Unified Mixture-of-Experts Framework for Joint Cardiac and Vascular Ultrasound Analysis and Report Generation

Bin Pu[1], Jiewen Yang[2], Xingguo Lv[1], Kai Xu[3], Kenli Li[1]*

[1]Hunan University, Changsha, China
[2]The Hong Kong University of Science and Technology, HKSAR, China
[3]Yunnan University, Kunming, China
{pubin,lvxg,lkl}@hnu.edu.cn, jyangcu@connect.ust.hk

Abstract

Echocardiography and vascular ultrasound are essential for comprehensive cardiovascular assessment, yet manual evaluation and writing reports are labor-intensive, time-consuming, and require expertise from both cardiology and vascular surgery departments. Current automated report generation systems mainly focus on X-ray or CT, often neglecting echocardiographic modalities and critical quantitative parameters like aortic diameter and main pulmonary artery diameter, limiting their clinical utility. Moreover, the interdependence between cardiac and peripheral vascular health necessitates cross-departmental insights, which existing methods fail to incorporate. To address these limitations, we first propose the vision-language framework named the Echo-Cardiac-Vascular (**ECV**), for joint cardiac and vascular ultrasound report generation and parameter measurements. ECV introduces a Mixture-of-Experts vision encoder tailored for distinct ultrasound subtypes, a structured parameter measurement module for accurate quantification, and task-specific decoders that generate interpretable, multimodal diagnostic reports. Our framework, trained on 10K+ paired records, achieves high accuracy, improving diagnostic efficiency, consistency, and cross-disciplinary clinical applicability.

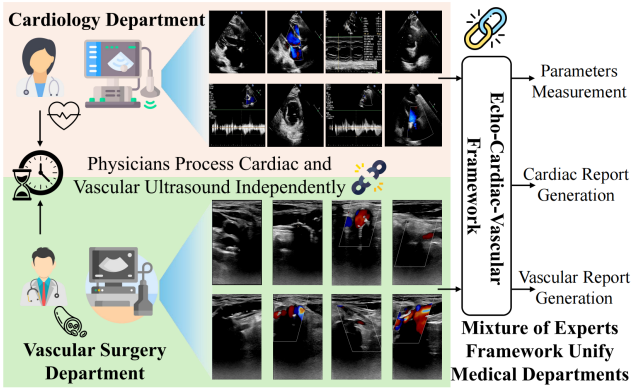

Figure 1: The motivation of our ECV framework. Previous cardiac and vascular measurements require manual assessment by specialists like echocardiographers and sonographers, whereas our ECV framework uses a vision-language model with the mixture-of-expert architecture to automatically deliver accurate multi-domain measurements, simulating cross-departmental collaboration.

Introduction

Echocardiogram and carotid artery screening play a pivotal role in diagnosing heart and vascular diseases, respectively, requiring comprehensive evaluation of numerous parameters such as aortic diameter (AOD), main pulmonary artery diameter (MPAD), and Right Atrial Diameter (RAD), leading to a final diagnosis (Deng et al. 2024; Yang et al. 2023a). Consequently, cardiac screening report and vascular report are inherently lengthy and detailed, encompassing multifaceted assessments of heart and blood vessel function and biological parameters, (Pillai et al. 2024). Compared to radiologists, cardiology sonographers and Vascular surgeons face a heavier workload, as each examination demands meticulous analysis of over a dozen parameters and the generation of extensive documentation. Compounded by a global shortage of trained personnel, these challenges underscore the urgent need for automated cardiac report generation and vascular report generation to enhance workflow efficiency and reduce clinician burden.

Recently, automated medical report generation (Wang et al. 2023) has emerged as a promising solution to mitigate these challenges and enhance the overall efficiency of the diagnostic process, attracting significant research interest. For instance, Li *et al.* (Li et al. 2023b) proposed a dynamic knowledge graph with a contrastive learning method for chest X-ray report generation, which adaptively updates graph structures and nodes while leveraging contrastive learning to improve visual-textual alignment. Despite the notable progress achieved by these latest studies, they remain limited in two key aspects. (1) *Existing studies primarily focus on X-ray (Li et al. 2023b; Huang, Zhang, and Zhang 2023; Yang et al. 2022, 2023b) and CT (Liu et al. 2021; Hamamci, Er, and Menze 2024) report generation, with very limited exploration of other imaging modalities such as echocardiogram and carotid artery.* (2) *Current automated report-generation methods fail to incorporate the evaluation and analysis of key biological parameters. However, these parameters (e.g., AOD, MPAD) are*

*Corresponding author.

critical for comprehensive cardiac and vascular screening. The omission of such essential metrics in prior approaches limits their clinical applicability, rendering them potentially unsuitable for echocardiographic report generation.

In clinical practice, one observation has shown that cardiac function, which is responsible for systemic blood supply, often exhibits concomitant abnormalities in peripheral vasculature when impaired (Mounsey et al. 2025; Joyce and Wang 2020; Wright et al. 2004). *Therefore, accurate diagnosis of cardiac pathologies not only requires cardiovascular physicians (from cardiology department) to possess an in-depth understanding of cardiac pathophysiology but may also necessitate comprehensive analysis incorporating knowledge of peripheral vascular systems (from vascular surgery department).* For instance, in the diagnosis and treatment of coronary artery disease, while cardiologists evaluate cardiac function, vascular surgeons frequently utilize carotid ultrasound to assess plaque burden and stenosis severity, enabling collaborative determination of disease progression and more precise therapeutic strategies (Naghavi et al. 2024). However, training physicians proficient in both cardiac and peripheral vascular ultrasound examinations typically requires over a decade, making it challenging to meet the growing clinical demands (Speranza et al. 2025; Lafitte et al. 2025). Consequently, there is an urgent need to develop novel automatic generating diagnostic report technologies capable of synchronously analyzing cardiac and peripheral vascular ultrasound images.

