

EchoMed Agent: A Foundation Model–Based Agentic Framework for Echocardiographic Left Ventricular Segmentation and Functional Assessment

Leshan Sun

Mayo Clinic Arizona

Juan Farina

Mayo Clinic Arizona

Songlin Zhao

Lehigh University

Shuoyang Wei

Mayo Clinic Arizona

Yuzhen Ding

Mayo Clinic Arizona

Reza Arsanjani

Mayo Clinic Arizona

Win-Kuang Shen

Mayo Clinic Arizona

Wei Liu

Liu.Wei@mayo.edu

Mayo Clinic Arizona

Article

Keywords: echocardiography, left ventricular segmentation, foundation model, video segmentation, MedSAM2, ejection fraction, agentic AI

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11209308/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Accurate left ventricular (LV) segmentation and ejection fraction (EF) estimation from echocardiographic videos are important for cardiovascular risk stratification, but automated methods remain sensitive to domain shift and limited site-specific annotations. We present EchoMed Agent, an imaging-driven framework that integrates MedSAM2, a video segmentation foundation model adapted using rank-4 low-rank adaptation in the mask-decoder attention layers, with a constrained GPT-5.2–based natural-language tool-routing interface for prompt-guided LV segmentation and automated functional assessment. A lightweight axis-regression module, AxisHead, applies mask-guided pooling to frozen vision-transformer features to predict the LV long axis for single-plane area-length volume and EF estimation. The framework was initially trained on EchoNet-Dynamic and subsequently evaluated using few-shot adaptation on an independent Mayo Clinic cohort of 98 studies. With 20 labeled site-specific studies, EchoMed Agent achieved a Dice coefficient of 0.940 and a median absolute EF error of 4.33 percentage points on the Mayo cohort. This Dice score exceeded the 0.932 obtained by an EchoNet-Dynamic–trained MedSAM2 model transferred to Mayo without site-specific adaptation. Overall, 83.3% of evaluated cases had an absolute EF error of no more than 10 percentage points. These findings suggest that parameter-efficient site-specific adaptation can reduce the number of required annotations while maintaining clinically relevant segmentation and EF-estimation accuracy on an independent institutional cohort.

Introduction

Echocardiography is the most widely used cardiac imaging modality in clinical practice, offering real-time, non-invasive assessment of cardiac structure and function^{1,2}. Among its derived parameters, left ventricular ejection fraction (LVEF) is a primary biomarker for diagnosing heart failure, stratifying cardiovascular risk, and guiding therapy^{3,4}. LVEF is conventionally estimated via the biplane method of disks, requiring LV boundary delineation at end-diastole (ED) and end-systole (ES) across apical views^{5,6}; when only a single view is available, the area-length method offers an alternative, computing volume from cavity area and long-axis length at each phase⁷. In practice, ED/ES identification and boundary delineation are performed manually, a process that is time-consuming and subject to substantial inter- and intra-observer variability — on the order of 10–13 percentage points even among experienced readers — that can affect clinical classification^{8–10}. This challenge is compounded by the dynamic nature of echocardiographic video, which requires consistent boundary tracking under acoustic shadowing, limited field of view, and anatomical variation, motivating automated methods that reduce observer burden and improve reproducibility at scale.

Automated LV segmentation has advanced from frame-level convolutional approaches to spatiotemporal architectures^{11–13} and, more recently, vision foundation models capable of prompt-guided segmentation without task-specific retraining¹⁴. However, ultrasound's speckle noise, low-contrast boundaries, and view-dependent artifacts fall outside the natural-image distribution such

models are pretrained on, making domain adaptation essential for clinical deployment — and integrating adapted segmentation into a coherent, clinician-facing workflow introduces systems-level challenges existing approaches have not fully addressed.

Even when accurate segmentation and quantification algorithms exist, physicians do not interact with analysis pipelines through code or structured queries — a translational barrier that is often underappreciated. Agentic AI systems, in which a large language model (LLM) orchestrates specialized tools through conversational interaction, offer a promising solution by translating a natural language query into a coordinated sequence of computational steps and interpretable outputs^{15, 16}. To our knowledge, its application to echocardiographic video analysis — where segmentation, temporal reasoning, and quantification must be tightly coupled — remains largely unexplored.

Early LV segmentation approaches relied on model-based methods such as active appearance and active shape models, which encoded prior cardiac geometry but required careful initialization and struggled under poor image quality^{17, 18}. Deep learning shifted the field toward data-driven approaches, with fully convolutional and U-Net-based architectures^{13, 19} and, subsequently, spatiotemporal convolutional networks exploiting the cyclic motion of the cardiac cycle^{12, 20} substantially improving segmentation accuracy. EchoNet-Dynamic established a widely adopted benchmark, with leading fully supervised methods reporting strong segmentation and EF estimation performance when trained on the complete dataset²¹ — but these models assume abundant target-institution labeled data, a condition that rarely holds in practice. Few-shot and semi-supervised approaches have shown annotation-efficient adaptation is achievable in general medical image segmentation^{22, 23}, yet their systematic application to echocardiographic video across institutional boundaries remains limited, leaving reliable generalization to new sites with minimal labeling an unmet need.

The Segment Anything Model (SAM)¹⁴ demonstrated that a vision transformer trained on large-scale natural image data could support prompt-guided segmentation across diverse visual domains with minimal task-specific supervision. MedSAM²⁴ extended this to medical imaging via fine-tuning on a broad collection of medical segmentation datasets. SAM2²⁵ introduced a streaming memory architecture enabling video-level segmentation via frame-to-frame mask propagation — directly relevant to echocardiography, where consistent LV boundary tracking across the cardiac cycle is essential — and MedSAM2²⁶ adapted this video foundation model to the medical domain, improving boundary delineation over zero-shot SAM2 in several clinical imaging modalities. However, ultrasound's speckle noise, reverberation artifacts, and low-contrast boundaries present domain adaptation challenges beyond other medical imaging modalities, and few-shot generalization to unseen clinical sites remains an open problem for this model family. A concurrent multi-institutional effort (Mayo Clinic Rochester and Stanford)²⁷ has independently applied SAM-family models to echocardiographic segmentation but without extending to EF estimation

Agentic AI systems use an LLM as a reasoning backbone that dynamically selects and invokes external tools in response to natural language instructions²⁸. In radiology and clinical NLP, LLM-based agents have been applied to automated report generation, structured data extraction, and multimodal question answering over imaging studies^{29,30}, as well as clinical decision support, medication management, and EHR summarization^{31,32}. Our group recently demonstrated that autonomous LLM agents can meet the accuracy and scalability demands of real-world clinical workflows, with RadOnc-GPT performing iterative evidence assessment and structured outcomes labeling across radiation oncology patient cohorts¹⁶. However, existing clinical agentic systems have predominantly addressed text-based and structured data tasks, and their extension to video-based medical imaging analysis with tightly coupled perception and quantification components remains largely unexplored.

To address these challenges, we present EchoMed Agent, an imaging-driven framework for LV segmentation and EF estimation in echocardiographic video that couples a domain-adapted video segmentation model, a mask-guided long-axis regression module, and an LLM-based clinician-facing interface. Figure 1 summarizes the end-to-end workflow. Echocardiographic video and physician-provided spatial prompts are routed through the agentic interface to MedSAM2 for LV segmentation; the resulting masks are then used for automated ED/ES identification and single-plane area-length EF estimation, with quantitative findings returned as a structured physician-readable report. The key contributions of this work are threefold: (i) a parameter-efficient domain adaptation strategy for video foundation-model segmentation in echocardiography that generalizes to a new clinical site with minimal labeled data; (ii) AxisHead, a mask-guided long-axis regression module enabling single-view area-length volumetric estimation; and (iii) a constrained LLM-orchestrated interface that exposes the segmentation-to-quantification pipeline through natural-language interaction without requiring programmatic interaction from the clinician.

