MUSE: MULTI-VIEW ECHOCARDIOGRAPHY SEGMENTATION VIA VIEW-ADAPTIVE EXPERTS AND BOUNDARY-AWARE MULTI-DOMAIN FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Echocardiography segmentation (ES) can substantially reduce radiologists' heavy clinical workload in image analysis. Despite recent advances, accurate multi-view segmentation remains challenging because echocardiograms are complex and ambiguous, whereas existing models mainly focus on a single echocardiographic view, limiting the clinical value of current models. To address these issues, we propose Multi-View Echocardiography Segmentation via View-Adaptive Experts and Boundary-Aware Multi-Domain Fusion (MUSE), a novel framework that aims to improve multi-view echocardiogram segmentation through view-adaptive expert paths. Specifically, MUSE adopts an encoder-decoder architecture. To handle ambiguous and complex echocardiographic images, MUSE constructs boundary priors from bottleneck features and performs multi-domain fusion through boundary-aware gating, which selectively enhances the wavelet, spatial-feature, and adaptive multiband Fourier branches. To further improve view-specific segmentation, we design a multi-expert routing mechanism that predicts view types from image features and employs shared and view-specific experts to extract common and distinctive features for segmentation. Additionally, to alleviate the limited view variety of existing datasets, we collect and annotate a multi-view, multi-structure echocardiography dataset, thereby improving clinical relevance through broader view coverage. Experimental results on two public datasets and one private dataset demonstrate the effectiveness of MUSE, which achieves state-of-the-art segmentation performance on both public and private datasets across diverse views and anatomical structures.

1 INTRODUCTION

Echocardiography is widely used for cardiovascular diagnosis and cardiac function assessment because it is non-invasive, accessible, inexpensive, and provides real-time cardiac imaging. Clinicians integrate multiple standard views to assess cardiac morphology, motion, and anatomical relationships, but manual interpretation and measurement require substantial expertise and effort. Automatic cardiac-structure segmentation can therefore reduce workload and support efficient echocardiographic analysis.

Deep learning has advanced echocardiographic segmentation. Early studies focused on left ventricular (LV) segmentation in a single view for estimating ventricular volume and ejection fraction. EchoNet-Dynamic (Ouyang et al., 2019) demonstrated automated LV segmentation and ejection-fraction estimation from apical four-chamber videos. Later methods used temporal modeling, self- or weakly supervised learning, and global context (Maani et al., 2024; Saeed et al., 2022), but usually addressed one structure and one view.Research then expanded to multiple structures and views. CAMUS (Leclerc et al., 2019) provides apical two- and four-chamber images with LV, myocardium, and left-atrium annotations, enabling evaluation of U-Net and U-Net++ (Ronneberger et al., 2015; Zhou et al., 2018). EchoONE (Hu et al., 2025) recently explored unified segmentation across imaging planes.

Nevertheless, clinically applicable multi-view, multi-structure segmentation faces two main challenges.First, standard echocardiographic views show substantial anatomical heterogeneity. Changes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

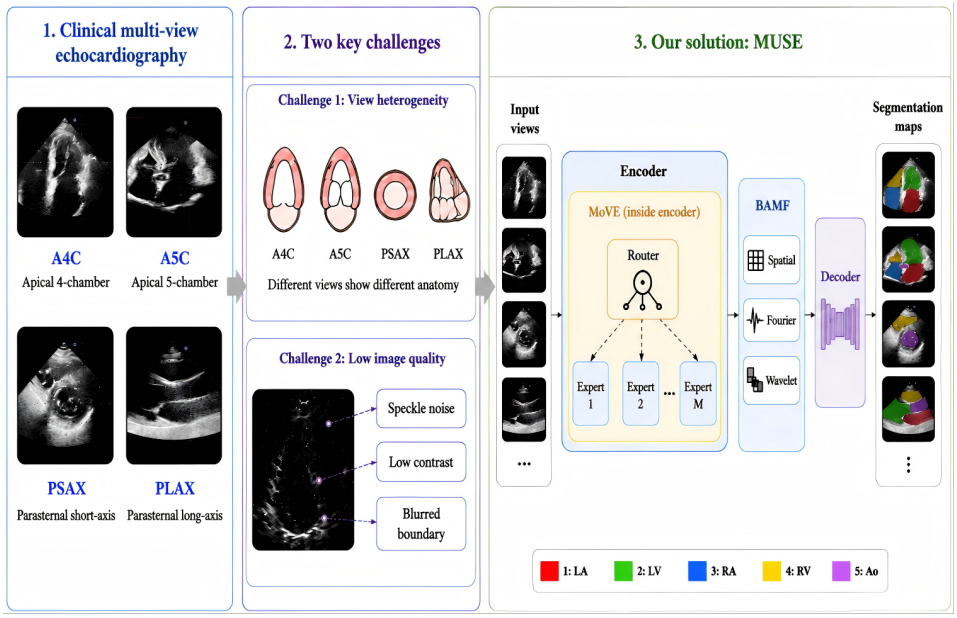


Figure 1: Motivation and overview of MUSE. View-adaptive expert specialization and boundary-aware multi-domain fusion address anatomical variation, speckle noise, and ambiguous boundaries in unified multi-view and multi-structure segmentation.

