[48] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[49] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022.

[50] A. B. Abacha, W. wai Yim, G. Michalopoulos, and T. Lin, "An investigation of evaluation metrics for automated medical note generation," 2023.

[51] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec, "Med-flamingo: a multimodal medical few-shot learner," 2023. [Online]. Available: https://arxiv.org/abs/2307.15189

[52] V. T. Phan, K. T. Nguyen, A. D. D. Q. Hoang, T. Q. Phan, and B. T. N. Tat, "Medblip: Multimodal medical image captioning using blip," EasyChair Preprint 13501, EasyChair, 2024.

[53] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. H. Chen, X. Wang, R. Zhang, Z. Cai, K. Ji, G. Yu, X. Wan, and B. Wang, "Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale," 2024. [Online]. Available: https://arxiv.org/abs/2406.19280

Authorized licensed use limited to: Morgan State University. Downloaded on December 15,2025 at 18:43:28 UTC from IEEE Xplore. Restrictions apply.