Results

Segmentation accuracy across datasets

Table 1 summarizes LV segmentation accuracy, measured by Dice similarity coefficient, across all methods and three evaluation datasets: EchoNet-Dynamic, the Mayo Clinic cohort, and the CAMUS external validation cohort. Among zero-shot baselines, SAM2 achieves Dice scores of 0.836 and 0.848 on EchoNet-Dynamic and the Mayo cohort respectively. MedSAM2 in the zero-shot setting improves upon SAM2 across all three datasets, achieving Dice of 0.893 on EchoNet-Dynamic, 0.910 on Mayo, and 0.955 on CAMUS — the strongest zero-shot transfer performance observed, reflecting the benefit of medical-domain pretraining even prior to any dataset-specific adaptation.

Among fully supervised models, MedSAM2 trained on the complete training pool for each dataset establishes the upper-bound reference: Dice of 0.941 on EchoNet-Dynamic (~ 8,000 training studies) and 0.969 on CAMUS (720 training cases); the same EchoNet-Dynamic-trained model, evaluated by zero-shot transfer without any Mayo-specific fine-tuning, achieves Dice of 0.932 on the Mayo cohort. EchoMed

Agent, adapted using only 20 labeled site-specific cases, achieves Dice of 0.934 on EchoNet-Dynamic, 0.940 on the Mayo cohort – exceeding the zero-shot-transfer baseline despite far fewer labeled examples – and 0.960 on CAMUS, closing most of the gap to the fully supervised CAMUS reference (0.969) using 36-fold fewer labeled training cases. These results demonstrate that few-shot LoRA adaptation generalizes segmentation performance beyond the institutional cohort to an independent, publicly benchmarked echocardiography dataset.

Table 1
Segmentation accuracy (Dice similarity coefficient)

Method	EchoNet-Dynamic Dice ↑	Mayo Cohort Dice ↑	CAMUS Dice ↑
SAM2 (zero-shot)	0.836 ± 0.128	0.848 ± 0.124	0.890 ± 0.038
MedSAM2 (zero-shot)	0.893 ± 0.054	0.910 ± 0.034	0.955 ± 0.020
MedSAM2 (fine-tuned, full)*	0.941 ± 0.026	0.932 ± 0.028	0.969 ± 0.013
EchoMed Agent (ours, 20-shot)	0.934 ± 0.034	0.940 ± 0.023	0.960 ± 0.016

↑ higher is better. *Fully supervised reference: full EchoNet-Dynamic training (~ 8,000 studies) for EchoNet-Dynamic; zero-shot transfer (no Mayo fine-tuning) for Mayo; full CAMUS training pool (720 cases) for CAMUS.

Ejection fraction estimation accuracy

Table 2 summarizes EF estimation accuracy on EchoNet-Dynamic and the Mayo cohort; CAMUS is not evaluated for EF, as its biplane reference standard is not directly comparable to the single-plane area-length method used in this work (Methods). Among zero-shot baselines, SAM2 yields median absolute EF errors of 9.87% on EchoNet-Dynamic and 8.19% on the Mayo cohort, with fewer than 57% of cases falling within the ± 10 percentage-point clinical acceptability threshold. MedSAM2 in the zero-shot setting improves upon SAM2, achieving median EF errors of 9.66% and 7.06% on EchoNet-Dynamic and Mayo respectively, though still well outside clinically acceptable bounds.

The fully supervised MedSAM2 model achieves the strongest EchoNet-Dynamic EF performance (median AE 3.55%, 91.6% of cases within ± 10 percentage points). EchoMed Agent, trained with only 20 labeled cases from the Mayo cohort, achieves a median EF absolute error of 4.24% on EchoNet-Dynamic and outperforms the same model's zero-shot-transfer performance on the Mayo cohort – achieving a median EF absolute error of 4.33% and 83.3% of cases within the ± 10 percentage-point threshold, compared to 5.47% and 80.6% respectively without Mayo-specific adaptation. These results demonstrate that EchoMed Agent achieves clinically competitive EF estimation on an independent institutional cohort with substantially less site-specific annotation than full supervision requires.

Table 2
Ejection fraction estimation accuracy

Method	EchoNet-Dynamic Median AE (%) ↓	EchoNet-Dynamic Cases ≤ 10pp ↑	Mayo Cohort Median AE (%) ↓	Mayo Cohort Cases ≤ 10pp ↑
SAM2 (zero-shot)	9.87	50.8%	8.19	56.1%
MedSAM2 (zero-shot)	9.66	51.9%	7.06	66.3%
MedSAM2 (fine-tuned, full)*	3.55	91.6%	5.47	80.6%
EchoMed Agent (ours, 20-shot)	4.24	85.1%	4.33	83.3%

↓ lower is better; ↑ higher is better. pp = percentage points. Median AE = median absolute EF error. Cases ≤ 10pp = percentage of cases with $|EF_{\text{pred}} - EF_{\text{ref}}| < 10$ pp (ASE inter-vendor tolerance guidelines). CAMUS not evaluated for EF (incompatible biplane reference, Methods). *Zero-shot transfer to Mayo, no fine-tuning. For the 20-shot EchoMed Agent condition, seed-to-seed dispersion across five random seeds is reported in Fig. 2 as ± SD.

Few-shot scaling analysis

Figure 2 illustrates the few-shot scaling behavior of EchoMed Agent across six annotation levels (0, 1, 3, 6, 10, and 20 shots) on both EchoNet-Dynamic and the Mayo cohort, with the MedSAM2 fully fine-tuned reference line shown for comparison. At zero shots, EchoMed Agent achieves Dice scores of approximately 0.893 and 0.910 on EchoNet-Dynamic and Mayo respectively, with median EF errors near 9.7% and 7.1%. Performance improves consistently and rapidly with increasing annotation: by 3 shots, Dice exceeds 0.913 on EchoNet-Dynamic and 0.928 on Mayo, with median EF errors falling to approximately 7.3% and 5.0% respectively. By 20 shots, both datasets reach Dice scores above 0.930 and median EF errors below 4.5%, approaching the fully supervised reference line.

The Mayo cohort achieves higher Dice scores and lower median EF errors than EchoNet-Dynamic at matched shot levels, suggesting amenability to rapid adaptation despite representing an unseen institutional distribution. Error bars remain narrow across all shot levels, indicating stable performance across random seeds.

Qualitative segmentation comparison

Figure 3 presents representative segmentation outputs across four methods — SAM2 zero-shot baseline, MedSAM2 zero-shot baseline, MedSAM2 fully fine-tuned, and EchoMed Agent at 20 shots — stratified by EF stratum (normal EF > 55%, borderline EF 40–55%, and reduced EF < 40%), with ED and ES frame pairs

shown for each case. SAM2 in the zero-shot setting produces visibly irregular boundary estimates with notable under-segmentation artifacts, particularly at ES frames where the LV cavity is smallest. MedSAM2 in the zero-shot setting yields smoother contours but still exhibits boundary leakage and shape inconsistencies relative to ground truth, most prominently in the reduced EF stratum where wall motion abnormalities challenge boundary localization. MedSAM2 fully fine-tuned produces contours closely matching ground truth across all three EF strata, consistent with its strong quantitative performance under full supervision. EchoMed Agent at 20 shots produces segmentation quality visually comparable to the fully fine-tuned reference across normal and borderline EF cases, with minor boundary discrepancies observed in isolated reduced EF cases at ES frames — consistent with the modest quantitative gap in the $\pm 10\%$ acceptability rate observed in Table 2.

