Multimodal Carotid Risk Stratification with Large Vision-Language Models: Benchmarking, Fine-Tuning, and Clinical Insights

1st Daphne Tsolissou
National Technical University of Athens
and
Archimedes, Athena Research Center
Athens, Greece
dtsolisou@biosim.ntua.gr

2nd Theofanis Ganitidis
National Technical University of Athens
Athens, Greece
orcid.org/0009-0006-7794-9793

3rd Konstantinos Mitsis
National Technical University of Athens
Athens, Greece
orcid.org/0000-0002-4629-2163

4th Stergios Christodoulidis
CentraleSupélec, Université Paris-Saclay
Paris, France
orcid.org/0000-0002-8773-1070

5th Maria Vakalopoulou
CentraleSupélec, Université Paris-Saclay
and
Archimedes, Athena Research Center
Paris, France
maria.vakalopoulou@centralesupelec.fr

6th Konstantina Nikita
National Technical University of Athens
Athens, Greece
orcid.org/0000-0001-8255-4354

2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-5899-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBE66822.2025.00122

***Abstract*—Reliable risk assessment for carotid atheromatous disease requires integrating diverse clinical and imaging information in a transparent and interpretable manner. This study investigates state-of-the-art large vision–language models (LVLMs) for multimodal carotid plaque assessment by integrating ultrasound imaging with structured clinical, demographic, laboratory, and biomarker data. A framework that simulates realistic diagnostic scenarios through interview-style question sequences is proposed, comparing open-source LVLMs including general-purpose and medically tuned models. Zero-shot experiments reveal that while most LVLMs accurately identify imaging modality and anatomy, all perform poorly in risk stratification. To address this, LLaVa-NeXT-Vicuna is adapted using low-rank adaptation (LoRA), achieving substantial improvements in stroke risk stratification. Integrating multimodal tabular data as text further enhances specificity and balanced accuracy, yielding competitive performance compared to prior CNN baselines. Our findings highlight both promise and limitations of LVLMs in ultrasound-based cardiovascular risk prediction, underscoring the importance of multimodal integration, domain adaptation, and model calibration for clinical translation.**

***Index Terms*—carotid ultrasound, risk stratification, large vision-language models, zero-shot evaluation, model adaptation**

I. Introduction

Recent advances in Artificial Intelligence (AI) have led to the emergence of Large Vision-Language Models (LVLMs), multimodal systems that jointly process images and text [1]. Typically combining Vision Transformers (ViT) [2] for image encoding with a Large Language Model (LLM) for text processing, LVLMs enable cross-modal reasoning. Models such as LLaVa-NeXT [3] achieve strong zero-shot performance on diverse natural image tasks including captioning, visual question answering (VQA), and report generation. Their reasoning ability can be further enhanced through prompting strategies such as chain-of-thought (COT), which elicit stepwise explanations [4], [5].

In medicine, LVLMs show growing potential for integrating multimodal clinical data, including imaging, laboratory, and demographic information, to support diagnostic reasoning [6]. However, their application to ultrasound imaging (USI) remains limited. USI presents unique challenges due to operator dependency, image variability, and the lack of large paired image–text datasets [7], [8]. These constraints hinder the transfer of knowledge from natural-image pretraining and emphasize the need to evaluate how current LVLMs generalize to this domain.

Several efforts have addressed these challenges. Open-source LVLMs such as LLaVa [3], [9]–[11] and PaliGemma [12] rival proprietary systems like GPT-4V [13] and Gemini [14], while medical variants, including MedGemma [15] and LLaVa-Med [16], introduce domain-specific tuning. QUILT-LLaVA [17] extends this direction to histopathology, incorporating spatial grounding. However, ultrasound-focused LVLMs remain scarce, with LLAUS [7] among the few instruction-tuned models available.

Benchmarking efforts such as GMAI-MMBench [18] and U2-Bench [8] provide valuable evaluations of medical LVLMs, though ultrasound coverage and specialized tasks like carotid plaque risk stratification remain limited. This is critical, as cardiovascular diseases (CVD) are a leading global cause of

This work was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union under NextGenerationEU and administered by the Archimedes Unit of the Athena Research Center.