in probe position, insonation angle, and acquisition produce different two-dimensional appearances of the same heart. A2C, A4C, and short-axis views differ markedly in visible structures, shape, scale, and spatial arrangement. A fully shared model may overlook view-specific characteristics and suffer from interference between heterogeneous distributions, whereas separate models increase complexity and restrict knowledge sharing. A unified model must therefore capture both common anatomical knowledge and view-specific information.Second, echocardiograms commonly contain speckle noise, low contrast, intensity inhomogeneity, and ambiguous boundaries. Shallow skip connections can propagate low-level noise, while repeated downsampling weakens boundary and texture information. These problems are particularly serious when adjacent chambers, myocardium, vessels, and outflow tracts have poorly defined interfaces. Spatial features alone may not preserve global anatomy while recovering fine boundaries, motivating the joint use of spatial, frequency, and boundary information.

To address these challenges, we propose Multi-View Echocardiography Segmentation via View-Adaptive Experts and Boundary-Aware Multi-Domain Fusion (MUSE), a unified end-to-end framework for multi-view and multi-structure echocardiographic segmentation (Figure 1). MUSE combines view-adaptive feature specialization with boundary-aware multi-domain enhancement. Its Mixture of View Experts (MoVE) module estimates a soft view distribution and routes each sample to a sparse combination of experts. A shared expert learns anatomical patterns common across views, while view-specific experts capture residual characteristics of individual imaging planes. View supervision and load-balancing constraints promote stable specialization under imbalanced view distributions. The Boundary-Aware Multi-domain Fusion (BAMF) module improves deep features under low signal-to-noise ratios and uncertain boundaries. Its spatial branch preserves local semantic structure, its multiband Fourier branch models frequency components at different scales, and its learnable wavelet branch captures multiscale boundaries and directional textures. A boundary prior selectively gates wavelet responses to avoid amplifying speckle, after which competitive fusion integrates the three domains. Thus, MoVE models how cardiac views differ, while BAMF improves how ambiguous structures are represented.

Public benchmarks cover only part of the views and structures used clinically. EchoNet-Dynamic focuses on LV analysis in A4C images, and CAMUS includes A2C and A4C images. We therefore construct an in-house dataset covering five standard views and a broader set of structures, including both atria, both ventricles, great vessels, and outflow tracts. Experiments on CAMUS and our clin-

ical dataset show that MUSE achieves state-of-the-art or competitive region-overlap and boundary performance across heterogeneous views.

Our main contributions are: (1) We propose MUSE, a unified framework for shared, view-specific, and multi-domain representation learning; **(2)** We introduce MoVE for view-aware expert routing and BAMF for boundary-aware fusion of spatial, Fourier, and wavelet features; **(3)** We construct a clinically oriented multi-view, multi-structure dataset with broader view and anatomical coverage than existing public benchmarks; and **(4)** Experiments on two public datasets and one private dataset demonstrate the effectiveness and generalizability of our proposed method.

2 Related Work

Echocardiographic Segmentation Deep learning has substantially advanced automatic echocardiographic segmentation. Early work focused mainly on the left ventricle (LV), whose contour is required to estimate ventricular volume and ejection fraction. EchoNet-Dynamic (Ouyang et al., 2019) provides 10,030 apical four-chamber videos with expert LV annotations; its baseline combines DeepLabV3 (Chen et al., 2018) with temporal information for segmentation and ejection-fraction prediction. Subsequent methods use local attention, Transformer bridges, graph convolutions, attention-based denoising, or weak supervision to improve temporal context, global modeling, and annotation efficiency (Liu et al., 2021; Thomas et al., 2022; Chen et al., 2022). However, these methods generally segment only the LV in one view.

