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

1st Daphne Tsolissou
National Technical University of Athens
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
stergios.christodoulidis@centralesupelec.fr

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

6th Konstantina Nikita
National Technical University of Athens
Athens, Greece
orcid.org/0000-0002-4629-2163

***Abstract*—Reliable risk assessment for carotid atheromatous disease remains a major clinical challenge, as it requires integrating diverse clinical and imaging information in a manner that is transparent and interpretable to clinicians. This study investigates the potential of state-of-the-art and recent large vision–language models (LVLMs) for multimodal carotid plaque assessment by integrating ultrasound imaging (USI) with structured clinical, demographic, laboratory, and protein biomarker data. A framework that simulates realistic diagnostic scenarios through interview-style question sequences is proposed, comparing a range of open-source LVLMs, including both general-purpose and medically tuned models. Zero-shot experiments reveal that even if they are very powerful, not all LVLMs can accurately identify imaging modality and anatomy, while all of them perform poorly in accurate risk classification. To address this limitation, LLaVa-NeXT-Vicuna is adapted to the ultrasound domain using low-rank adaptation (LoRA), resulting in substantial improvements in stroke risk stratification. The integration of multimodal tabular data in the form of text further enhances specificity and balanced accuracy, yielding competitive performance compared to prior convolutional neural network (CNN) baselines trained on the same dataset. Our findings highlight both the promise and limitations of LVLMs in ultrasound-based cardiovascular risk prediction, underscoring the importance of multimodal integration, model calibration, and domain adaptation 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 development of Large Vision-Language Models (LVLMs), a class of multimodal systems designed to jointly process visual and textual information [1]. Such models combine encoders, typically Vision Transformers (ViT) [2] specialized in image inputs, with Large Language Models (LLMs), to encode text inputs. This combination has shown remarkable capabilities in processing and integrating information across multiple modalities [1]. In natural images, models such as LLaVa-NeXT [3] have been successfully applied to a wide range of tasks, including image captioning, visual question answering (VQA), report generation, and multimodal retrieval. These reasoning abilities emerge from a process called chain-of-thought (COT) prompting, a process where a model generates intermediate natural language reasoning steps before producing a final answer [4]. Interview-style prompting, in which the model is guided through step-by-step analysis, has been shown to elicit deeper modality-specific reasoning and improve performance in zero-shot settings [5].

In medicine, the ability of LVLMs to leverage multimodal information makes them particularly promising for complex diagnostic tasks where clinical decision-making depends on the synthesis of heterogeneous data types, such as imaging features, laboratory results, demographic factors, and clinical information [6]. However, despite this potential, their application in this domain, particularly in the field of Ultrasound imaging (USI), remains largely unexplored and challenging [7], [8]. USI is a particularly challenging modality, especially for AI-based systems. It is highly operator-dependent, susceptible to artifacts such as speckle noise and low contrast, and its appearance can vary considerably with acquisition settings, patient anatomy, and device manufacturer [7], [8]. Furthermore, most LVLMs are pre-trained on natural images and everyday concepts [3], [9], [10], making direct transfer to USI analysis not easy. The lack of large and high-quality paired image–text datasets in ultrasound, coupled with the fact that most existing multimodal medical datasets contain little or no USI data alongside rich clinical context, further hinders progress. This creates a pressing need to assess how well current LVLMs can adapt to the ultrasound domain