2471-7819/25/$31.00 ©2025 IEEE
DOI 10.1109/BIBE66822.2025.00122

mortality, with carotid atheromatous plaques being a key source of ischemic stroke [19]. Atherosclerosis involves progressive arterial inflammation and plaque formation, which can destabilize and rupture, leading to cerebrovascular events such as stroke [20], [21]. Traditional approaches rely on single-modality imaging (USI or CT angiography) combined with risk factors [22]–[24], however the complexity of the disease requires the integration of multimodal data for robust, interpretable predictions [25]–[27].

This study evaluates state-of-the-art LVLMs for vulnerable carotid plaque detection from multimodal data, integrating USI with clinical, demographic, laboratory, and protein analysis information. Using open-source models such as LLaVA-Next [3], we design tailored prompts that simulate realistic diagnostic scenarios by incorporating patient-specific context. Our evaluation framework (Fig. 1) progresses from interviewing-style prompts that assess visual understanding and reasoning to stroke risk stratification tasks in both zero-shot and fine-tuned settings. The main contributions are: 1) a comprehensive evaluation of general-purpose and medically tuned LVLMs for carotid plaque risk stratification, and 2) task-specific fine-tuning of LLaVa-NeXT-Vicuna on carotid USI data, addressing the gap in domain-adapted LVLMs for ultrasound-based cardiovascular risk assessment.

II. Methods

A. Dataset and image strategy

An anonymized dataset was used from Attikon General University Hospital of Athens, comprising B-mode carotid USI videos from 72 patients, expertly annotated as high (59) or low (13) risk, with accompanying clinical and laboratory data (demographics, medications, blood tests, and protein analysis).

Two sampling strategies generated image-text pairs: 1) **Random Sampling:** five frames per video, sampled from equal temporal segments to capture cardiac cycle variability, used for zero-shot evaluation. 2) **Full-frame Sequence:** all frames used for fine-tuning with stratified 3-fold cross-validation at the patient-level to prevent data leakage and address frame-count imbalance.

B. Prompt strategies for LVLMs

Three types of text prompts were designed to simulate clinical reasoning:

1) **Context-Free Interview**: An image-only prompt that asked models to describe imaging modality, anatomy, and diagnostic features (Fig. 2).
2) **Imaging Context**: Image-only prompt that also included class definitions for risk stratification.
3) **Image and Tabular Context**: A multimodal prompt that extended the imaging context prompt by appending patient-specific clinical and laboratory data in free-text form.

All image-text pairs were tokenized using Hugging Face [28] preprocessing modules per model requirements.

C. Benchmarking of LVLMs

Six recent open-source LVLMs were benchmarked on zero-shot visual reasoning and risk stratification tasks using randomly sampled frames and the mentioned prompt. Three zero-shot tasks were evaluated: 1) imaging modality recognition, 2) anatomy identification, and 3) stroke risk classification.

The following general purpose, trained primarily on natural image–text pairs were used: *Paligemma-Mix (3B) [29], Paligemma2-Mix (3B) [12], LLaVa-NeXT Vicuna (7B and 13B), LLaVa-NeXT Mistral (7B) [3]*, selected for their strong vision-language benchmark performance comparable to *GPT-4V* [13]. Additionally, *MedGemma (4B)* [15], a medically tuned foundation model with strong medical imaging performance [30], [31] and no ultrasound training data [30], was included for unbiased domain assessment.

D. LVLM Adaptation to USI data and Carotid Risk Stratification

Following zero-shot evaluation, the best-performing general-purpose model, *LLaVa-NeXT Vicuna (7B)*, was adapted for risk stratification using LoRA [32]. This adaptation strategy, combining a general-purpose foundation model with parameter-efficient fine-tuning, was designed to maintain broad visual-language capabilities while specializing to the ultrasound domain, thereby promoting generalization to unseen examples.

The LLaVa-Next Vicuna (7B) model was adapted in two scenarios: 1) single-modality risk stratification using only image data as input and, 2) multimodal risk stratification combining image and tabular data as input. In both cases, the full frame sequences of each patient were used. The aim was to investigate whether a LVLM can leverage multimodal datasets that combine images with structured patient information.