Building upon this analysis, we first collected a large-scale ***C*ardiac and *V*ascular *U*ltrasound (CVU)** multimodal dataset comprising 11,276 paired patient records from cardiology and vascular surgery departments with a total of 335,151 ultrasound images. Subsequently, we proposed **ECV** (see Figure 1), the first multimodal large language model specifically engineered for joint cardiac ultrasound and vascular examination report generation and parameter measurement that covered 25 key metrics (see Table A1 in Appendix). This framework pioneers the integration of cross-departmental data to enable comprehensive cardiac and vascular assessment. Specifically, our proposed ECV framework is a vision-language model (VLM)-based architecture tailored for the joint analysis of cardiac and vascular ultrasound images.

Our framework addresses key limitations in automated ultrasound report generation and parameter measurement through three integrated innovations. First, we introduce a Mixture-of-Experts-based vision encoder, where experts specialize in distinct anatomical regions or modalities (e.g., echocardiographic vs. vascular ultrasound). A lightweight gating mechanism dynamically selects and combines experts during inference, enabling adaptive, context-aware feature extraction with improved generalization and modularity. Second, unlike prior 'black-box' approaches (Moor et al. 2023; Zhu et al. 2023; Chen et al. 2024; Li et al. 2023a), our framework includes a structured parameter measurement module that accurately quantifies key cardiovascular metrics (e.g., AOD, MPAD). These interpretable measurements guide the task-specific visual-language model to generate clinically coherent diagnostic reports. Finally, by jointly modeling cardiac and vascular data, our framework supports a multimodal, cross-departmental approach to automated cardiovascular assessment—advancing the state of the art in precision and clinical interpretability. In summary, the primary contributions are outlined as follows:

- We introduce the first large-scale multimodal dataset of 11,276 paired cardiac and vascular ultrasound examinations with expert-annotated reports, enabling integrated modeling of cardiovascular and peripheral vascular health.
- We propose a MoE-based VLM framework named ECV that can dynamically adapt to diverse ultrasound modalities through specialized representations, enhancing accuracy and generalization in several tasks.
- Our ECV framework achieves precise joint cardiac and vascular parameter measurements and generates clinically reliable, interpretable reports by aligning textual findings with quantified data.

Related Work

Automated Parameter Measurement in Ultrasound

Automated measurement of clinical parameters in ultrasound imaging has made substantial progress, especially in echocardiography. Recently, segmentation-free regression frameworks such as EchoNet-Dynamic (Ouyang et al. 2020), EchoClip (Christensen et al. 2024), and EchoPrime (Vukadinovic et al. 2024) have attracted widespread attention, which can directly estimate functional indices from echocardiogram videos. Self-supervised pretraining and contrastive learning have further improved generalizability (Holste et al. 2024; Yang et al. 2023a; Deng et al. 2024). More advanced architectures employ uncertainty modeling or metric supervision, while parameter-efficient tuning techniques like LoRA (Hu et al. 2022) allow for the adaptation of large-scale visual encoders to ultrasound-specific domains. Meanwhile, sparse expert models such as Mixture-of-Experts (MoE) (Shazeer et al. 2017; Huai et al. 2025) offer modularity for handling heterogeneous input modalities. Despite these advances, most methods focus exclusively on cardiac measurements and operate in isolation from vascular assessments. In contrast, our work establishes a unified framework that simultaneously quantifies both cardiac and vascular parameter measurements, enabling a more physiologically coherent and clinically comprehensive evaluation.

VLMs for Medical Report Generation

Medical report generation has evolved from early CNN-RNN pipelines (Yin et al. 2019; Chen et al. 2020) to transformer-based architectures such as Clinical-bert (Yan and Pei 2022) and Metransformer (Wang et al. 2023) that leverage structured knowledge, clinical priors, and expert-designed tokens (Wang et al. 2023; Huang, Zhang, and Zhang 2023). While these models have shown success in modalities with static, well-localized findings in medical modalities such as chest X-rays (Wang et al. 2018). However, their application to ultrasound remains limited, where

ultrasound presents unique challenges due to its dynamic, multi-view acquisition, operator dependency, and the need for precise anatomical localization, compounded by the scarcity of large-scale, well-annotated report-image pairs. Recently, large vision-language models (VLMs) such as MiniGPT-4 (Zhu et al. 2023), LLaVA-Med (Li et al. 2023a), and PMC-VQA (Zhang et al. 2023) have demonstrated promising capabilities in medical visual question answering and image captioning. However, these models often generate descriptive text without grounding in quantitative measurements, resulting in reports that lack the specificity required for clinical decision-making. To bridge this gap, our ECV framework integrates structured parameter estimation with a vision-language decoder, combining modality-specific visual experts and cross-modal alignment to generate clinically coherent, quantitatively grounded reports for cardiac and vascular ultrasound, advancing from generic descriptions toward diagnostic-level precision.

Methodology

As illustrated in Figure 2, our proposed framework enables robust and automated analysis of ultrasound images, focusing on accurate parameter measurement and structured report generation. The methodology is composed of three core components: (1) Preprocessing and feature extraction, (2) Vision encoders with expert adaptors, and (3) Visual-Language Model for parameter measurement and report generation.