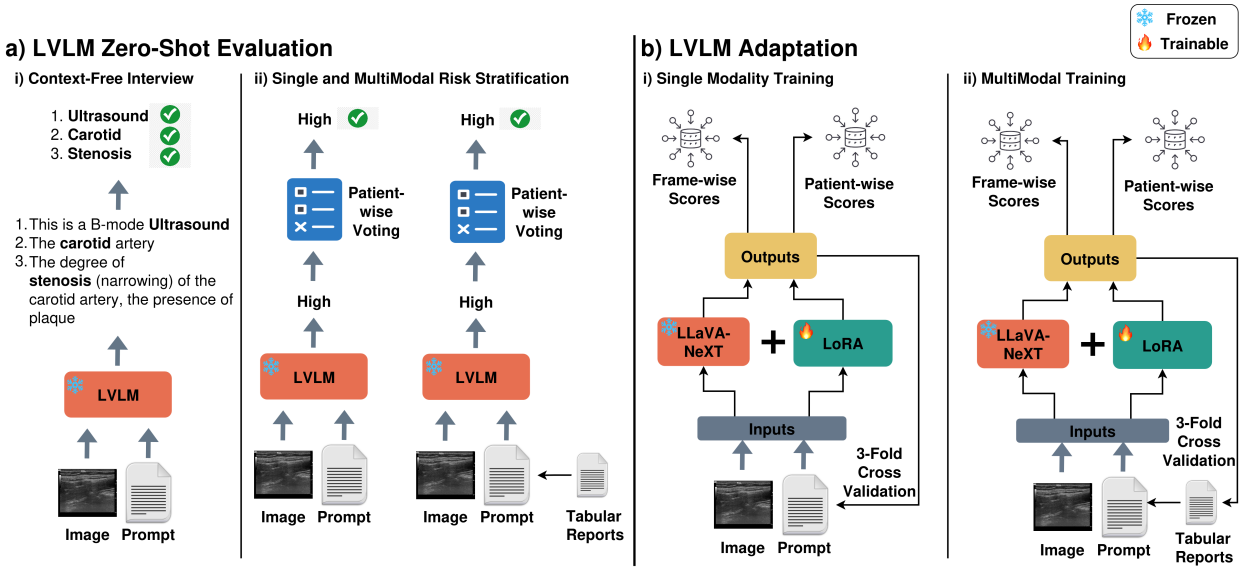


Fig. 1. Overview of the methods that were applied in this study. 1) A zero-shot evaluation framework was designed to assess state-of-the-art large vision language models in i) medical reasoning and ii) risk stratification using single and multimodal inputs from carotid ultrasound images and clinical, demographic, laboratory, and protein analysis tabular data. 2) The adaptation of LLaVA-NeXT is performed using Low-rank Adaptation (LoRA) and 3-fold cross validation in a single and multimodal setting.

and leverage associated textual data for clinically meaningful reasoning.

Several recent works have explored this direction. General-purpose open-source LVLMs such as PaliGemma [10] and LLaVa [9], [15], [16] have improved rapidly in recent years, in some cases approaching or even surpassing the performance of commercial models such as GPT-4V [17] and Gemini [18] on certain multimodal benchmarks. Their main advantage lies in the ability to be fine-tuned or adapted to specific datasets and domains, enabling improved performance in medical tasks and fostering research progress. MedGemma [19], developed by Google Deepmind, extends the Gemma 3 [20] language model with a medically tuned vision encoder, demonstrating strong multimodal medical reasoning capabilities across images and text, and serving as a solid foundation for downstream clinical applications. LLaVa-Med [21] is another example of a medical-domain LVLM, fine-tuned from LLaVA using biomedical figures, captions, and instruction-style data, and has been applied to build interactive medical assistants. In the histopathology domain, QUILT-LLaVA [22] is a LLaVa-based model fine-tuned on pathology image–text pairs derived from expert training videos, incorporating positional annotations of regions of interest to improve spatial grounding. Although these approaches have shown promise in radiology, pathology and ophthalmology, progress in USI has been slower, with only a few recent attempts such as LLAUS [7], an instruction-tuned LVLM for obstetric and general ultrasound.

Benchmarking efforts also play a crucial role in this field. In general medical multimodal reasoning, GMAI-MMBench [23] serves as a large-scale benchmark covering diverse medical imaging modalities and specialties, designed to test both perception and high-level reasoning in LVLMs. However, it contains a relatively small number of ultrasound cases. To address this, U2-Bench [8] was recently introduced as a comprehensive USI benchmark that offers a multi-task evaluation framework, spanning 15 anatomies such as breast, heart and lung, and 8 clinical tasks like classification (i.e. disease diagnosis) and text generation. This benchmark enables a more rigorous assessment of model capabilities in this modality. Even if U2-Bench is very diverse, it does not really evaluate the challenging task of carotid risk stratification.