Training was performed using the official `LLaVaTrainer` module (available on GitHub[1]) with LoRA rank $r = 128$ and alpha $\alpha = 128$ and default hyperparameters. Since LLaVa is transformer-based, low-rank matrices were inserted into the attention projection matrices of the query and value components (W_Q, W_V) in both the vision encoder and the language model. Images were resized and padded to 336×336 pixels using the default LLaVa preprocessing pipeline.

The AdamW [33] optimizer was used with a learning rate of $1\mathrm{e}{-5}$, a warm-up ratio of 3%, and a cosine learning rate scheduler. In LLaVA models, a Multilayer Perceptron (MLP) cross-modal connector aligns image and text features; this module was trained with a higher learning rate of $2\mathrm{e}{-4}$ to enable faster adaptation to the USI domain. Training employed the standard cross-entropy loss over output tokens. Given the iteration-based nature of the `LLaVaTrainer` and the sharp decrease in training loss observed during initial training, we stopped training after 460 iterations to prevent overfitting, as no validation set could be integrated into the training loop. Model selection followed a patient-level stratified 3-fold cross-validation scheme.

[1]https://github.com/haotian-liu/LLaVA

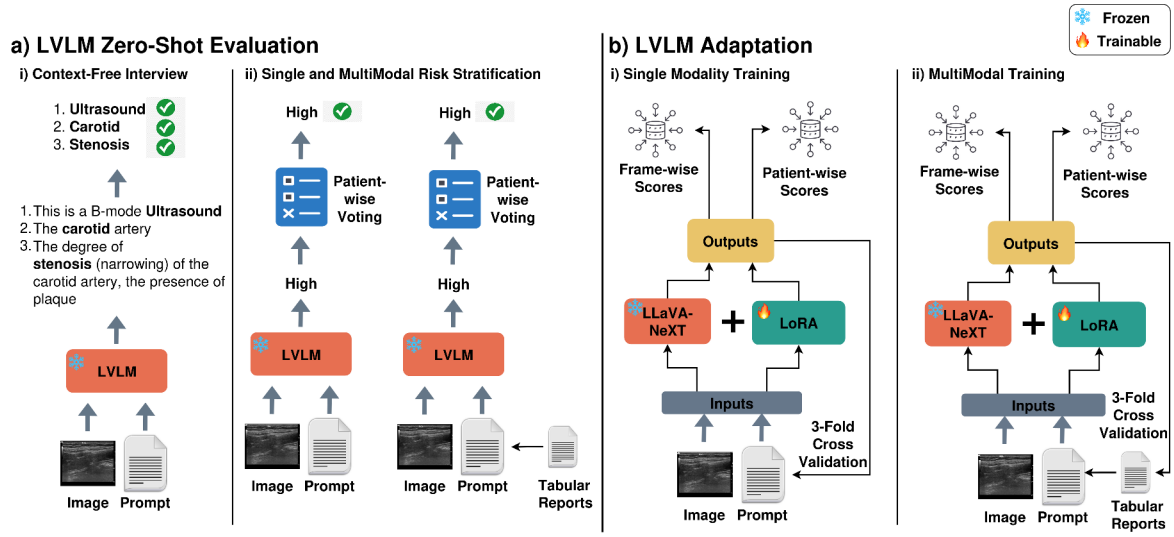

Fig. 1: Overview of methods applied in this study: 1) Zero-shot evaluation of LVLMs on medical reasoning and risk stratification using single-modality (ultrasound) and multimodal inputs (ultrasound with clinical, demographic, laboratory, and biomarker data); 2) LoRA adaptation of LLaVA-NeXT with 3-fold cross-validation in both settings.