Agentic workflow demonstration

Figure 4 demonstrates the complete EchoMed Agent output for a representative echocardiographic study. The system propagates LV segmentation through the video, derives the beat-to-beat cavity-area curve to identify ED and ES phases, overlays the AxisHead-predicted long axis on the selected frames, and returns the resulting EDV, ESV, and EF measurements in a structured report. This example illustrates how the segmentation, temporal phase detection, geometric estimation, and functional quantification components are combined in the final clinician-facing output.

Table 3 presents representative query-response pairs across four interaction categories, demonstrating the functional routing behavior of the EchoMed Agent interface. For well-formed clinical queries, the GPT-5.2 routing layer correctly maps natural language intent to the appropriate registered function: queries requesting cardiac phase identification are routed to `extract_ed_es_phases`, returning the requested ED or ES frame indices and corresponding area values, while queries requesting ejection fraction are routed to `compute_ejection_fraction`, returning a structured report containing EDV, ESV, and computed EF. Correct routing is maintained across variation in query phrasing — for example, both "Can you identify the ED/ES phase for me?" and phase-specific variants are handled without ambiguity.

For queries whose intent does not directly map to a registered function, the system exhibits two distinct behaviors. Semantically adjacent queries — such as a request for "pumping efficiency" — trigger a clarification prompt that proposes the nearest supported function and re-routes upon confirmation, preventing silent misrouting. Queries referencing capabilities entirely outside the registered function set — such as LV wall strain analysis — are declined with a direct response indicating that the function is not currently supported. Together these behaviors confirm that the agentic interface operates reliably across a range of clinical query types, including graceful handling of ambiguous and unsupported requests, without requiring programmatic interaction from the clinician.

Table 3
Representative agentic routing outcomes across query categories

Query Category	Example User Query	Routing Decision	Agent Response Summary
ED/ES Phase Extraction	"Can you identify the ED/ES phase for me?" "Can you extract the end-diastolic phase alone for me?"	extract_ed_es_phases	Returns detected ED frame index, ES frame index (joint query), or the requested phase index alone, with corresponding LV mask area values and beat-to-beat area curve.
EF Estimation	"Can you give me the ejection fraction for this case?"	compute_ejection_fraction	Returns EDV, ESV, and computed EF (%) alongside detected ED/ES frame indices and long-axis lengths in a structured report.
Ambiguous Query	"Can you analyse the heart's pumping efficiency for me?"	Clarification → compute_ejection_fraction	"Pumping efficiency is not a registered function. Did you mean ejection fraction calculation?" Re-routes upon confirmation.
Unsupported Query	"Can you do the left ventricle wall strain analysis for me?"	Declined	"Left ventricle wall strain analysis is not currently supported."

Discussion

EchoMed Agent demonstrates that a parameter-efficient video foundation model, adapted through few-shot LoRA fine-tuning and orchestrated via a natural language agentic interface, can achieve clinically competitive LV segmentation and EF estimation performance on an independent institutional cohort with a fraction of the annotation cost required by fully supervised approaches. The central finding is that meaningful clinical generalization is achievable with as few as 20 labeled cases — a regime practically accessible to institutions without large annotated echocardiographic archives — while preserving consistent performance across normal, borderline, and reduced EF strata, which is essential for deployment in real-world case mixes.

MedSAM2 currently requires geometric prompting — bounding box or point annotations — rather than free-text prompts. Geometric prompts provide explicit spatial constraints that yield tighter, more anatomically faithful boundary delineation than text-only prompting under challenging acoustic conditions²⁴, which motivates our choice of interaction modality here. As promptable foundation models increasingly unify spatial and language interaction (e.g., SAM3³⁶), integrating both modalities is a natural direction for future work.

This work is methodologically and empirically distinct from a concurrent Mayo Clinic–Stanford study that also applied SAM-family models to echocardiographic segmentation²⁷. That study fine-tuned frame-based SAM and MedSAM on an independently collected Mayo Clinic cohort of 99 patients (58 TTE from 2017–2018, 41 POCUS from 2022), evaluated segmentation accuracy alone, and did not address EF estimation or agentic clinician interaction. EchoMed Agent instead adapts the video-native MedSAM2 architecture via few-shot LoRA fine-tuning, couples segmentation to automated EF quantification through the AxisHead regressor, and exposes the full pipeline through a natural language agentic interface. Notably, our Mayo Clinic cohort (IRB 25-013540; 98 cases) was collected and annotated independently of the cohort used in²⁷; although both are drawn from the same institution, the two studies reflect non-overlapping patient populations, and their site-specific results are not directly comparable.

The modular agentic architecture places no structural constraint on the number of registered functions; additional cardiac structures and analyses — such as left atrial strain or right ventricular EF — can be integrated by training lightweight task-specific modules under the same few-shot protocol and registering them as callable tools.

The heavy-tailed EF error distribution was driven by a small subset of failure cases rather than by uniform degradation across the cohort. Qualitative inspection indicated that discrepancies were most apparent at end-systole, when the LV cavity is smallest and boundary localization is more sensitive to segmentation leakage or irregular contours; similar boundary discrepancies were observed in reduced-EF cases with abnormal wall motion (Fig. 3). We did not perform a prespecified disease-specific or image-quality subgroup analysis and therefore cannot determine from the present cohort whether these outliers are statistically associated with particular diagnoses, apical structural abnormalities, or poor acoustic windows. Future validation in larger, clinically characterized cohorts should stratify EF error by pathology, image quality, and acoustic-window adequacy to identify conditions in which the method is less reliable.

The current implementation has limitations. The agentic routing layer uses a structured hard-coded decision logic rather than full open-ended LLM reasoning, a deliberate choice that prioritizes routing robustness and functional reliability over conversational flexibility. Additionally, the Mayo Clinic cohort comprises 98 studies from a single institution; evaluation across additional sites and ultrasound vendors will be necessary to establish the generalizability of the few-shot adaptation strategy at scale.

We presented EchoMed Agent, an imaging-driven agentic framework that integrates parameter-efficient video foundation model segmentation with automated LV functional assessment through a clinician-facing natural language interface, and demonstrated segmentation generalization to an independent public benchmark (CAMUS) in addition to an institutional clinical cohort. The modular design supports straightforward extension to additional cardiac structures and functional analyses, offering a practical path toward comprehensive, annotation-efficient echocardiographic assessment in routine clinical environments.

Methods

MedSAM2: parameter-efficient fine-tuning for echocardiographic video segmentation

Base architecture

MedSAM2 is derived from the Segment Anything Model 2 (SAM2) ²⁶, a promptable visual foundation model for object segmentation in images and videos, comprising a hierarchical Hiera ViT image encoder, a lightweight memory module propagating object representations across frames, and a two-stage mask decoder. The mask decoder takes output tokens (with prompt tokens, $N_{\text{tokens}} \times 256$) and image embeddings ($256 \times 64 \times 64$), applies self-attention, token-to-image, and image-to-token attention layers, and produces the final mask via a $2\times$ convolutional upsampler and dot-product interaction with per-mask features (Fig. 5).