Cardiovascular diseases (CVD) represent a significant clinical challenge, with carotid atheromatous plaques being one of the leading causes of ischemic stroke worldwide [11]. Atherosclerosis is a chronic and progressive inflammatory disease of the arterial wall, characterized by lipid accumulation, immune cell infiltration, and the formation of a fibrous cap over a lipid-rich core [24], [25]. Over time, these plaques can become unstable due to several risk factors, increasing the risk of rupture and subsequent cerebrovascular events such as stroke. Accurate risk stratification of carotid plaques is crucial for determining appropriate therapeutic interventions and preventing cerebrovascular events. Traditional approaches to plaque assessment have relied primarily on single-modality imaging, typically USI [12] or computed tomography angiography [13], combined with clinical risk factors [14]. However, the complex nature of atherosclerotic disease necessitates a more comprehensive approach that integrates diverse data sources to achieve optimal risk prediction, while ensuring that the underlying reasoning remains transparent and comprehensible to humans [26]–[28].

This study investigates the capabilities of state-of-the-art (SOTA) LVLMs on detecting vulnerable atheromatous plaques from multimodal data. Open-source models, such as LLaVA-

Next [3], are employed, focusing on the integration of USI with free text-based information, including clinical, demographic, laboratory, and protein analysis data. To simulate realistic diagnostic scenarios, a set of tailored prompts is designed that incorporates both image and textual patient-specific clinical context. The presented evaluation framework begins with an interviewing prompt sequence, investigating each model's visual understanding of the USI modality, the organ displayed, and the reasoning behind patient-specific risk factors. Subsequently, the models are further assessed in stroke risk stratification, following a zero-shot, as well as a task-specific fine-tuning approach.

An overview of the study can be seen in Fig. 1. In particular, the main contributions of this study are twofold: 1) the first in-depth comparative evaluation of general-purpose and medically tuned LVLMs for medical reasoning and inference for carotid plaque risk stratification, and 2) the task-specific fine-tuning of LLaVa-NeXT-Vicuna on carotid USI data, addressing the current scarcity of domain-adapted LVLMs for ultrasound-based CVD risk stratification. Finally, the study provides empirical insights into the strengths and limitations of existing LVLMs in medical inference, highlighting key challenges for generalization to real-world clinical settings.

II. Methods

A. Dataset and image strategy

In this study, an anonymized private dataset was used, consisting of B-mode carotid USI sequences (videos) with corresponding tabular data from 72 patients, aged 56 to 80 years, all presenting with $> 30\%$ atherosclerotic plaque. All videos were acquired under standardized conditions (including patient preparation, examination position, room temperature, and transducer settings) with a minimum duration of 3 *seconds*. A frame rate exceeding 25 *frames per second* was used, capturing approximately 2 to 3 cardiac cycles. The image resolution was 12 *pixels per millimetre* in both the radial and longitudinal directions. Each video was labeled by the degree of stenosis and the presence of a stroke or transient ischaemic attack within the six months preceding the ultrasound examination. Each sample was classified as high risk if it met one of the following criteria:

1) carotid artery stenosis exceeding 70%, or
2) occurrence of stroke or transient ischaemic attack symptoms within the preceding six months.

The dataset includes 59 high-risk and 13 low-risk patients, making it suitable for binary classification but inherently imbalanced. The tabular data included clinical and demographic information such as age, sex, and prescribed medications with dosages, blood test results, and protein analysis. None of the patients exhibited additional health conditions, such as cancer, chronic disease, or heart failure.

This video-based dataset was pre-processed to generate image-text pairs suitable for LVLM input. To maximize the use of each patient's video and to extract as much as possible visual information for zero-shot evaluation and training, multiple frames were extracted using two distinct sampling strategies:

1) **Random Sampling.** Five frames were randomly selected from each patient's video, after first dividing the video into five equal segments using a deterministic process. This ensured coverage across different phases of the cardiac cycle—during which the plaque might be more visible—and increased dataset variability at the frame level. These frames were then used for the zero-shot evaluation of the selected LVLMs.
2) **Full-frame Sequence.** In the second sampling strategy, all frames from each video were used for model adaptation. Training and testing were performed on disjoint patient subsets to prevent data leakage. Since videos had varying numbers of frames, some patients were overrepresented; this was mitigated with patient-level 3-fold cross-validation, and results are reported as the average across folds.