E. Evaluation Protocol

LVLMs generate categorical textual responses, not calibrated probabilities, so only threshold-free metrics are directly applicable. Interview answers were parsed using regular expressions to identify relevant keywords. For risk stratification, textual responses were converted to binary labels and evaluated using metrics suitable for imbalanced clinical datasets: Sensitivity (Sens.), Specificity (Spec.), Balanced Accuracy (BAcc) as well as Mathews Correlation Coefficient (MCC), and Area Under the Curve (AUC). Since frame-level predictions are hard class assignments, patient-level predictions used majority voting across frames. The mean vote fraction was used as a pseudo-probability estimate for AUC calculation. This method serves as a proxy and does not reflect true probabilistic confidence.

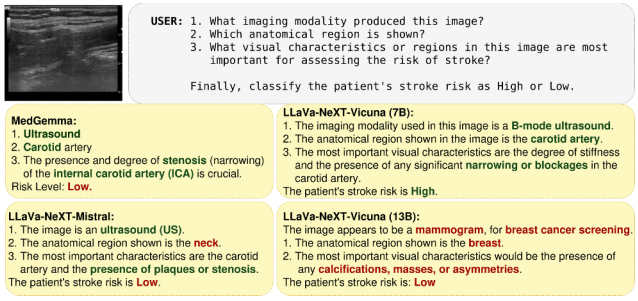

Fig. 2: Example of the context-free interview prompt. Models receive only the image and interview-style text, with responses illustrating zero-shot medical reasoning ability.

III. Results

A. Zero-Shot Evaluation

1) Imaging Modality and Anatomy Detection: Table I presents modality and anatomy identification results. MedGemma, correctly identified USI in all cases and the carotid artery in 98.33% of cases, outperforming all general-purpose models. LLaVa-NeXT-Vicuna-7B achieved the highest accuracy, correctly identifying USI in 95% of cases and the carotid artery in 91.94% of cases. LLaVa-NeXT-Mistral also detected USI correctly in 95% of cases but identified the neck rather than the carotid artery in 94.16% of cases. The larger LLaVa-NeXT-Vicuna-13B consistently failed to detect the correct modality. The Paligemma models proved unsuitable for zero-shot USI tasks, exhibiting extreme sensitivity to prompt wording.

Qualitative analysis revealed that both top-performing models mentioned clinical keywords such as ”stenosis”, ”narrowing”, and ”plaque” when identifying the carotid artery. Notably, both models sometimes detected textual labels embedded in the ultrasound images, particularly ”ICA”, ”LICA”, and ”RICA” (denoting Left/Right Internal Carotid Artery). However, LLaVa-NeXT-Vicuna-7B incorrectly expanded ”LICA” as Lateral Intra-Cranial Artery every time. In contrast, MedGemma frequently identified these labels correctly and appeared to base its predictions primarily on this textual cue rather than on intrinsic visual features, suggesting potential label-induced bias.

2) Risk Stratification: All models failed the zero-shot risk stratification task, exhibiting strong class biases. LLaVa-NeXT-Vicuna-7B consistently predicted *High risk*, LLaVa-NeXT-Mistral predicted *Low risk* across all cases, and MedGemma predicted *Low risk* in 99.44% of cases. This validates that USI risk stratification requires model adaptation beyond zero-shot prompting.

TABLE I: Prediction rates for the most mentioned imaging modalities and anatomies across four of the six LVLMs. Paligemma models were excluded as they failed to identify either modality or anatomy.

Models	Imaging Modalities				Anatomies			
	MRI	X-Ray	CT	USI	Brain	Lung	Neck	Carotid
LLaVa-Next-Vicuna-7B [3]	0.56%	0%	3.89%	**95%**	2.22%	0.28%	0.28%	**91.94%**
LLaVa-Next-Vicuna-13B [3]	0%	96.94%	3.06%	0%	-	-	-	-
LLaVa-Next-Mistral [3]	2.78%	2.22%	0%	**95%**	2.78%	2.22%	94.16%	0.83%
MedGemma [15]	0%	0%	0%	**100%**	0.28%	0%	0.56%	**98.33%**

B. Model Adaptation

Table II the stratified 3-fold cross-validation results of adapted LLaVa-NeXT-Vicuna-7B on held-out patients under *single-modality* and *multimodal* settings. Incorporating tabular metadata improved risk stratification, particularly at the patient-level where multimodal specificity ($63.3 \pm 32.1\%$) exceeded single-modality ($55 \pm 18\%$), demonstrating that the general-purpose foundation model successfully adapted to leverage multimodal ultrasound data without overfitting to training cases.