Research subsequently expanded to multiple cardiac structures. CAMUS (Leclerc et al., 2019) contains apical two- and four-chamber images from 500 patients, with annotations of the LV, myocardium, and left atrium at end diastole and end systole. It supports comparisons of U-Net (Ronneberger et al., 2015), U-Net++ (Zhou et al., 2018), and the two-stage LU-Net (Leclerc et al., 2020). Later methods further develop multiscale, residual, attention, and temporal representations (Lin et al., 2017; Li et al., 2020; Wei et al., 2023). Nevertheless, EchoNet-Dynamic covers one view and CAMUS only two. Existing multi-view methods often use separate branches followed by feature sharing or cross-view fusion (Liu et al., 2022; Martín-Isla et al., 2023; Mazurowski et al., 2023), increasing complexity and limiting generalization. In contrast, our unified model addresses view heterogeneity while learning from a dataset with broader view and anatomical coverage.

Mixture of Experts Model Mixture-of-Experts (MoE) models dynamically route inputs to specialized subnetworks. Early work by Jacobs et al. (Jacobs et al., 1991) established the framework, later reviewed by Yuksel et al. (Yuksel et al., 2012). Sparse gating activates only a few experts per input, increasing capacity at moderate computational cost (Shazeer et al., 2017). GShard and Switch Transformer (Lepikhin et al., 2020; Fedus et al., 2022) demonstrated this scalability, while Soft MoE (Puigcerver et al., 2024) alleviates instability and expert imbalance through soft routing.

MoE has also been extended to vision and multimodal learning. V-MoE (Riquelme et al., 2021) applies sparse experts to visual representations, and related work scales vision-language training (Shen et al., 2023). VLMo (Bao et al., 2022) combines modality-specific and shared paths, while MoAI, MoVA, and MoE-LLaVA (Lee et al., 2024; Zong et al., 2024; Lin et al., 2026) use expert modules for multimodal processing. Large models, including DeepSeek-V3 and Qwen3 (Liu et al., 2024; Yang et al., 2025), further show the effectiveness of top- and weighted expert fusion (Cai et al., 2024; Rajbhandari et al., 2022; Li et al., 2025).

Clinical MoE methods address heterogeneous data, missing modalities, and interpretability (Zhu et al., 2024; Shen et al., 2024; Zhang et al., 2025b), but rarely consider multi-view echocardiographic segmentation. Because standard views differ in anatomy and layout, MoVE routes features to shared and view-specific experts to combine common knowledge with specialized modeling.

3 Method

Unless otherwise specified, feature maps lie in $\mathbb{R}^{B\times C\times H\times W}$, where B, C, H, and W denote batch size, channels, height, and width; c, m, and i index channels, experts, and samples, respectively; and $\odot$, $\|$, and Concat denote element-wise multiplication, concatenation of two tensors, and concatenation of multiple tensors.

3.1 FRAMEWORK

As shown in Figure 2, the network comprises a MoVE-enhanced encoder, a BAMF bottleneck, and a decoder. Given an ultrasound image I, its prediction is

$$P = \mathcal{D}(\mathcal{T}(\mathcal{E}_{\text{MoVE}}(I))), \tag{1}$$

where $\mathcal{E}_{\text{MoVE}}$, $\mathcal{T}$, and $\mathcal{D}$ denote the encoder, BAMF, and decoder, respectively, and P is the segmentation prediction.