B. Prompt strategies for LVLMs

Three text prompts were designed to explore and compare the LVLMs' capabilities in both zero-shot and fine-tuning settings:

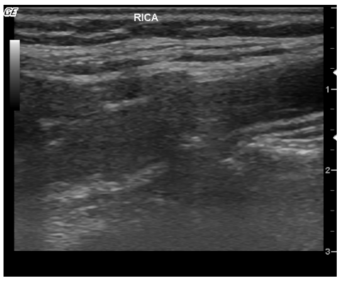
1) **Context-Free Interview**: this prompt provided only the medical image and instructed the model to act as an expert radiologist (Fig. 2). The prompt contained COT-style questions about the modality that produced the input image, the depicted organ, and diagnostically relevant visual features.
2) **Imaging Context**: This prompt included context about the imaging modality and organ (Fig: 3). It also defined high and low risk classes using the criteria we described before and requested a classification into one of these classes.
3) **Image and Tabular Context**: This prompt extended the image-only prompt, with patient specific demographic, clinical, and laboratory information converted to free text (Fig. 4). Binary values replaced with descriptive terms, e.g. 0 and 1 turned to male and female, and blood test acronyms expanded to their full names.

All image–text pairs were tokenized and normalized according to each VLM's input requirements using the relevant Hugging Face [29] pre-processing modules.

C. Benchmarking of the LVLMs models

Our study first focuses on the investigation of the capabilities of LVLMs for ultrasound image understanding. Each model was prompted with the frames selected using the random sampling strategy for each patient, accompanied by the interview text. Three tasks were evaluated: 1) modality identification, 2) anatomy identification, and 3) risk stratification.

Models that achieved the highest accuracy in modality and anatomy identification without additional context were further tested on the risk stratification task using two prompt settings: single-modality setting using only image data and multimodal



USER: You are shown a single image from a patient's imaging results.

Please answer the following questions clearly and concisely:

1. What imaging modality produced this image?
2. Which anatomical region is shown?
3. What visual characteristics or regions in this image are most important for assessing the risk of stroke?

Finally, based on your expert evaluation, classify the patient's stroke risk:

- **High** – The image shows features that suggest the patient is at high risk of stroke in the near future and may require urgent medical intervention.
- **Low** – The image shows features that suggest the patient has low risk of stroke in the near future.

Respond with **only the risk level**: `"High"` or `"Low"`.

Note: This prompt is for research purposes only and not for clinical use.

MedGemma:
1. **Ultrasound**
2. **Carotid** artery
3. The presence and degree of **stenosis** (narrowing) of the **internal carotid artery (ICA)** is crucial. The **RICA label** indicates the region of interest.

Risk Level: **Low**

LLaVa-NeXT-Vicuna (7B):
1. The imaging modality used in this image is a **B-mode ultrasound**, specifically a B-mode image of the carotid artery.
2. The anatomical region shown in the image is the **carotid artery**.
3. The most important visual characteristics for assessing the risk of stroke in this image are the degree of stiffness and the presence of any significant **narrowing or blockages** in the carotid artery.

Based on the image provided, the patient's stroke risk is ****High****.

LLaVa-NeXT-Mistral:
1. The image is an **ultrasound (US)** of the neck.
2. The anatomical region shown is the **neck**.
3. The most important visual characteristics or regions in this image for assessing the risk of stroke are the carotid artery and the **presence of plaques or stenosis**.

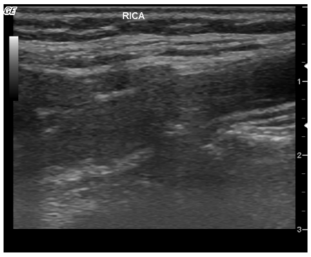
Based on my expert evaluation, the patient's stroke risk is `"**Low**"`.

LLaVa-NeXT-Vicuna (13B):
The image provided appears to be a **mammogram**, which is an imaging modality used for **breast cancer screening**.

1. The anatomical region shown is the **breast**.
2. The most important visual characteristics for assessing stroke risk in this image would be the presence of any **calcifications, masses, or asymmetries** that could indicate a potential risk for breast cancer.

Based on the expert evaluation, the patient's stroke risk is: "**Low**"

Fig. 2. An example of the context-free interview process. The model is prompted with the image and the interview-style text. Each model produces different responses, indicating its zero-shot capabilities in medical reasoning.



USER: You are an expert radiologist specialized in medical image analysis and stroke risk assessment from carotid ultrasounds.

You are given:
- A single frame from a patient's carotid ultrasound video.

Based on your expert assessment of the ultrasound image classify the patient's stroke risk as one of the following:

1. High – The image shows features that suggest the patient has a high likelihood of experiencing a stroke soon and should undergo surgery immediately.
2. Low – The image shows features that suggest the patient has a low likelihood of experiencing a stroke in the near future and does not require surgery right now.