However, large standard deviations in specificity across folds indicate high variability in low-risk detection, likely due to severe class imbalance. Sensitivity remained consistently high ($> 90\%$) with low variance, revealing a bias toward predicting *High risk*. Compared to prior SOTA work [22], which trained domain-specific CNN ensembles from scratch on radiologist-segmented images using 4-fold cross-validation with additional patient cases, the adapted model achieved higher AUC but lower specificity. While direct comparison is limited by dataset and validation differences, the adapted general-purpose model demonstrates competitive performance and the ability to generalize to unseen patients, though detecting the under-represented low-risk class remains challenging due to dataset imbalance rather than model overfitting.

IV. Discussion

This work provides a systematic evaluation of state-of-the-art open-source LVLMs on ultrasound-based cardiovascular risk stratification, integrating USI with structured clinical, demographic, and laboratory data in textual form.

Zero-shot evaluation revealed that while LVLMs can identify imaging modality and anatomy with high accuracy, their performance on clinically meaningful classification tasks remains poor. MedGemma and LLaVa-NeXT-Vicuna-7B demonstrated strong ultrasound modality detection and carotid artery recognition. However, MedGemma's performance was partly attributed to embedded textual labels in images, raising concerns about label-induced bias rather than true visual reasoning. All models failed zero-shot risk stratification, assigning nearly all cases to a single class, reflecting the absence of domain-specific priors and the difficulty of relating ultrasound features to clinical risk without supervision.

Adapting the general-purpose LLaVa-NeXT-Vicuna-7B via LoRA fine-tuning significantly improved performance while maintaining generalization to unseen patients. Integrating tabular data consistently increased specificity and balanced accuracy, demonstrating effective multimodal fusion. However, specificity showed high variance across folds due to severe class imbalance, while sensitivity remained above 90% with low variance, indicating a bias toward high-risk prediction.

Compared to prior CNN ensembles trained on expert-segmented images, the adapted LVLM achieved comparable or higher AUC but lower specificity, matching overall discriminative power with less calibrated predictions. Critically, the LVLM approach eliminates the need for time-consuming expert annotations and manual segmentation, offering a more scalable solution for real-world clinical deployment where such annotations are rarely available.

This study demonstrates that general-purpose LVLMs can achieve competitive performance in specialized medical tasks with minimal domain-specific adaptation, while generalizing to unseen cases. The ability to integrate imaging and clinical data highlights the strength of multimodal learning. Nevertheless, the study is limited by the relatively small and unbalanced dataset, potential instability of treating videos as independent frames, and suboptimal multimodal fusion due to the lack of medical domain–specific architectural design. Additionally, LVLMs' text-based outputs hamper calibrated probability estimates, requiring pseudo-probability approximations for threshold-dependent metrics. In-image text labels may have also provided unintended cues. Future work should address these issues through larger and more balanced multimodal datasets, incorporation of temporal dynamics from ultrasound sequences, systematic assessment of text masking impact, and exploration of methods to extract calibrated confidence scores from LVLMs to improve clinical integration and interpretability.

V. Conclusion

This study evaluates open-source LVLMs for stroke risk stratification from ultrasound images and clinical data. While zero-shot evaluation showed strong modality and anatomy identification but failed classification, adapting the general-purpose LLaVa-NeXT-Vicuna-7B via parameter-efficient LoRA fine-tuning achieved competitive performance with improved generalization to unseen patients. Multimodal integration significantly enhanced results, demonstrating that general-purpose foundation models can adapt to specialized medical tasks without requiring expert annotations, offering a scalable path toward flexible diagnostic systems.

Acknowledgment

The authors acknowledge GRNET – National Infrastructures for Research and Technology for AWS cloud infrastructure access.