Preprocessing and Feature Extraction

This initial stage standardizes input images for downstream processing through cropping, padding, and proportional resizing, these critical steps given the high variability of ultrasound images across devices and protocols. Feeding raw images directly into the vision encoder risks information loss or distortion, especially when preserving anatomical proportions is essential for diagnosis. As shown in Figure 2, cropping isolates the region of interest (ROI), removing irrelevant structures to reduce noise and improve efficiency. Given an original image $I \in \mathbb{R}^{H\times W\times C}$, the ROI is a sub-image $I_{\text{ROI}} \in \mathbb{R}^{h\times w\times C}$. Padding preserves the aspect ratio during resizing by adding zero-padding if I_{ROI} does not match the target input size (H_s, W_s), preventing geometric distortions. Proportional resizing then maintains spatial relationships within the ROI, crucial for tasks like chamber segmentation or valve detection.

Given the padded image $I_{\text{pad}} \in \mathbb{R}^{h_{\text{pad}}\times w_{\text{pad}}\times C}$, the resized image $I_{\text{resize}} \in \mathbb{R}^{H_s\times W_s\times C}$ is obtained by $I_{\text{resize}} = \text{resize}(I_{\text{pad}}, \min(\frac{H_s}{h_{\text{pad}}}, \frac{W_s}{w_{\text{pad}}}))$. Finally, window partitioning allows local feature extraction with attention mechanisms, mimicking how clinicians examine specific regions of an image in detail. To facilitate efficient feature extraction by the vision encoder, the resized image is divided into non-overlapping windows. Each window $W_{ij} \in \mathbb{R}^{P\times P\times C}$ is extracted from I_{resize} using a sliding window approach: $W_{ij} = I_{\text{resize}}[iP:(i+1)P, jP:(j+1)P, :], \quad i = 0, 1, ..., \lfloor \frac{H_s}{P} \rfloor - 1, \quad j = 0, 1, ..., \lfloor \frac{W_s}{P} \rfloor - 1$, where P is the window size.

Vision Encoders with Expert Adaptors

This module uses Vision Transformers to extract semantic features from processed image windows. Given limited annotated ultrasound data, full fine-tuning is impractical due to computational cost and overfitting; instead, we use Low-Rank Adaptation (LoRA)-based (Hu et al. 2022) Expert Adaptors for efficient, effective adaptation. Their modular design enables transfer across modalities (e.g., echocardiography and vascular ultrasound) without full retraining. A Mixture-of-Experts architecture further enhances flexibility, with each expert specializing in a specific modality or anatomy, and a lightweight gating mechanism dynamically selects relevant experts during inference for context-aware feature extraction.

Given an image window $W_{ij} \in \mathbb{R}^{H_s\times W_s\times C}$, it is first flattened into patches and linearly embedded to form a sequence of tokens $X \in \mathbb{R}^{(N+1)\times d}$, where $N = \frac{H_s\times W_s}{P^2}$ denotes the number of patches and d is the embedding dimension. Positional embeddings are added to preserve spatial information before passing the input through transformer blocks. Instead of fully fine-tuning the entire vision encoder (Bai et al. 2025), which is both computationally expensive and prone to overfitting due to limited annotated ultrasound data, we employ LoRA modules. These insert low-rank matrices $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{r\times d}$ into the weight matrices W of the attention layers with $W' = W + UV$, where $r \ll d$ is the rank of the adaptation matrix. This significantly reduces the number of trainable parameters while preserving model expressiveness.

To enable specialized adaptation to different ultrasound imaging domains, we implement a MoE mechanism. Unlike conventional fine-tuning strategies that apply uniform parameter updates across all modalities, our approach assigns dedicated adaptors as expert modules, each tailored to a specific anatomical structure or imaging context. Let $\mathcal{E} = \{E_1, E_2, ..., E_K\}$ denote a set of K expert low-rank modules, each responsible for adapting the base vision encoder to a particular ultrasound domain. A lightweight gating network that is denoted as $G(\cdot)$ dynamically selects a sparse combination of these experts based on the input image characteristics:

$$\alpha_k = G(W_{ij})_k, \quad \sum\nolimits_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0, \qquad (1)$$

where α_k represents the weight assigned to expert E_k. The adapted weight matrix becomes:

$$W' = W + \sum\nolimits_{k=1}^{K} \alpha_k U_k V_k. \qquad (2)$$

This MoE-based design offers several key advantages. (i) Each expert adapts only to its designated domain, leading to better generalization and robustness compared to single-domain fine-tuning. (ii) Parameter Efficiency: Only a small subset of the total parameters is activated per inference step, reducing computational load and memory usage. (iii) Transferability: Pre-trained experts can be reused or recombined for new tasks, supporting rapid deployment in unseen ultrasound applications. (iv) Interpretability: The gating mechanism provides insights into which expert(s) are most relevant for a given input, aligning with clinical expectations