LoRA fine-tuning strategy

Full fine-tuning of the SAM2 backbone is computationally prohibitive and risks catastrophic forgetting. We therefore adopt Low-Rank Adaptation (LoRA) ³³, which injects trainable low-rank matrices into a frozen model's attention projections: for weight matrix $W_0 \in \mathbb{R}^{(d \times k)}$, LoRA introduces update $\Delta W = BA$ ($B \in \mathbb{R}^{(d \times r)}$, $A \in \mathbb{R}^{(r \times k)}$, $r \ll \min(d, k)$), giving adapted weight $W = W_0 + \Delta W$, with only A and B trained.

As shown in Fig. 5, we insert rank-4 LoRA adapters ($r = 4$) into all three mask-decoder attention layers (self-attention, token-to-image, image-to-token), applied to the query, key, value, and output projection matrices — the standard LoRA target in transformer fine-tuning. All other backbone weights, including the image encoder and memory module, remain frozen throughout training.

MedSAM2 was first fine-tuned on the EchoNet-Dynamic dataset, which comprises 10,030 apical four-chamber echocardiographic video studies with expert LV contour annotations at ED and ES frames. Training used bounding box prompts derived from ground-truth contours, with random jitter augmentation applied to the box coordinates for robustness. Point prompts (positive and negative) were sampled from within and outside the LV mask, respectively, to simulate the range of physician-issued annotations encountered in clinical practice.

To evaluate generalization to an independent clinical distribution, MedSAM2 is adapted to the Mayo cohort via few-shot fine-tuning: a small number of cases are randomly sampled from the 98 usable Mayo studies to continue LoRA fine-tuning from the EchoNet-Dynamic-trained checkpoint, with all other weights frozen. We evaluate six shot levels (0, 1, 3, 6, 10, 20), where 0-shot corresponds to direct inference with EchoNet-Dynamic-trained weights and no Mayo adaptation. Each shot level is repeated over five random seeds, and results are reported as mean $\pm$ standard deviation (SD); remaining Mayo cases are held out strictly for evaluation. This protocol applies identically to AxisHead.

AxisHead: LV long-axis estimation for volume calculation

Motivation and design

Ejection fraction estimation via the single-plane area-length method requires, beyond LV mask area A , the long-axis length L — the distance from LV apex to mitral valve center³⁴. Deriving L via ellipse-fitting or the mask's major axis is sensitive to boundary irregularities and inconsistent across frames. We therefore introduce AxisHead, a lightweight regression module that directly predicts the long-axis endpoint coordinates (X_1, Y_1, X_2, Y_2) from mask-conditioned image features.

Architecture

As illustrated in Fig. 5, AxisHead operates downstream of MedSAM2, taking the predicted binary mask and the image encoder's ViT feature map as input. The mask computes a soft spatial attention weight via mask-guided pooling, gating the ViT features to emphasize regions within and around the LV contour; the resulting pooled vector is passed to a three-layer MLP regressor ($\sim 33K$ parameters) outputting the four long-axis endpoint coordinates. This lightweight design is deliberate: AxisHead is an intermediate geometric estimator, not a standalone segmentation component, and its accuracy is evaluated implicitly through downstream EF error.

Training strategy

AxisHead is trained sequentially after MedSAM2 fine-tuning, with the MedSAM2 encoder and decoder frozen and only AxisHead parameters updated. In the first stage, AxisHead trains on EchoNet-Dynamic using pseudo-labels — the two contour points maximizing Euclidean distance along the base-to-apex direction — and is then calibrated on the Mayo cohort under the same few-shot protocol as MedSAM2 (Methods). This sequential design avoids co-training interactions that could destabilize segmentation performance.

LLM-based agentic interface

Overview and design rationale

The agentic coordination layer is implemented using GPT-5.2, a large language model (LLM) capable of natural language understanding and structured function invocation, accessed through the Mayo Clinic AI Factory — an institutionally governed Azure OpenAI deployment (via Apigee X) that keeps all clinical data within HIPAA-compliant infrastructure and never transmitted to external endpoints³⁵. This component serves as the system's conversational front-end: it receives free-form physician prompts, interprets intent, and dispatches the appropriate backend function call — for example, "segment the LV in this study and compute ejection fraction" — without requiring the clinician to know the underlying pipeline structure. Adding new analysis capabilities in future iterations requires only updating the routing logic and system prompt, not retraining a task-specific classifier.

Two-pass interaction protocol

The agent operates in two sequential interaction passes, each involving a separate API call to the GPT-5.2 endpoint (Fig. 1). In the first pass, the physician provides the echocardiographic video together with segmentation prompts specifying a bounding box around the LV and, optionally, one or more positive and negative point annotations. The GPT-5.2 model parses this combined input and invokes the `run_segmentation` function, passing the video identifier and structured prompt parameters to the MedSAM2 segmentation server. MedSAM2 processes the video and returns frame-by-frame binary LV masks, which the API assembles into an annotated video and returns to the physician.

In the second pass, the physician issues a post-segmentation natural language query — for example, "identify the ED and ES frames" or "calculate the ejection fraction." The GPT-5.2 model routes this query to one of two registered functions: `extract_ed_es_phases`, which identifies the cardiac phase frames from the beat-to-beat LV area curve, or `compute_ejection_fraction`, which additionally runs the area-length volume estimation and EF computation. The model reads the returned structured result and synthesizes a physician-readable natural language summary as the final output.

Function registration and routing

The two backend analysis functions are implemented as Python routines served through a Flask-based REST API running on a local Windows server. Clinician interaction is mediated through a custom web interface that handles both prompt submission and results display. Upon receiving a physician query, the GPT-5.2 model performs schema-constrained function invocation, returning a structured function-call object that identifies the inferred intent — either LV phase extraction or EF computation — based on semantic matching between the query and the predefined routing logic specified in the system prompt. This structured output is validated against a fixed function allowlist and dispatched to the corresponding Flask API endpoint over an authenticated channel to the local server; the model never generates or transmits free-form executable text. Results are returned through the same channel, fetched by the web interface, and rendered as a structured physician-readable output.

The system prompt provided to GPT-5.2 at each session establishes its role as a cardiac imaging analysis assistant, instructs it to restrict its reasoning to the registered function set, and explicitly prohibits it from generating clinical conclusions beyond those produced by the deterministic backend pipelines. This scoping is intentional: the model performs language understanding and routing, not autonomous clinical reasoning.

Automated ED/ES phase extraction and EF estimation

Beat-to-beat area curve construction

Following video segmentation by MedSAM2, the LV mask area is computed for every frame of the echocardiographic video by summing the binary mask pixels and converting to physical area units using the pixel spacing value provided by the operator at the time of analysis. This produces a temporally resolved beat-to-beat LV area curve $A(t)$, where t indexes the frame number, as illustrated in Fig. 4b. A

one-dimensional Gaussian smoothing kernel ($\sigma = 2$ frames) is applied to the area curve prior to extremum detection to suppress frame-to-frame boundary noise.

Cardiac phase detection

End-diastolic (ED) frames, corresponding to maximum LV filling, are identified as local maxima of the smoothed area curve $\tilde{A}(t)$; end-systolic (ES) frames, corresponding to maximum LV emptying, are identified as local minima. Formally, frame t^* is classified as an ED frame if for all t within a neighborhood of half-width w (i.e., $t \in [t^* - w, t^* + w]$), and analogously for ES frames at local minima. The neighborhood half-width w is set to 8 frames, corresponding to approximately one-third of a typical cardiac cycle at standard echocardiographic frame rates (30–60 fps at 60–80 bpm), which prevents spurious detections from minor area oscillations while remaining sensitive to physiologically plausible cycle lengths. Multiple ED–ES pairs may be detected across a multi-cycle video; the first detected pair is selected for downstream volume estimation, consistent with the EchoNet-Dynamic annotation convention of providing a single ED and ES frame per study.