Respond with only the risk level: **"High"** or **"Low"**.

Note: This prompt is for research purposes only and not for clinical use.

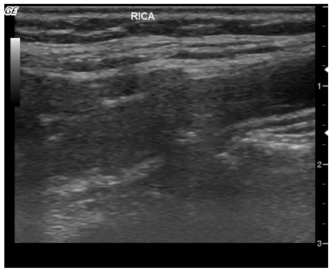
MedGemma: Low

LLaVa-NeXT-Vicuna (7B): High

LLaVa-NeXT-Mistral: Low

LLaVa-NeXT-Vicuna (13B): Low

Fig. 3. An example of the imaging context (single modal) prompt. The model is prompted with the image and user text that provides context for the image and asks a question that should be answered with a single word. This way of prompting is designed to assess the model's diagnosis capabilities from a single image.



USER: You are an expert radiologist specialized in medical image analysis and stroke risk assessment from carotid ultrasounds.

You are given:
- A single frame from a carotid ultrasound video.
- The patient's clinical data, including blood test results, protein levels, demographics, existing conditions, and medications.

{exams}

Based on your expert assessment of the ultrasound image and the provided patient data, classify the patient's stroke risk as one of the following:

1. High – The image shows features that suggest the patient has a high likelihood of experiencing a stroke soon and should undergo surgery immediately.
2. Low – The image shows features that suggest the patient has a low likelihood of experiencing a stroke in the near future and does not require surgery right now.

Respond with only the risk level: **"High"** or **"Low"**.

Note: This prompt is for research purposes only and not for clinical use.

MedGemma: Low

LLaVa-NeXT-Vicuna (7B): High

LLaVa-NeXT-Mistral: Low

LLaVa-NeXT-Vicuna (13B): Low

Fig. 4. An example of the imaging and tabular (multimodal) context prompt. The model is prompted with the image and user text that provides context for the image, demographic, clinical, and laboratory information, inside the test and asks a question that should be answered with a single word. This way of prompting is designed to assess the model's diagnosis capabilities from a multimodal prompt.

setting using image combined with tabular data. This aimed to investigate whether providing contextual information related to demographics, clinical risk factors, and laboratory examinations could improve decision-making in risk stratification.

In this study, six open-source LVLMs were explored. Five were general-purpose models—trained primarily on natural image–text pairs and suitable for inference with free-text prompts—namely: *Paligemma-Mix (3B) [30], Paligemma2-Mix (3B) [10], LLaVa-NeXT Vicuna (7B and 13B), LLaVa-NeXT Mistral (7B) [3]*. These models were selected for their strong performance on vision–language benchmarks, approaching that of leading commercial models such as *GPT-4* [17]. In addition, one medically tuned vision–language foundation model, *MedGemma (4B)* [19] was included. This model was chosen for its improved performance on a range of medical imaging benchmarks [31], [32] and because, according to its documented training data composition [31], it had not been trained on any ultrasound datasets, allowing for an unbiased assessment in this domain.

D. LVLM Adaptation to USI data and Carotid Risk Stratification

Following the zero-shot evaluation, the best-performing general-purpose model, *LLaVa-NeXT Vicuna (7B)*, was selected and adapted for risk stratification using LoRA adapter [33]. This model was selected such that we will test the challenges and opportunities of general-purpose models. Moreover, LoRA adaptation is well-suited for large transformer-based models as it freezes pre-trained weights, injects small low-rank matrices into selected layers, and trains only these matrices. This approach is more memory- and time-efficient than full fine-tuning.

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 $\alpha = 128$. 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 [34] 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, with all other hyperparameters kept as in the official LLaVA fine-tuning script. Model selection followed a patient-level 3-fold cross-validation scheme.

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

E. Evaluation Protocol

For zero-shot evaluation, interview responses were parsed using regular expressions to extract relevant information, and the proportion of cases in which the models explicitly mentioned the terms "ultrasound" and "carotid" was computed. For binary classification, Sensitivity, Specificity, Balanced Accuracy as well as Mathews Correlation Coefficient (MCC), and Area Under the Curve (AUC), known for their suitability in imbalanced data sets, were used as evaluation metrics.

These metrics were computed both *frame-wise* and *patient-wise*. For patient-wise metrics, majority voting over frame-level predictions was applied. AUC was calculated only at the patient level. In order to obtain a probability prediction, the mean value across *frame-wise* label predictions was calculated and used for the AUC evaluation.