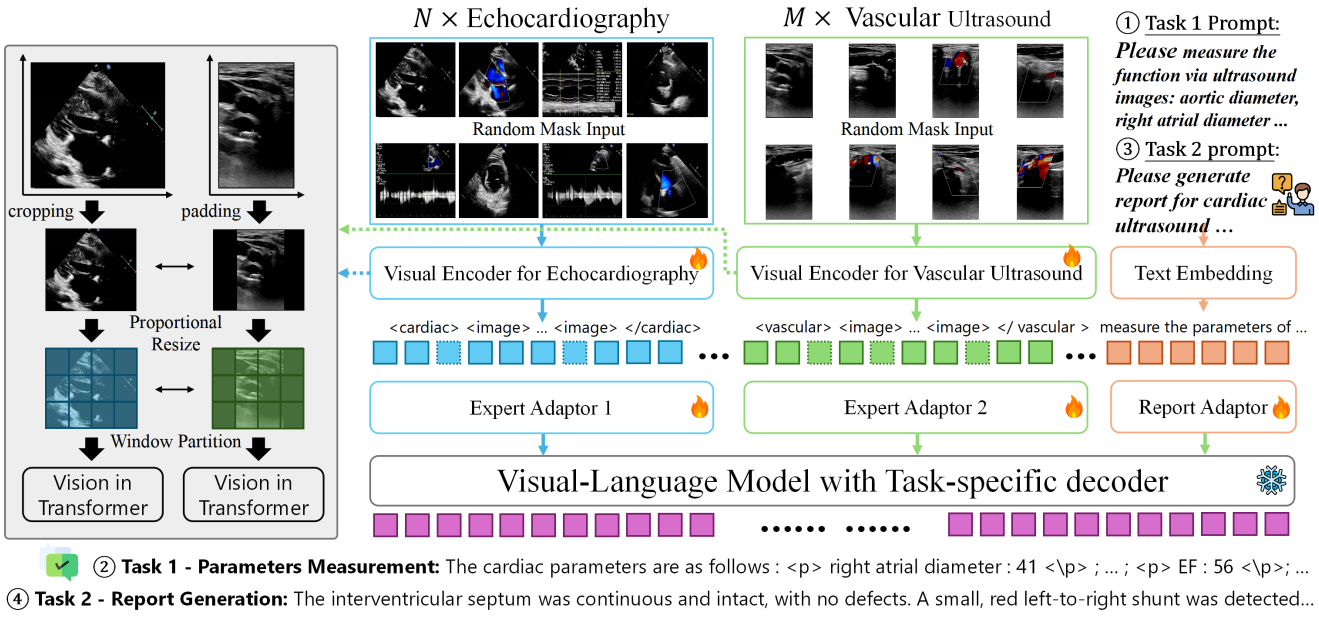

Figure 2: Overall pipeline of ECV framework. Our framework comprises vision encoders for different types of ultrasound and a language model decoder to process multimodal inputs. The vision encoder processes native-resolution inputs with dynamic length, preserves aspect ratios via augmentation, routes features through echocardiography or vascular ultrasound LoRA experts, and jointly decodes them with prompts to generate Electronic Health Record (EHR) content or parameter measurements.

of modality-specific expertise. By combining expert adaptors with MoE principles, our architecture achieves efficient, scalable, and interpretable adaptation of vision encoders to the heterogeneous nature of medical ultrasound imaging.

Parameter Measurement and Report Generation

The final output module integrates visual and textual modalities to generate structured diagnostic reports. Given that ultrasound interpretation combines visual assessment with standardized reporting, we employ a visual-language model to align visual embeddings from MoE vision encoders with text embeddings from diagnostic prompts. Cross-modal attention enables contextual understanding of visual findings through clinical language, facilitating the generation of coherent and clinically meaningful reports. This module bridges low-level visual features and high-level clinical interpretation. Following feature extraction and multimodal fusion via the MoE mechanism, the system performs precise parameter estimation—covering anatomical dimensions (e.g., ventricular diameters), functional metrics (e.g., ejection fraction), and qualitative descriptors (e.g., valve morphology)—which are seamlessly incorporated into the generated reports.

Let $Z_v \in \mathbb{R}^{N\times D}$ denote the sequence of visual embeddings extracted from the image windows, and $Z_t \in \mathbb{R}^{M\times D}$ represent the text embeddings derived from structured prompts or templates. The fused representation $Z_f \in \mathbb{R}^{D}$ is obtained via cross-modal attention (Bai et al. 2025). This fused representation is then used for two downstream tasks with task-specific decoders. Firstly, for continuous clinical measurements, such as ventricular size or ejection fraction, the system employs a lightweight multi-layer perceptron (MLP) to regress the quantitative values: The regression head is trained using a mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T} \|y_{\text{reg}}^{(t)} - y_{\text{gt}}^{(t)}\|^2, \quad y_{\text{reg}} = \text{MLP}_{\text{reg}}(Z_f), \tag{3}$$

where T is the number of numerical parameters, $y_{\text{reg}}^{(t)}$ is the predicted value, and $y_{\text{gt}}^{(t)}$ is the corresponding ground truth. Then, to generate structured diagnostic reports, the fused representation Z_f is fed into the decoder-based language model, which autoregressively generates a textual report $R = \{r_1, r_2, ..., r_T\}$ conditioned on the visual input:

$$R = \text{LLM}(Z_f). \tag{4}$$

The language model is trained using a standard cross-entropy loss over the generated tokens:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T}\sum_{t=1}^{T} \log p(r_t^{\text{gt}} \mid r_{<t}^{\text{gt}}, Z_f), \tag{5}$$

where r_t^{gt} denotes the ground-truth token at position t. The overall training objective combines both tasks into a unified multi-task loss as $\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{MSE}} + \lambda_2 \cdot \mathcal{L}_{\text{CE}}$, where λ_1 and λ_2 are hyperparameters balancing the contributions of the regression and generation objectives.

Our ECV framework supports automated ultrasound interpretation in real-world settings. To ensure clinical accuracy, report generation is guided by domain-specific templates and controlled decoding, including keyword constraints and syntactic structure enforcement. Our design combines quantitative measurements with structured narrative reports to ensure that the system supports both objective analysis and subjective interpretation, aligning closely with real-world clinical workflows.