Volume estimation and EF computation

LV volumes at the identified ED and ES frames are estimated using the single-plane area-length formula:

$$V = 8A^2 / (3\pi L)$$

where A is the LV mask area and L is the long-axis length predicted by AxisHead at the corresponding frame (Fig. 4c). This formula is the single-plane area-length method for LV volume estimation, consistent with American Society of Echocardiography (ASE) recommendations for two-dimensional echocardiographic quantification when only one acoustic window is available⁵. Ejection fraction (EF) is then computed as:

$$EF = (EDV - ESV) / EDV \times 100\%$$

where end-diastolic volume (EDV) and end-systolic volume (ESV) correspond to the volumes estimated at the detected ED and ES frames respectively. The agent compiles the detected ED/ES frame indices, LV mask areas, long-axis lengths, EDV, ESV, and computed EF into a structured physician-readable report (Fig. 4d).

Datasets

The primary training and validation dataset is EchoNet-Dynamic, a publicly available collection of 10,030 apical four-chamber echocardiographic video studies from Stanford University Medical Center, each with a de-identified MP4 video, expert LV contours at ED and ES, and a clinically measured EF value; we followed the dataset's official train/validation/test splits.

To evaluate few-shot generalization to an independent clinical distribution, we assembled a de-identified cohort of 98 echocardiographic studies from Mayo Clinic (IRB 25-013540), spanning normal, mildly reduced, and severely reduced EF to reflect routine clinical case mix. Each study was reviewed by a

board-certified echocardiographer, who provided LV contour annotations at ED and ES and verified the corresponding EF measurement.

To assess generalization beyond the training distribution, we additionally evaluate MedSAM2 on CAMUS¹³, a public benchmark of paired apical two- and four-chamber (2CH/4CH) acquisitions with expert LV contours at ED and ES. Because CAMUS provides frame-level contours only, without full-cycle video, this evaluation is restricted to segmentation accuracy at ED/ES; EF is not evaluated, as CAMUS's biplane reference is not comparable to the single-plane area-length estimate used elsewhere in this work.

Evaluation metrics

Segmentation performance was assessed using the Dice similarity coefficient (DSC) computed between predicted and ground-truth LV masks at ED and ES frames. EF estimation accuracy was characterized by the median absolute error (MedAE) and the proportion of cases falling within $\pm 10\%$ absolute EF of the ground-truth value — a threshold aligned with published inter-vendor tolerance guidelines for echocardiographic EF measurement⁵. Median rather than mean absolute error was chosen as the primary continuous metric because the EF error distribution exhibits a heavy tail attributable to a subset of outlier cases, and the median provides a more robust summary of central tendency in this setting. For the few-shot experiments, each shot level was repeated over five random seeds; Fig. 2 reports mean $\pm$ standard deviation across seeds. Table 2 reports the median absolute EF error point estimates for the specified model settings, with the corresponding seed-to-seed variability for the 20-shot EchoMed Agent condition shown in Fig. 2.

Declarations

Competing Interests

[Please confirm whether the authors have any competing financial or non-financial interests.]

Author Contribution

Leshan Sun developed the models, conducted the experiments, analyzed the results, and drafted the manuscript. Juan Farina performed clinical echocardiographic annotation and contributed to data collection. Songlin Zhao contributed to model development. Shuoyang Wei and Yuzhen Ding assisted with development of the agentic framework. Reza Arsanjani provided clinical guidance and contributed to project administration. Win-Kuang Shen provided clinical supervision and contributed to project administration. Wei Liu conceived and supervised the study, contributed to manuscript preparation, and serves as the corresponding author. All authors reviewed and approved the final manuscript.

Acknowledgement

The authors acknowledge the investigators and institutions responsible for the development and public release of the EchoNet-Dynamic and CAMUS datasets used in this study. This work received no specific external funding.

Data Availability

EchoNet-Dynamic and CAMUS are publicly available echocardiography datasets and can be accessed through their respective repositories. The de-identified Mayo Clinic cohort used in this study was assembled under IRB 25-013540 and is not publicly distributed. Access to the institutional data is subject to Mayo Clinic policies, applicable IRB requirements, and data-use agreements.

References

1. Gillam, L. D. & Marcoff, L. Echocardiography: past, present, and future. *Circulation: Cardiovascular Imaging* 17, e016517 (2024).
2. Keller, M., Magunia, H., Rosenberger, P. & Koeppen, M. Echocardiography as a tool to assess cardiac function in critical care—a review. *Diagnostics* 13, 839 (2023).
3. Nagueh, S. F. et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography and for heart failure with preserved ejection fraction diagnosis: an update from the American Society of Echocardiography. *Journal of the American Society of Echocardiography* 38, 537–569 (2025).
4. Potter, E. & Marwick, T. H. Assessment of left ventricular function by echocardiography. *JACC: Cardiovascular Imaging* 11, 260–274 (2018).
5. Lang, R. M. et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Journal of the American Society of Echocardiography* 28, 1–39.e14 (2015).
6. Shahgaldi, K., Gudmundsson, P., Manouras, A., Brodin, L. A. & Winter, R. Visually estimated ejection fraction by two dimensional and triplane echocardiography is closely correlated with quantitative ejection fraction by real-time three dimensional echocardiography. *Cardiovascular Ultrasound* 7, 41 (2009).
7. Wahr, D. W., Wang, Y. S. & Schiller, N. B. Left ventricular volumes determined by two-dimensional echocardiography in a normal adult population. *Journal of the American College of Cardiology* 1, 863–868 (1983).
8. Cole, G. D. et al. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. *International Journal of Cardiovascular Imaging* 31, 1303–1314 (2015).
9. Pellikka, P. A. et al. Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. *JAMA Network Open* 1, e181456 (2018).

10. Garcia-Elcano, L. et al. Accurate estimation of interobserver variability in echocardiography segmentation measured from 35 experts and impact on derived clinical metrics. *European Heart Journal - Cardiovascular Imaging* 26, Supplement_1 (2025).
11. Zhang, L. et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging* 39, 2531–2540 (2020).
12. Lei, Y. et al. Echocardiographic image multi-structure segmentation using Cardiac-SegNet. *Medical Physics* 48, 2426–2437 (2021).
13. Leclerc, S. et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Transactions on Medical Imaging* 38, 2198–2210 (2019).
14. Kirillov, A. et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026 (2023).
15. Luo, J. et al. Large language model agent: a survey on methodology, applications and challenges (2025).
16. Holmes, J. et al. RadOnc-GPT: an autonomous LLM agent for real-time patient outcomes labeling at scale (2025).
17. Cootes, T. F., Taylor, C. J., Cooper, D. H. & Graham, J. Active shape models—their training and application. *Computer Vision and Image Understanding* 61, 38–59 (1995).
18. Bosch, J. G. et al. Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Transactions on Medical Imaging* 21, 1374–1383 (2002).
19. Smistad, E. et al. Real-time automatic ejection fraction and foreshortening detection using deep learning. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 67, 2595–2604 (2020).
20. Wei, H. et al. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 623–632 (2020).
21. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 580, 252–256 (2020).
22. Tang, Y. et al. Self-supervised pre-training of Swin Transformers for 3D medical image analysis (2022).
23. Alsaleh, A. M., Albalawi, E., Algosaibi, A., Albakheet, S. S. & Khan, S. B. Few-shot learning for medical image segmentation using 3D U-Net and model-agnostic meta-learning (MAML). *Diagnostics* 14 (2024).
24. Ma, J., He, Y., Li, F., Han, L., You, C. & Wang, B. Segment anything in medical images. *Nature Communications* 15, 654 (2024).
25. Ravi, N. et al. SAM 2: Segment Anything in Images and Videos (2024).
26. Ma, J. et al. MedSAM2: Segment Anything in 3D Medical Images and Videos (2025).