III. Results

A. Zero-Shot Evaluation

1) Imaging Modality Detection: Table I presents the frequency with which different medical imaging modalities were mentioned by the models during the zero-shot interview evaluation. The medically tuned model, MedGemma, correctly identified USI in all cases. Among the general-purpose models, the smaller LLaVa variants (LLaVa-NeXT-Vicuna-7B and LLaVa-NeXT-Mistral) also achieved high accuracy, correctly identifying USI in 95% of cases; in the remaining 5%, they predicted other modalities such as magnetic resonance imaging (MRI), X-ray, or computed tomography (CT). In contrast, the larger LLaVa-NeXT-Vicuna-13B model consistently failed to detect the correct modality.

The Paligemma family of models is not included in this table, as both models consistently returned the message: *"Sorry, as a base VLM I am not trained to answer this question."*—even when a disclaimer was included in the prompt specifying that responses were for research purposes and not medical advice. Further zero-shot testing revealed that these models were highly sensitive to prompt wording, with minor changes in phrasing or syntax dramatically altering their outputs. When asked solely to describe the image content, they correctly identified the modality in roughly 35% of the cases. This highlights that Paligemma models are not suitable for USI tasks in a zero-shot setting.

2) Anatomy Detection: Table I shows the frequency of some of the most common anatomies mentioned by the models that correctly predicted the USI modality. The two LLaVa variants demonstrated completely different patterns in their answers. LLaVa-NeXT-Vicuna-7B correctly mentioning the carotid artery 91.94% of cases, whereas LLaVa-NeXT-Mistral did so in only 0.83% of cases, instead referring to the neck in 94.16% of cases. Overall, LLaVa-NeXT-Vicuna-7B was the most accurate general-purpose model in this study.

TABLE I

PREDICTION RATE OF MOST MENTIONED IMAGING MODALITIES AND ANATOMIES FOR 4 OUT OF THE 6 LVLMS THAT WERE USED IN THIS STUDY. THE PALIGEMMA FAMILY OF MODELS HAD FAILED TO IDENTIFY BOTH THE MODALITY AND ANATOMY OF THE INPUT AND FOR THIS REASON WERE EXCLUDED FROM THE TABLE.

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 [19]	0%	0%	0%	**100%**	0.28%	0%	0.56%	**98.33%**

TABLE II

AVERAGE EVALUATION SCORES FOR 3-FOLD CROSS VALIDATION AFTER THE ADAPTATION OF THE LLAVA-NEXT-VICUNA-7B MODEL. THE EVALUATION WAS PERFORMED BOTH IN FRAME LEVEL AND PATIENT LEVEL AFTER AVERAGING THE PROBABILITIES RECEIVED FOR THE FRAME LEVEL. THE EVALUATION WAS PERFORMED IN TERMS OF SENSITIVITY (SENS.), SPECIFICITY (SPEC.), BALANCED ACCURACY (BACC) AS WELL AS MATHEWS CORRELATION COEFFICIENT (MCC), AND AREA UNDER THE CURVE (AUC).

Tasks	**Frame-Level Evaluation**				**Patient-Level Evaluation**				
	bAcc	**Sens.**	**Spec.**	**MCC**	**bAcc**	**Sens.**	**Spec.**	**MCC**	**AUC**
Single-modality (Image)	$71.3 \pm 11.8\%$	$94.2 \pm 2.6\%$	$48.4 \pm 25\%$	$42.9 \pm 16.4\%$	$73.2 \pm 8.1\%$	$91.5 \pm 3.1\%$	$55 \pm 18\%$	$47.4 \pm 1.1\%$	$82.6 \pm 11\%$
Multimodal (Image & Tabular)	$74.5 \pm 17\%$	$94.7 \pm 6.1\%$	$54.3 \pm 40\%$	$50 \pm 12\%$	$77.5 \pm 13.3\%$	$91.6 \pm 5\%$	$63.3 \pm 32.1\%$	$54.4 \pm 13.8\%$	$\mathbf{84.3 \pm 13.5}$
CNN-Ensemble [12]	–	–	–	–	$72.5 \pm 6\%$	$75 \pm 17.6\%$	$70 \pm 10.3\%$	–	$73 \pm 10.7\%$

MedGemma again outperformed all others, identifying the correct anatomy in 98.33% of cases.