Tasks	Metrics	Med-Flamingo (Moor et al. 2023)	MiniGPT-Med (Zhu et al. 2023)	HuaTuo-7B (Chen et al. 2024)	Med-Llava (Li et al. 2023a)	HuaTuo-34B (Chen et al. 2024)	Ours
PM	Precision ↑	-	-	75.68	75.68	77.42	**100.00**
	Recall ↑	-	-	43.55	21.16	23.68	**99.45**
	F1 Score ↑	-	-	53.05	31.05	36.26	**99.72**
	MAE ↓	-	-	16.05	15.55	13.25	**5.72**
	RMSE ↓	-	-	23.71	27.57	19.61	**12.43**
CRG	BLEU-1 ↑	11.99	15.29	28.62	18.25	31.99	**34.90**
	BLEU-4 ↑	0.80	1.21	4.81	2.63	**13.05**	5.82
	ROUGE-1 ↑	17.16	14.16	27.92	21.37	52.55	**54.70**
	ROUGE-L ↑	12.34	10.79	18.69	14.21	47.82	**51.40**
	BertScore ↑	59.92	59.08	65.09	59.24	82.78	**87.37**
VRG	BLEU-1 ↑	11.98	16.22	19.95	14.29	12.02	**24.58**
	BLEU-4 ↑	1.07	1.28	3.23	1.87	2.15	**6.67**
	ROUGE-1 ↑	13.84	14.44	21.73	17.12	7.86	**29.32**
	ROUGE-L ↑	10.18	10.37	16.23	11.83	7.79	**28.42**
	BertScore ↑	60.69	59.68	61.81	57.14	71.90	**81.98**

Table 1: The comparison between different methods in the report generation task. Results[1] reported in metrics Precision, Recall, F1 Score, MAE, and RMSE for Parameter Measurement (PM) task, while BLEU-1, BLEU-4, METEOR, ROUGE-L, and BertScore for Cardiac Report Generation (CRG) and Vascular Report Generation (VRG) tasks.

Results

Dataset

The proposed ***C*ardiac and *V*ascular *U*ltrasound (CVU)** dataset was collected from the Clinical Medical College of Dali University, the First Affiliated Hospital of Dali University, with a total of 11,276 patients with both Cardiac and Vascular ultrasound. Each part contains 10-30 ultrasound images, including echocardiogram images and Doppler spectrum images, which are collected by professional ultrasound physicians and saved according to specifications and standards of guidelines (Mitchell et al. 2019). The collection process was approved by the ethics committees of local hospitals. A senior and experienced sonographer from the hospital will perform manual parameter measurements based on ultrasound images of these patients. In total, over 25 key parameters that reflect cardiac and vascular functions are annotated, and abbreviations of these parameters are illustrated in Table A1 of our *Appendix*. Moreover, EHRs are written to evaluate the cardiac/carotid condition of patients. Images are captured by local sonographers from various ultrasound devices, such as Samsung and SonoScape. There are a total of 201,273 cardiac ultrasound images and 133,878 Vascular Surgery ultrasound images in CVU dataset. In total, our dataset was split with 10,276 cases for training and 1,000 cases for testing. This study has been approved by the hospital ethics committee (approval number: LLYJ2024-202-089). Each case and its metadata were anonymized, and all personally identifiable information was removed to prevent any leakage of identifying details.

Experimental Details

Training. We employ the Qwen-2.5-VL-3B model (Bai et al. 2025) as the basic model. Fine-tuning with a batch size of 128, employing the cosine learning rate schedule with a warmup ratio of 0.03. The learning rates for the main model, merger, and vision components were set to 1e-4, 1e-5, and 2e-6, respectively. We applied adaptors with a rank of 64 and a dropout rate of 0.05. In our fine-tuning process, we meticulously handled image data to enhance model performance. Images were resized and decoupled into patches to ensure a minimum and maximum of $256 \times 28 \times 28$ and $512 \times 28 \times 28$ pixels, respectively. We employed random sampling with each sample containing between 5 to 20 images.

Method	HuaTuo-7B (Chen et al. 2024)	Med-Llava (Li et al. 2023a)	HuaTuo-34B (Chen et al. 2024)	ECV (Ours)
AoAnn	20.40	21.41	13.91	**2.62**
AoD	22.20	22.18	25.44	**1.76**
AR Vmax	1.34	1.67	1.21	**0.76**
CI	1.07	1.08	0.69	**0.19**
CO	1.08	0.92	1.17	**0.28**
E/A	4.10	9.52	6.02	**0.76**
EF	9.48	9.44	6.98	**4.27**
IVST	6.70	6.95	8.30	**0.60**
LAD	34.05	42.89	47.12	**1.86**
LVPWT	6.62	6.66	5.81	**1.15**
LVEDd	30.33	22.27	16.92	**2.22**
LVESD	22.17	22.36	19.02	**3.94**
LVFS	24.23	26.39	29.22	**2.89**
MPAD	16.27	17.43	11.43	**1.47**
PASP	-	-	-	**12.03**
PR Vmax	-	-	-	**0.57**
PV	1.18	3.02	2.89	**0.68**
RAD	23.52	23.95	18.07	**1.93**
RVD	19.84	21.05	15.55	**3.33**
RVOT	16.86	12.74	13.86	**1.64**
Supra-Ao	4.91	7.71	4.14	**0.05**
Supra-Pulm	0.92	-	-	**0.25**
TRPG	25.00	7.33	**5.29**	10.02
TR $_{Vmax}$	1.53	1.66	2.01	**0.16**

Table 2: Parameter measurement result performed by different methods. Results[1] are reported in MAE. Med-Flamingo and MiniGPT-Med fail in measuring parameters.