27. Chao, C. J. et al. Foundation versus domain-specific models for left ventricular segmentation on cardiac ultrasound. *npj Digital Medicine* 8, 341 (2025).
28. Schick, T. et al. Toolformer: language models can teach themselves to use tools (2023).
29. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023).
30. Bhayana, R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology* 310, e232756 (2024).
31. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* 30, 1134–1142 (2024).
32. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems (2023).
33. Hu, E. J. et al. LoRA: low-rank adaptation of large language models (2021).
34. Schiller, N. B. et al. Recommendations for quantitation of the left ventricle by two-dimensional echocardiography. American Society of Echocardiography Committee on Standards, Subcommittee on Quantitation of Two-Dimensional Echocardiograms. *Journal of the American Society of Echocardiography* 2, 358–367 (1989).
35. Mayo Clinic. AI Factory. <https://docs.mcc.mayo.edu/docs/ai-factory>.
36. Carion, N. et al. SAM 3: Segment Anything with Concepts (2026).

Figures

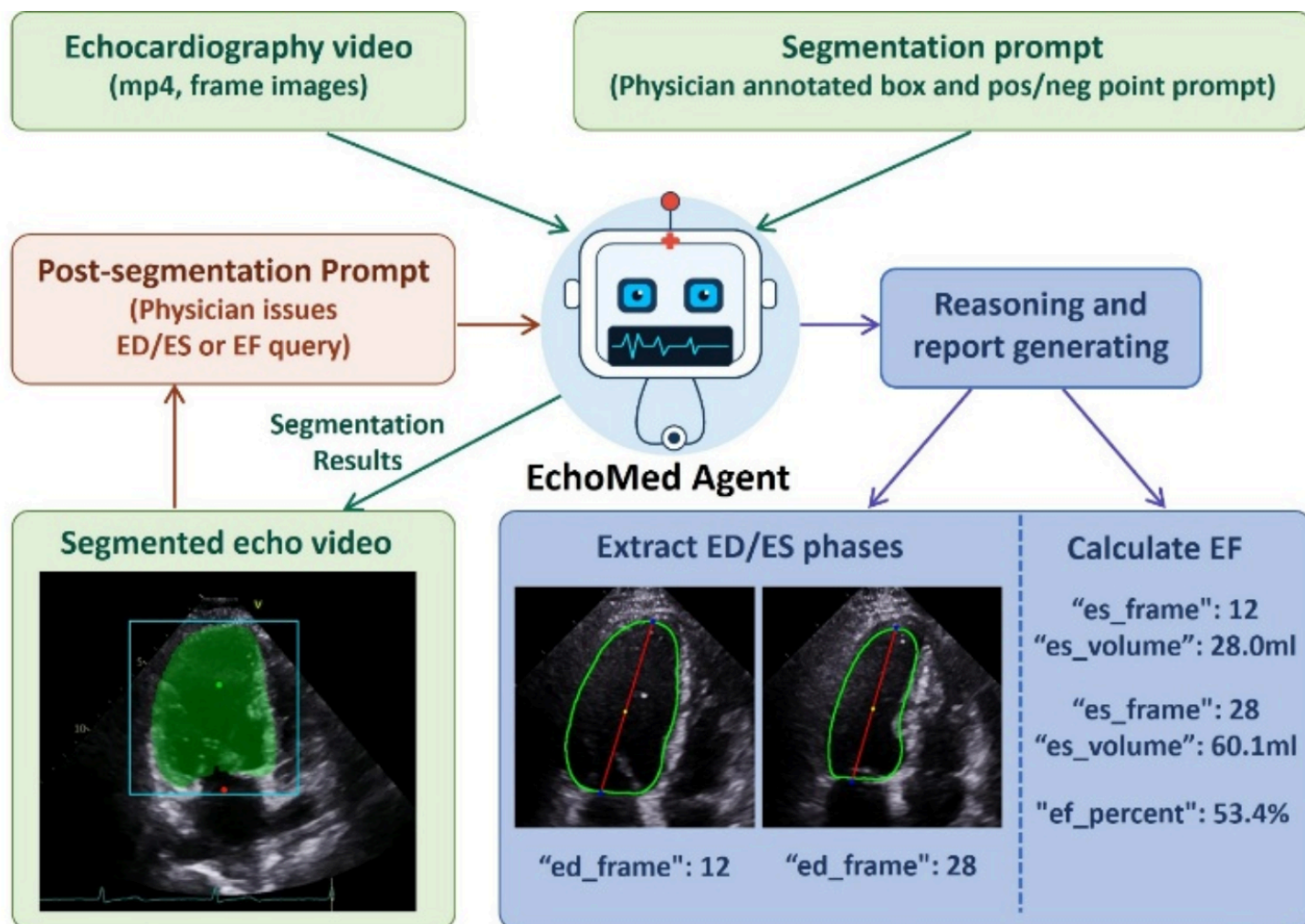
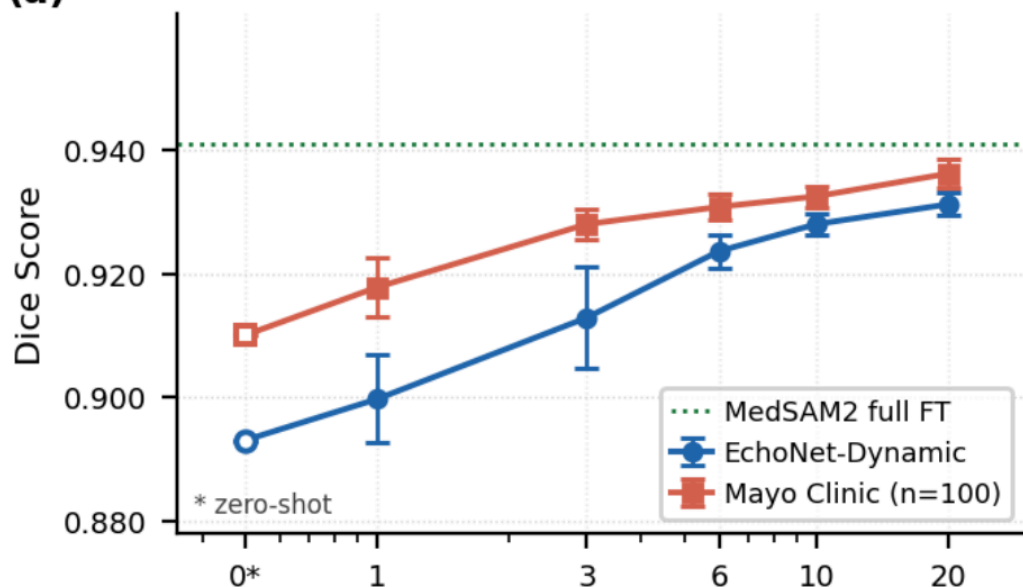


Figure 1

EchoMed Agent workflow. Echocardiography video and physician-issued prompts (bounding box and point annotations) are routed by a GPT-5.2-based agentic interface to the MedSAM2 segmentation backbone. The physician may then issue post-segmentation natural-language queries for ED/ES phase identification or EF estimation, which the agent resolves and compiles into a structured report.

