**MULTIMODAL LLM FOR REPORT GENERATION IN HEALTHCARE**

**Abstract**

This study focuses on building a Multimodal Large Language Model (MLLM) to automate clinical report generation by combining medical images, clinical notes, and structured healthcare data. Traditional unimodal approaches often lead to fragmented insights, thus, we need a models that can interpret many types of healthcare data combined to enhance diagnostic accuracy and reduce the physician’s workload.

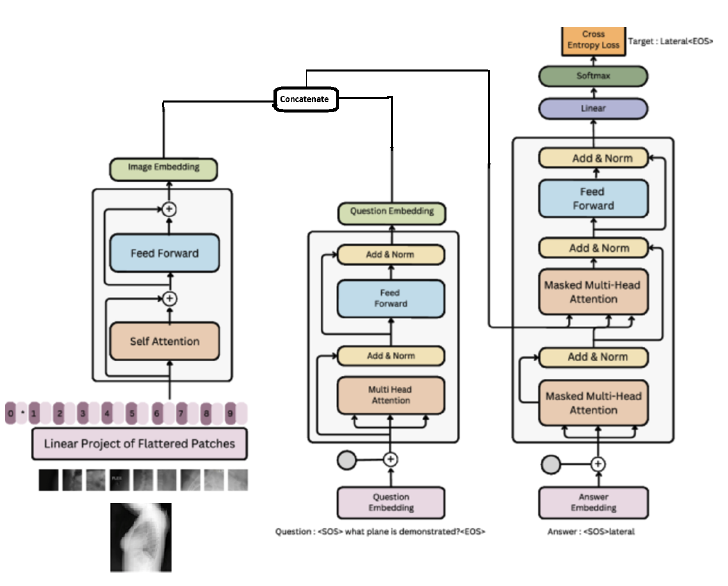
Earlier models like CLIP and BLIP lack precise healthcare adaptations. This research proposes using a Vision Transformer for image features and a transformer LLM for clinical text, fused via cross-attention and enhanced using retrieval-augmented generation.

Fine-tuning will leverage datasets such as MIMIC-CXR and IU-Xray. Evaluation will use BLEU, ROUGE, and Meteor metrics. The outcome aims for accurate, context-rich clinical reports and reduced clinician burden.

**References**

1. **1. Park, K., Sayres, R., Sellergren, A., Pollard, T., Jamil, F., Kohlberger, T., Lau, C. and Kiraly, A.**  
   Application of Med-PaLM 2 in the refinement of MIMIC-CXR labels. *(Multimodal LLM for chest X-ray dataset.)*
2. **2. Van, M.H., Verma, P. and Wu, X., 2024.**  
   On large visual language models for medical imaging analysis: An empirical study. In: *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp.172–176. IEEE. *(Study on VLMs applied to medical images.)*
3. **3. Mohsan, M.M., Akram, M.U., Rasool, G., Alghamdi, N.S., Baqai, M.A.A. and Abbas, M., 2022.**  
   Vision transformer and language model based radiology report generation. *IEEE Access*, 11, pp.1814–1824. *(Vision Transformer + LLMs for radiology reports.)*
4. **4. Ho, T.C., Kharrat, F., Abid, A., Karray, F. and Koubaa, A., 2024.**  
   Remoni: An autonomous system integrating wearables and multimodal large language models for enhanced remote health monitoring. In: *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp.1–6. IEEE. *(Multimodal LLMs integrating wearable data and medical images.)*
5. **5. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., 2013.**  
   OverFeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint* arXiv:1312.6229. *(Important for multimodal VLM foundations.)*
6. **6. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021.**  
   Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.10012–10022. *(Modern backbone for multimodal vision-language models.)*
7. **7. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., et al., 2024.**  
   Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4), pp.1134–1142. *(Highly relevant to medical LLMs working on clinical multimodal tasks.)*
8. **8. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M. and Dufour, R., 2024.**  
   BioMistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint* arXiv:2402.10373. *(Multimodal LLM foundation models for medical use.)*
9. **9. Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., et al., 2023.**  
   HuatuoGPT, towards taming language model to be a doctor. *arXiv preprint* arXiv:2305.15075. *(Training LLMs toward doctor-level reasoning, applied to multimodal data.)*
10. **10. Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., et al., 2024.**  
    Almanac—Retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2), p.AIoa2300068. *(RAG + LLM systems for clinical multimodal tasks.)*

**High-Level Block Diagram (Flow Description)**

****

**Abstract**

This work presents a Multimodal Large Language Model (MLLM) designed to automate clinical report generation by combining medical images, clinical notes, and structured health data. Traditional single-modal approaches often result in incomplete insights; thus, integrating multiple data types is critical for improving diagnostic precision and easing clinician workload. The model architecture incorporates a Vision Transformer for extracting image features and a transformer-based LLM for processing clinical text, connected through cross-attention and enhanced by retrieval-augmented generation techniques. Training will be conducted on datasets like MIMIC-CXR and IU-Xray, with performance assessed using BLEU, ROUGE, and Meteor scores. The goal is to generate accurate, contextually rich clinical reports while minimizing the burden on healthcare providers.

**SCI/Scopus Indexed Journals List**

1. IEEE Transactions on Medical Imaging
2. Artificial Intelligence in Medicine
3. Medical Image Analysis
4. Journal of Biomedical Informatics
5. Computer Methods and Programs in Biomedicine
6. Nature Communications (Healthcare AI Section)
7. Frontiers in Artificial Intelligence - Healthcare Section
8. Computers in Biology and Medicine
9. Scientific Reports (Healthcare Research)
10. Sensors (AI Applications in Healthcare Special Issue)