		Modality Ablation			Component Fine-tuning Ablation			Training Strategy Ablation		
Task	Metric	Cardiac Only	Vascular Only	Ours (Fused)	LLM Decoder Only	Vision Encoder Only	Ours (Both)	Parameter-Only	Report-Only	Ours (Joint)
PM	Precision ↑	82.10	79.34	**100.00**	74.14	83.27	**100.00**	84.28	-	**100.00**
	Recall ↑	98.20	97.01	**99.45**	23.97	90.70	**99.45**	**99.98**	-	99.45
	F1 Score ↑	83.71	81.23	**99.72**	38.21	82.36	**99.72**	85.02	-	**99.72**
	MAE ↓	6.15	6.89	**5.72**	7.48	6.77	**5.72**	6.01	-	**5.72**
	RMSE ↓	42.30	45.12	**12.43**	61.56	55.03	**12.43**	40.46	-	**12.43**
CRG	BLEU-1 ↑	16.01	**35.85**	34.90	9.52	29.89	**34.90**	12.59	30.63	**34.90**
	BLEU-4 ↑	2.54	**6.19**	5.82	1.94	4.41	**5.82**	1.42	**5.84**	5.82
	ROUGE-1 ↑	25.90	**58.58**	54.70	14.30	46.70	**54.70**	16.79	49.72	**54.70**
	ROUGE-L ↑	24.66	**55.52**	51.40	13.79	44.97	**51.40**	14.60	48.07	**51.40**
	BertScore ↑	86.35	86.54	**87.37**	66.96	78.05	**87.37**	63.67	**88.43**	87.37
VRG	BLEU-1 ↑	4.24	19.41	**24.58**	4.83	16.82	**24.58**	7.92	8.46	**24.58**
	BLEU-4 ↑	0.82	3.84	**6.67**	0.37	2.90	**6.67**	1.83	0.56	**6.67**
	ROUGE-1 ↑	4.09	19.19	**29.32**	4.88	20.11	**29.32**	3.88	8.11	**29.32**
	ROUGE-L ↑	3.94	18.62	**28.42**	4.79	19.67	**28.42**	3.03	8.03	**28.42**
	BertScore ↑	78.64	79.88	**81.98**	68.12	74.69	**81.98**	61.80	71.93	**81.98**

Table 3: The ablation studies of ECV. Results[1] reported in metrics Precision, Recall, F1 Score, MAE, and RMSE for Parameter Measurement (PM) task, while BLEU-1, BLEU-4, METEOR, ROUGE-L, and BertScore for Cardiac Report Generation (CRG) and Vascular Report Generation (VRG) tasks. For each ablation experiment, all other modules are fully enabled (*e.g.,* in component fine-tuning ablation, both cardiac and vascular modalities are input, with parameter measurement and report generation tasks jointly trained).

Inference and Metrics. During the inference stage, all images were resized and decoupled to patches with a fixed size of $256 \times 28 \times 28$ pixels and included in the input sequence. A unified prompt template was applied across all comparison methods to ensure fair evaluation. For parameter measurement, we report metrics Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson correlation in the parameter measurement task. For report generation, we employed BLEU, ROUGE, and BERT-Score to evaluate the quality of generated reports *(see Appendix Section A1)*. Precision and Recall evaluate whether the model regresses exactly the set of predefined clinical parameters, ensuring no hallucinated metrics, instead of regression accuracy (assessed by MAE/RMSE). High precision/recall means all reported number corresponds to a valid, expected parameter.

Quantitative Analysis

Compared with Echocardiographers. As shown in Table 2 and Figure 3, our ECV framework closely matches manual measurements by experienced echocardiographers. It achieves notably low MAE for MPAD (1.47), IVST (0.60), and Aod (1.76), with similarly low RMSE, demonstrating high precision. Performance remains strong for LVEDd (2.22), LVFS (2.89), EF (4.27), and RAD (1.93), with clinically negligible errors. This accuracy is enabled by expert adaptors in our visual-language model, which address domain-specific challenges and support structured reporting. See Section A2 and Figures A1–A6 in the *Appendix* for case studies.

Compared with State-of-the-art Methods in Parameter Measurement. Tables 1 and 2 show that ECV significantly outperforms existing methods (Moor et al. 2023; Zhu et al. 2023; Chen et al. 2024; Li et al. 2023a). For AoAnn, ECV achieves MAE of 1.66 versus 20.40 (HuaTuo-7B) and 21.41 (Med-Llava); for MPAD, MAE is 1.47 versus 16.27 and 17.43. Similar gains are seen for LVEDd (2.22) and EF (4.27). ECV also excels in vascular metrics (Supra-Ao: 0.048; TR Vmax: 0.16). Many baselines return "NA" due to black-box designs that prioritize text over numerical outputs, limiting clinical utility. ECV's Mixture-of-Experts vision encoder and structured measurement module enable consistent, accurate quantification across cardiac and vascular parameters.

Compared with State-of-the-art Methods in Report Generation. In report generation, ECV surpasses all baselines (Table 1). For cardiac reports, it achieves BLEU-1: 34.90 and BLEU-4: 5.82—substantially higher than HuaTuo-7B (28.62, 3.10) and Med-Llava (18.25, 1.05). For vascular reports, BLEU-1 is 24.58 and BLEU-4 is 6.67. ROUGE-L scores (54.70 cardiac, 29.32 vascular) further confirm better content alignment. Competing models suffer from weak multimodal integration and lack explicit parameter modeling: Med-Flamingo and MiniGPT-Med produce unstructured reports, while HuaTuo-7B and Med-Llava miss clinical nuances due to single-expert designs. By unifying structured measurement with a visual-language model, ECV delivers precise, fluent, and clinically grounded reports across departments.