Fig. 3. Few-shot scaling of EchoMed Agent

(a)



(b)

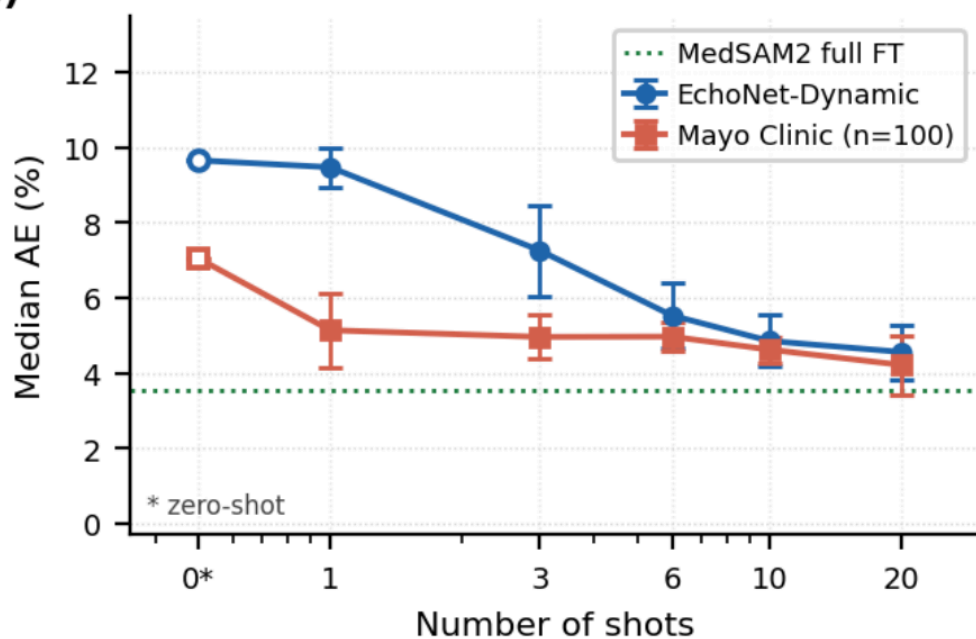


Figure 2

Few-shot scaling performance of EchoMed Agent. (a) Dice score and (b) median absolute EF error as a function of labeled adaptation examples (0–20 shots) on EchoNet-Dynamic and Mayo Clinic cohorts. The dotted line denotes MedSAM2 fine-tuned on the full EchoNet-Dynamic training set. Error bars indicate $\pm$ SD over five random seeds.

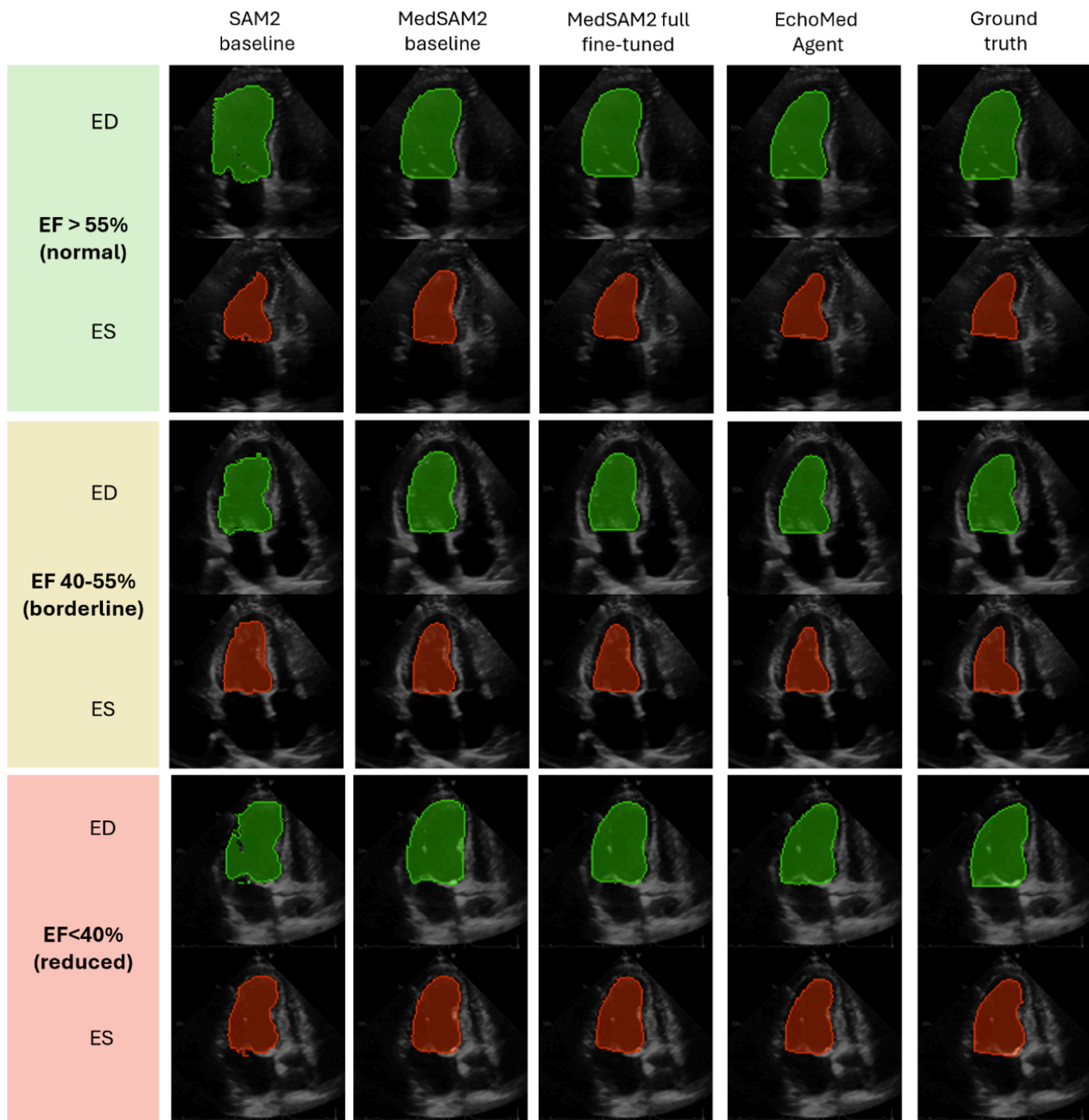


Figure 3

Qualitative LV segmentation comparison stratified by ejection fraction (EF) category. Each row pair shows end-diastolic (ED, green overlay) and end-systolic (ES, red overlay) frames for a representative case from three strata: normal (EF > 55%), borderline (EF 40–55%), and reduced (EF < 40%). Columns show SAM2 zero-shot, MedSAM2 zero-shot, MedSAM2 fully fine-tuned on EchoNet-Dynamic, EchoMed Agent (20-shot), and ground truth.