TABLE II: Results (in %) from 3-fold cross-validation after adapting the LLaVa-NeXT-Vicuna-7B model. Per-fold and aggregated (*mean ± standard deviation (SD)*) metrics are reported at frame and patient levels (via majority voting). Metrics: Sensitivity (Sens.), Specificity (Spec.), Balanced Accuracy (bAcc), Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC).

Tasks	**Folds #**	**Frame-Level Evaluation**				**Patient-Level Evaluation**				
		bAcc	**Sens.**	**Spec.**	**MCC**	**bAcc**	**Sens.**	**Spec.**	**MCC**	**AUC**
Single-modality	1	68.88	96.93	40.83	47.9	67.5	95	40	43.08	83.5
	2	60.89	93.73	28.04	24.58	69.74	89.47	50	39.47	71.05
	3	84.12	91.91	76.33	56.28	82.5	90	75	59.65	93.13
	Mean (±SD)	71.3 ± 11.8	94.2 ± 2.6	48.4 ± 25	42.9 ± 16.4	73.25 ± 8.1	91.5 ± 3.1	55 ± 18	47.4 ± 1.1	82.6 ± 11
Multimodal	1	67.45	97.75	37.16	47.48	67.5	95	40	43.08	85
	2	62.22	98.71	25.72	39.34	72.37	94.74	50	50.35	70.39
	3	93.86	87.71	100	63.02	92.5	85	100	69.69	97.5
	Mean (±SD)	74.5 ± 17	94.7 ± 6.1	54.3 ± 40	50 ± 12	77.5 ± 13.3	91.6 ± 5	63.3 ± 32.1	54.4 ± 13.8	$\mathbf{84.3 \pm 13.5}$
CNN-Ensemble [22]	**Mean (±SD)**	–	–	–	–	72.5 ± 6	75 ± 17.6	70 ± 10.3	–	73 ± 10.7

REFERENCES

[1] X. Li et al., "Vision-Language Models in medical image analysis: From simple fusion to general large models," Information Fusion, p. 102995, 2025.

[2] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[3] H. Liu et al., "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge." Jan. 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[4] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan. 10, 2023, arXiv: arXiv:2201.11903. doi: 10.48550/arXiv.2201.11903.

[5] G. Xu et al., "LLaVA-CoT: Let Vision Language Models Reason Step-by-Step," July 21, 2025, arXiv: arXiv:2411.10440. doi: 10.48550/arXiv.2411.10440.

[6] M. A. Shaaban, A. Khan, and M. Yaqub, "MedPromptX: Grounded Multimodal Prompting for Chest X-Ray Diagnosis," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops, MICCAI 2024 Workshops, A. Schroder et al., Eds., Cham: Springer Nature Switzerland, 2025, pp. 211–222. doi: 10.1007/978-3-031-84525-3_18.

[7] J. Guo, X. Shan, G. Wang, D. Chen, R. Lu, and S. Tang, "LLAUS: A High-Quality Instruction-Tuned Large Vision Language Assistant for UltraSound," in Proceedings of the 2025 International Conference on Multimedia Retrieval, Chicago IL USA: ACM, June 2025, pp. 398–406. doi: 10.1145/3731715.3733374.

[8] A. Le et al., "U2-BENCH: Benchmarking Large Vision-Language Models on Ultrasound Understanding," arXiv preprint arXiv:2505.17779, 2025, unpublished.

[9] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, arXiv: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," Dec. 11, 2023, arXiv: arXiv:2304.08485. doi: 10.48550/arXiv.2304.08485.

[11] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," May 15, 2024, arXiv: arXiv:2310.03744. doi: 10.48550/arXiv.2310.03744.

[12] A. Steiner et al., "Paligemma 2: A family of versatile vlms for transfer," arXiv preprint arXiv:2412.03555, 2024.

[13] OpenAI et al., "GPT-4 Technical Report," Mar. 04, 2024, arXiv: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[14] G. Team et al., "Gemini: A Family of Highly Capable Multimodal Models," May 09, 2025, arXiv: arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805.

[15] A. Sellergren et al., "Medgemma technical report," arXiv preprint arXiv:2507.05201, 2025.