Ablation Studies

Ablation of Modality. Table 3 (modality ablation) shows the impact of ultrasound modalities in training the model. The joint training achieved the best performance (parameter MAE: 5.72; BertScore: 87.37 and 81.98 for cardiac and vascular reports). Training on cardiac ultrasound alone yielded slightly lower results (MAE: 6.15; BertScore: 86.35, 78.64), while vascular-only training performed significantly worse, particularly in cardiac parameter estimation (MAE: 6.89). These results show the complementary nature of cardiac and vascular data: cardiac ultrasound captures detailed cardiac function, while vascular ultrasound reflects periph-

The appendices and supplementary materials are available at https://github.com/Yore0/ECV.

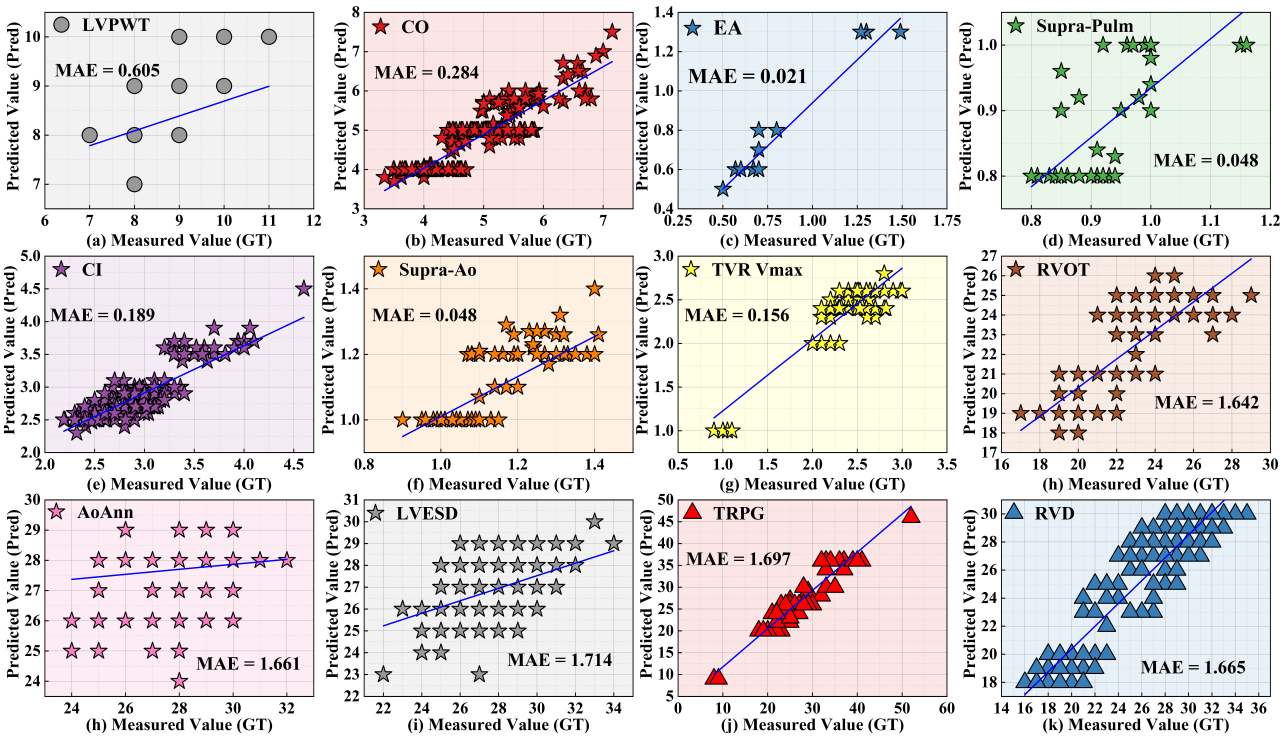

Figure 3: The cardiac parameters predicted by our ECV framework (y axis) and the parameters measured by echocardiographers (x axis). Plots with different colors and shapes in each subfigure denote different types of cardiac parameters.

eral hemodynamics. Integrating both enables cross-domain learning and improves diagnostic accuracy. While cardiac data alone supports robust performance due to its rich morphological content, relying solely on vascular data severely limits cardiac assessment, underscoring the necessity of multimodal integration for reliable clinical interpretation.

Ablation of Vision Encoder and LLM Decoder. We conducted an ablation study on the Vision Encoder and LLM decoder. As shown in component ablation of Table 3, joint fine-tuning achieved the best performance, with a parameter MAE of 5.72 and BertScore values of 87.37 and 81.98 for cardiac and vascular reports. Fine-tuning only the LLM decoder yielded the worst results (MAE: 7.48; BertScore: 66.96, 68.12), while optimizing only on one modal also dropped the performance (MAE: 6.89; BertScore: 86.54, 79.88). These results highlight the synergy between visual feature extraction and textual generation. Fine-tuning the vision encoder alone improves image representation but fails to produce coherent reports, whereas optimizing the LLM decoder enhances language generation but cannot overcome limitations from poor visual features. Joint fine-tuning enables end-to-end optimization, aligning visual representations with language modeling, thereby improving both measurement accuracy and report quality.

Multi-task and Single-task Finetuning. We evaluate multi-task versus single-task fine-tuning in the ECV framework in Table 3. Joint training achieves the best performance. Parameter-only training yields high accuracy (MAE: 6.01) but poor report quality (BertScore: 63.67, 61.80), resulting in incoherent narratives. Report-only training generates readable text but lacks parameter measurement capability and exhibits incomplete clinical coverage.

The advantage of multi-task learning lies in task synergy: accurate parameter estimation provides quantitative grounding for reports, while report generation enhances visual understanding, improving measurement. This mutual reinforcement highlights the effectiveness of joint fine-tuning for accurate, clinically meaningful ultrasound interpretation.