A closer qualitative inspection of the best-performing models revealed the specific visual cues that informed their predictions. When the carotid artery was correctly identified, the models often referenced clinical keywords such as "stenosis", "narrowing", and "plaque". Notably, both models sometimes leveraged textual labels embedded in the ultrasound images themselves. In particular, the label "ICA" (and its variants "LICA" and "RICA"), originating from the raw videos and denoting the Left/Right Internal Carotid Artery, was occasionally detected and used in reasoning.

However, the models differed in how they interpreted this label. LLaVa-NeXT-Vicuna-7B detected "LICA" in only 7 out of 360 images, and incorrectly expanded it as Lateral Intra-Cranial Artery. In contrast, MedGemma frequently identified the ICA/LICA/RICA labels correctly and appeared to base almost all of its predictions primarily on this textual cue rather than on the intrinsic visual features of the carotid artery. This reliance suggests a potential label-induced bias, meaning the model's strong performance in most cases may have been aided by incidental on-image text rather than true anatomical recognition.

B. Zero-shot Classification

All models failed the zero-shot risk stratification task. LLaVa-NeXT-Vicuna-7B consistently predicted *High risk* for every case, while LLaVa-NeXT-Mistral predicted *Low risk* across the board. Even MedGemma exhibited a similar bias, predicting *Low risk* in 99.44% of cases. The Paligemma models were prompted with slightly different instructions for the same tasks to encourage them to respond with the class label; however, their classification performance was worse than that of a random classifier. This finding validates that USI data are not easy to be handled in a zero-shot fashion from LVLMs and their adaptation is needed for more specific tasks.

C. Model Adaptation

Table II presents the average evaluation metrics and standard deviation from the 3-fold cross-validation of LLaVa-NeXT-Vicuna-7B on the held-out patient cases, for both frame- and patient-level evaluation, under the single-modality and *multimodal* settings. The results indicate that incorporating tabular metadata into the multimodal framework improves risk stratification performance. In particular, at the patient-level (Table II, right), the specificity of the *multimodal* setting ($63.3 \pm 32.1\%$) exceeds that of the *single-modality* setting ($55 \pm 18\%$), suggesting that tabular data help the model better identify the under-represented class. Overall, these findings show that the adapted model can effectively leverage the image-text pairs provided during training.

An interesting result is that, across both frame-wise and patient-wise evaluations, specificity values demonstrate large standard deviations, indicating high-variability in low-risk detection across folds. This is likely the result of the strong class imbalance present in the dataset. In contrast, the sensitivity values remained consistently high ($> 90\%$) with low standard deviations. This reveals a strong bias toward predicting *High risk* class.

For context, Table II also includes results from prior SOTA work [12] on the same task, which trained a set of CNN models from scratch on the same dataset using radiologist-segmented cropped images, and used them in an ensemble setting for inference. That study employed 4-fold cross-validation and included additional patient cases (patients with multiple videos), which were excluded in the current experiments to reduce bias. While the differences in dataset composition and validation setup prevent direct comparison, these prior results

provide a reference for the performance of an expert, domain-specific model on the same task.

Comparing the reported scores, the adapted LLaVa-NeXT-Vicuna-7B achieves a higher AUC to the CNN ensemble but exhibits lower specificity, indicating greater difficulty in detecting the low-risk class. Combined with the relatively large standard deviations observed across folds and the high sensitivity scores, these findings suggest that the adapted model's generalizability remains limited.

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.

Our findings in the zero-shot setting, show that while current LVLMs can identify imaging modality and anatomical region with high accuracy in most cases, their performance on clinically meaningful classification tasks remains poor. MedGemma and LLaVa-NeXT-Vicuna-7B demonstrated strong performance in ultrasound modality detection and carotid artery recognition. Qualitative inspection revealed that when the models correctly recognized the carotid artery, they related it with common clinical keywords. However, for the MedGemma model, part of this performance can be attributed to textual cues such as embedded textual labels in the ultrasound images. This label-induced bias raises concerns about overestimating true visual reasoning ability when models rely on such cues, rather than visual features. Importantly, all evaluated LVLMs failed in the zero-shot classification of plaque vulnerability, predicting almost the entire test set to a single class. This might reflect the absence of domain-specific priors in pretraining and the difficulty to relate ultrasound features to clinical risk factors without explicit supervision.