[16] C. Li et al., "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," June 01, 2023, arXiv: arXiv:2306.00890. doi: 10.48550/arXiv.2306.00890.

[17] M. S. Seyfioglu, W. O. Ikezogwo, F. Ghezloo, R. Krishna, and L. Shapiro, "Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos," Jan. 13, 2025, arXiv: arXiv:2312.04746. doi: 10.48550/arXiv.2312.04746.

[18] J. Ye et al., "Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai," Advances in Neural Information Processing Systems, vol. 37, pp. 94327–94427, 2024.

[19] D. Bos et al., "Atherosclerotic Carotid Plaque Composition and Incident Stroke and Coronary Events," JACC, vol. 77, no. 11, pp. 1426–1435, Mar. 2021, doi: 10.1016/j.jacc.2021.01.038.

[20] L. Melaku and A. Dabi, "The cellular biology of atherosclerosis with atherosclerotic lesion classification and biomarkers," Bulletin of the National Research Centre, vol. 45, no. 1, p. 225, Dec. 2021, doi: 10.1186/s42269-021-00685-w.

[21] K. Malekmohammad, E. E. Bezsonov, and M. Rafieian-Kopaei, "Role of Lipid Accumulation and Inflammation in Atherosclerosis: Focus on Molecular and Cellular Mechanisms," Front. Cardiovasc. Med., vol. 8, Sep. 2021, doi: 10.3389/fcvm.2021.707529.

[22] T. Ganitidis, M. Athanasiou, K. Dalakleidi, N. Melanitis, S. Golemati, and K. S. Nikita, "Stratification of carotid atheromatous plaque using interpretable deep learning methods on B-mode ultrasound images," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 3902–3905. doi: 10.1109/EMBC46164.2021.9630402.

[23] A. S. Antonopoulos, A. Angelopoulos, K. Tsioufis, C. Antoniades, and D. Tousoulis, "Cardiovascular risk stratification by coronary computed tomography angiography imaging: current state-of-the-art," Eur. J. Prev. Cardiol., vol. 29, no. 4, pp. 608–624, Mar. 2022, doi: 10.1093/eurjpc/zwab067.

[24] Z.-L. Li et al., "Research on ischemic stroke risk assessment based on CTA radiomics and machine learning," BMC Med Imaging, vol. 25, no. 1, p. 206, Jun. 2025, doi: 10.1186/s12880-025-01697-y.

[25] J. Shi et al., "Radiomics Signatures of Carotid Plaque on Computed Tomography Angiography," Clin Neuroradiol, vol. 33, no. 4, pp. 931–941, Dec. 2023, doi: 10.1007/s00062-023-01289-9.

[26] G. D. Liapi et al., "Carotid Plaque Motion Analysis in Ultrasound Videos to Discover Rupture-Prone Plaque Areas," in 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Feb. 2024, pp. 1–4. doi: 10.1109/ISBI56570.2024.10635865.

[27] Y. Wang, T. Wang, Y. Luo, and L. Jiao, "Identification Markers of Carotid Vulnerable Plaques: An Update," Biomolecules, vol. 12, no. 9, p. 1192, Sep. 2022, doi: 10.3390/biom12091192.

[28] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," July 14, 2020, arXiv: arXiv:1910.03771. doi: 10.48550/arXiv.1910.03771.

[29] L. Beyer et al., "Paligemma: A versatile 3b vlm for transfer," arXiv preprint arXiv:2407.07726, 2024.

[30] "MedGemma model card — Health AI Developer Foundations," Google for Developers. Accessed: Aug. 15, 2025. [Online]. Available: https://developers.google.com/health-ai-developer-foundations/medgemma/model-card

[31] "MedGemma: Our most capable open models for health AI development." Accessed: Aug. 15, 2025. [Online]. Available: https://research.google/blog/medgemma-our-most-capable-open-models-for-health-ai-development/

[32] E. J. Hu et al., "Lora: Low-rank adaptation of large language models.," ICLR, vol. 1, no. 2, p. 3, 2022.

[33] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 04, 2019, arXiv: arXiv:1711.05101. doi: 10.48550/arXiv.1711.05101.