Conclusion

In this paper, we propose EVC, a novel vision-language framework for automated report generation from cardiac and vascular ultrasound. Unlike prior methods lacking explicit parameter estimation, EVC enables modality-aware, adaptive feature learning across diverse ultrasound domains, integrates accurate imaging quantification with structured reporting, ensuring clinical relevance and interpretability. We also introduce the first large-scale multimodal dataset with paired images and expert-labeled reports, supporting cross-departmental cardiovascular assessment. Our EVC has strong potential to improve diagnostic efficiency and reduce clinician workload. Limitations include reliance on high-quality annotations and limited generalization across devices. Future work will focus on domain adaptation, dataset expansion to more centers, incorporation of temporal dynamics in echocardiographic sequences, and interactive report generation with clinician feedback.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62272159, Guangdong Basic and Applied Basic Research Foundation, under Grants 2025A1515011404, and in part by the Natural Science Foundation of Hunan Province under Grants 2024JJ5089. We acknowledge the dataset sourced from the Clinical Medical College of Dali University, the First Affiliated Hospital of Dali University.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

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

Christensen, M.; Vukadinovic, M.; Yuan, N.; and Ouyang, D. 2024. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5): 1481–1488.

Deng, X.; Wu, H.; Zeng, R.; and Qin, J. 2024. Memsam: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9622–9631.

Hamamci, I. E.; Er, S.; and Menze, B. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 476–486. Springer.

Holste, G.; Oikonomou, E. K.; Mortazavi, B. J.; Wang, Z.; and Khera, R. 2024. Efficient deep learning-based automated diagnosis from echocardiography with contrastive self-supervised learning. *Communications Medicine*, 4(1): 133.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huai, T.; Zhou, J.; Wu, X.; Chen, Q.; Bai, Q.; Zhou, Z.; and He, L. 2025. CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19608–19617.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19809–19818.

Joyce, W.; and Wang, T. 2020. What determines systemic blood flow in vertebrates? *Journal of Experimental Biology*, 223(4): jeb215335.

Lafitte, S.; Rodrigues, L.; Ong, C.; Dezellus, A.; Goldberg, Y.; Bouchat, M.; Roger, E.; Moal, O.; Singh, V.; Moal, B.; et al. 2025. Artificial intelligence empowers Novice Users to acquire Diagnostic-Quality echocardiography. *Archives of Cardiovascular Diseases*, 118(6-7): S238.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

Liu, G.; Liao, Y.; Wang, F.; Zhang, B.; Zhang, L.; Liang, X.; Wan, X.; Li, S.; Li, Z.; Zhang, S.; et al. 2021. Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9): 3786–3797.

Mitchell, C.; Rahko, P. S.; Blauwet, L. A.; Canaday, B.; Finstuen, J. A.; Foster, M. C.; Horton, K.; Ogunyankin, K. O.; Palma, R. A.; and Velazquez, E. J. 2019. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American Society of Echocardiography. *Journal of the American Society of Echocardiography*, 32(1): 1–64.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Mounsey, L. A.; Guo, M.; Lau, E. S.; and Ho, J. E. 2025. Exercise Training in Heart Failure: Clinical Benefits and Mechanisms. *Circulation Research*, 137(2): 273–289.

Naghavi, M.; Reeves, A. P.; Atlas, K.; Zhang, C.; Atlas, T.; Henschke, C. I.; Yankelevitz, D. F.; Budoff, M. J.; Li, D.; Roy, S. K.; et al. 2024. Artificial intelligence applied to coronary artery calcium scans (AI-CAC) significantly improves cardiovascular events prediction. *npj Digital Medicine*, 7(1): 309.

Ouyang, D.; He, B.; Ghorbani, A.; Yuan, N.; Ebinger, J.; Langlotz, C. P.; Heidenreich, P. A.; Harrington, R. A.; Liang, D. H.; Ashley, E. A.; et al. 2020. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802): 252–256.

Pillai, B.; Salerno, M.; Schnittger, I.; Cheng, S.; and Ouyang, D. 2024. Precision of echocardiographic measurements. *Journal of the American Society of Echocardiography*, 37(5): 562–563.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Speranza, G.; Mischkewitz, S.; Al-Noor, F.; and Kainz, B. 2025. Value of clinical review for AI-guided deep vein thrombosis diagnosis with ultrasound imaging by non-expert operators. *npj Digital Medicine*, 8(1): 135.

Vukadinovic, M.; Tang, X.; Yuan, N.; Cheng, P.; Li, D.; Cheng, S.; He, B.; and Ouyang, D. 2024. EchoPrime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation. *arXiv preprint arXiv:2410.09704*.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; and Summers, R. M. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9049–9058.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11558–11567.

Wright, S. P.; Prickett, T. C.; Doughty, R. N.; Frampton, C.; Gamble, G. D.; Yandle, T. G.; Sharpe, N.; and Richards, M. 2004. Amino-terminal pro–C-type natriuretic peptide in heart failure. *Hypertension*, 43(1): 94–100.

Yan, B.; and Pei, M. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2982–2990.

Yang, J.; Ding, X.; Zheng, Z.; Xu, X.; and Li, X. 2023a. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11878–11887.

Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S. K.; and Xiao, L. 2023b. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86: 102798.

Yang, S.; Wu, X.; Ge, S.; Zhou, S. K.; and Xiao, L. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80: 102510.

Yin, C.; Qian, B.; Wei, J.; Li, X.; Zhang, X.; Li, Y.; and Zheng, Q. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE international conference on data mining (ICDM)*, 728–737. IEEE.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.