Figure 2 — LV segmentation, beat-to-beat mask area & volumetric analysis

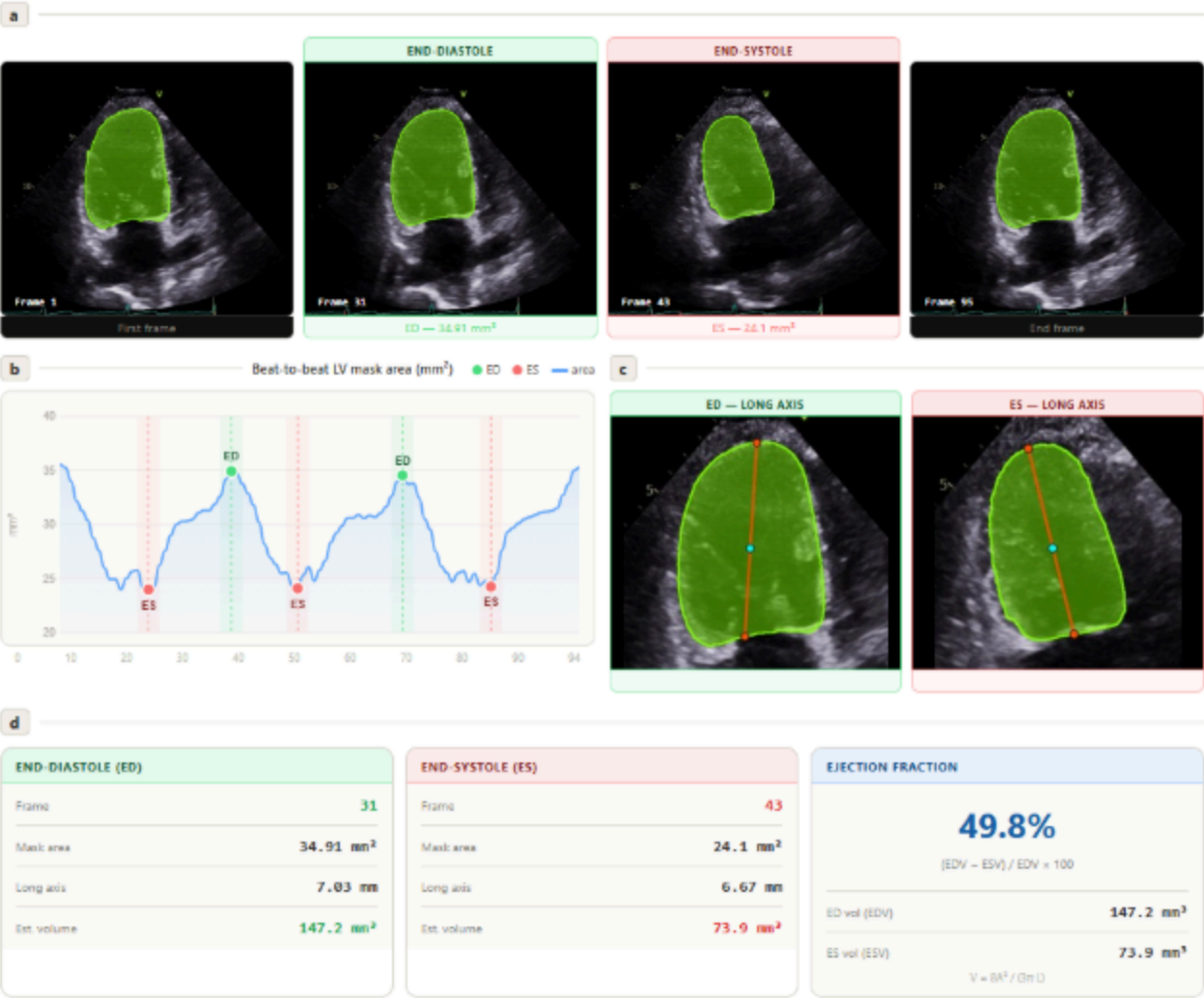


Figure 4

Representative EchoMed Agent output for a single echocardiographic case. (a) Frame-by-frame LV segmentation overlays (first, ED, ES, and final frames). (b) Beat-to-beat LV mask area curve with detected ED and ES markers. (c) ED and ES frames with predicted long-axis overlays. (d) Structured physician report summarizing detected ED/ES frame indices, LV mask areas, long-axis lengths, estimated EDV/ESV, and computed EF.

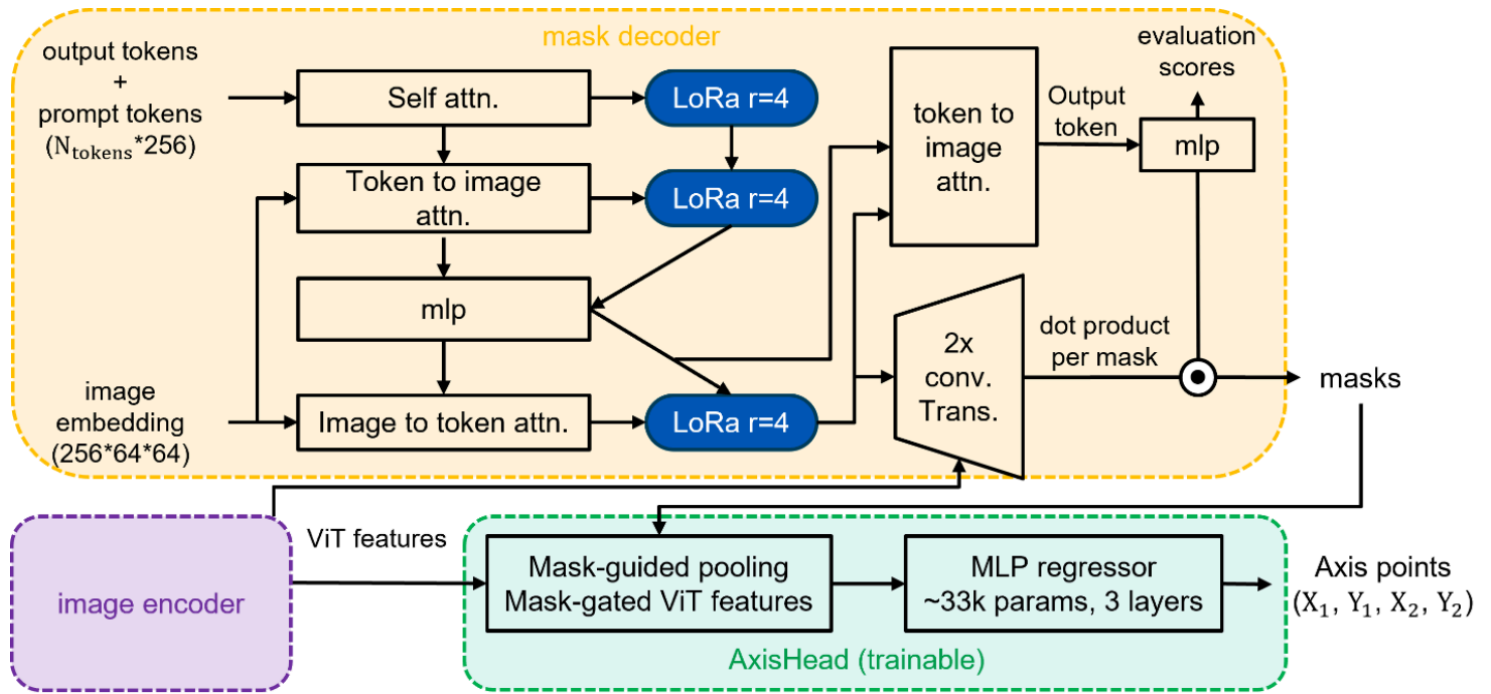


Figure 5

MedSAM2 mask decoder with LoRA fine-tuning and AxisHead integration. Rank-4 LoRA adapters are inserted after each attention layer for parameter-efficient fine-tuning with backbone weights frozen. AxisHead applies mask-guided pooling of frozen ViT image encoder features, followed by a lightweight MLP (~33K parameters), to regress the LV long-axis endpoint coordinates (X_1 , Y_1 , X_2 , Y_2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)