Adapting LLaVa-NeXT-Vicuna-7B via LoRA fine-tuning significantly improved risk stratification performance. The integration of tabular data consistently increased specificity and balanced accuracy compared to the image single-modality setting, suggesting that LVLMs can benefit from multimodal tabular data when available. Nevertheless, specificity exhibited high variance across folds, suggesting unstable low-risk detection likely driven by severe class imbalance. In contrast, sensitivity remained above 90% in all cases, with low variance, indicating strong high-risk class detection. This also suggests a tendency towards the high-risk class, caused by the class imbalance.

When compared to prior CNN-based models trained on cropped images from the same dataset, the fine-tuned LVLM achieved comparable or higher AUC but lower specificity. This suggests that the fine-tuned LVLM matches overall discriminative power, but remains less calibrated for label prediction. However, the ensemble's reliance on expert-annotated data poses a significant requirement for manual effort, domain expertise, and time-consuming segmentation workflows, limiting its scalability and applicability in real-world clinical environments where such annotations are rarely available.

This study shows that, despite the challenges pointed above, a fine-tuned LVLM can achieve competitive discriminative performance, even matching or exceeding the AUC of CNN baselines trained on expert-annotated data. The model's ability to leverage both imaging and structured clinical data demonstrates the strength of multimodal integration, with significant performance improvement when patient metadata are included. These results highlight the adaptability of general purpose LVLMs to medical domains with minimal domain-specific training, offering a scalable path toward flexible diagnostic systems.

Potential limitations of the proposed work include the dataset size and class imbalance, both of which, while realistic, impact the stability and generalizability of results. Additionally, processing a video-based dataset as a collection of static images overlooks temporal information that is often crucial for diagnosis, as medical experts frequently rely on motion cues when performing risk stratification. Another limitation lies in the heterogeneity of the input modalities, integrating ultrasound frames with tabular patient data poses challenges for model alignment, and the current LVLM architectures may not optimally handle such multimodal fusion without architectural modifications. Moreover, the benchmark evaluation may be influenced by prompt design, and differences in how models handle medical terminology, potentially affecting comparability. Finally, fine-tuning on a small, domain-specific dataset risks overfitting, especially given the high model capacity of LVLMs.

Building on the results of this study, several key areas should be addressed towards translating LVLMs into reliable clinical tools. First, expanding to larger, and more balanced multimodal datasets will be crucial to improve model robustness and mitigate biases originated from limited training data. Improving calibration, particularly for underrepresented classes, will enhance the reliability of predictions in real-world decision-making. Incorporating temporal features from ultrasound sequences could further enrich the models' understanding of vascular motion through the cardiac cycle and improve diagnostic accuracy. Finally, future directions should leverage the inherent interactive capabilities of LVLMs, such as question–answering sequences, to provide step-by-step reasoning. Such features can enhance interpretability, clarify diagnostic decisions, and support the integration of these models into real-world diagnostic workflows.

V. Conclusion

This study presents an in-depth evaluation of open-source LVLMs for carotid plaque risk stratification from ultrasound images combined with structured clinical data. Zero-shot evaluation revealed that while some models excel at identifying imaging modality and anatomy, they fail at clinically relevant classification without adaptation. Fine-tuning LLaVa-NeXT-Vicuna-7B using LoRA significantly improved performance, especially when multimodal tabular data were included, achieving competitive AUC against CNN baseline architectures.

ACKNOWLEDGMENT

The authors acknowledge support from GRNET – National Infrastructures for Research and Technology, which provided access to AWS cloud infrastructure. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union under the NextGenerationEU Program, administered by the Archimedes Unit of the Athena Research Center.

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, A. Schroder, X. Li, T. Syeda-Mahmood, N. P. Oxtoby, A. Young, A. Hering, T. S. Mathai, P. Mukherjee, S. Kuckertz, T. He, I. Llorente-Saguer, A. Maier, S. Kashyap, H. Greenspan, and A. Madabhushi, 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] A. Steiner et al., "Paligemma 2: A family of versatile vlms for transfer," arXiv preprint arXiv:2412.03555, 2024.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] G. Team et al., "Gemma 3 Technical Report," Mar. 25, 2025, arXiv: arXiv:2503.19786. doi: 10.48550/arXiv.2503.19786.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] "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

[32] "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/

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

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