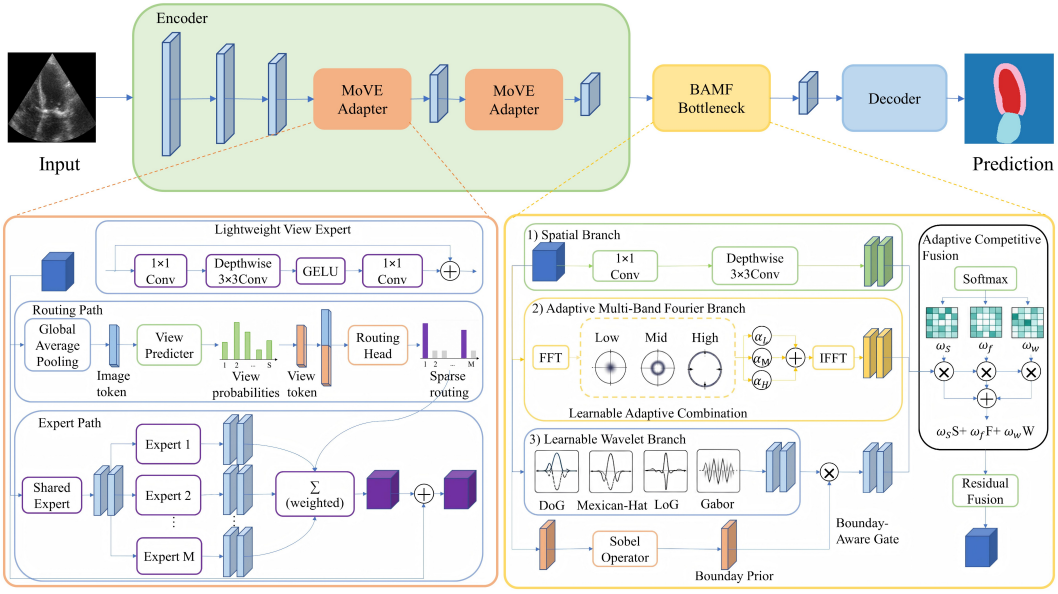


Figure 2: Overview of the proposed MUSE framework. MoVE is introduced into the encoder to model view-specific anatomical variations through adaptive expert routing, while BAMF is incorporated at the bottleneck to enhance ambiguous boundary representations through boundary-aware multi-domain fusion.

3.2 MIXTURE OF VIEW EXPERTS (MOVE)

Soft-View Awareness Cardiac views share common anatomy but retain view-specific appearance, making either fully independent or fully shared models suboptimal. MoVE therefore combines shared knowledge with view-specific residual specialization.

For an input feature F, we obtain an image token $z = \text{GAP}(\text{GN}(F))$ and blend its predicted view distribution with supervision:

$$\widetilde{q} = \rho y_s + (1-\rho)\,\text{softmax}(\mathcal{P}(z)/\tau_s), \tag{2}$$

where GN and GAP denote Group Normalization and global average pooling, $z \in \mathbb{R}^{B\times C}$ is the image token, $\widetilde{q} \in \mathbb{R}^{B\times S}$ is the guided distribution over S views, y_s is the one-hot view label, $\mathcal{P}$ is a two-layer predictor, τ_s is the view temperature, and $\rho \in [0,1]$ is a progressively reduced teacher-guidance coefficient. A continuous view token is then obtained as $v = \widetilde{q}V$, where $V \in \mathbb{R}^{S\times C}$ is a learnable embedding matrix and $v \in \mathbb{R}^{B\times C}$.

Sparse Expert Routing and Fusion The router combines the image and view tokens with a view–expert prior:

$$\pi = \text{softmax}\left(\frac{\mathcal{R}(z\|v) + \eta W_p \widetilde{q}}{\tau_r}\right), \tag{3}$$

where $\mathcal{R}$ is the routing network, $W_p \in \mathbb{R}^{M\times S}$ maps view probabilities to priors over M view-specific experts, η is a learnable prior strength, τ_r is the routing temperature, and $\pi \in \mathbb{R}^{B\times M}$ is the expert distribution. We retain and renormalize its top two entries:

$$\widetilde{\pi} = \text{TopKNorm}_K(\pi), \tag{4}$$

where $\mathrm{TopKNorm}_K$ sets all but the largest K entries to zero and renormalizes the retained entries, and $\widetilde{\pi}$ is the resulting sparse distribution. We set $K = 2$.

Each lightweight expert uses 1×1 compression, 3×3 depthwise convolution, and 1×1 restoration. The shared and routed expert residuals are directly integrated into the module output:

$$F_{\mathrm{out}} = F + \alpha\left[\mathcal{E}_{\mathrm{sh}}(\bar{F}) + \sum_{m=1}^{M} \widetilde{\pi}_m \mathcal{E}_m\big(\bar{F} + \mathcal{E}_{\mathrm{sh}}(\bar{F})\big)\right], \tag{5}$$

where $\bar{F} = \mathrm{GN}(F)$, $\mathcal{E}_{\mathrm{sh}}$ and $\mathcal{E}_m$ are the shared and m-th view-specific experts, $\widetilde{\pi}_m$ is the weight of expert m, α is a learnable scale bounded by $\alpha_{\max}$, and F_{out} is the enhanced feature.

3.3 Boundary-Aware Multi-domain Fusion (BAMF)

Local convolutions can lose fine boundaries and cannot efficiently capture global cardiac morphology. BAMF processes a bottleneck feature X through spatial, Fourier, and wavelet branches, producing S, F_f, and W, respectively, and uses boundary-aware competitive fusion to suppress speckle noise.

Spatial and Fourier Modeling The spatial branch applies a 1×1 pointwise convolution for channel mixing, followed by GroupNorm, GELU, and a 3×3 depthwise convolution for efficient local modeling. For global context, the Fourier branch combines low-, mid-, and high-frequency soft bands with channel-wise weights:

$$M_c^{\mathrm{eff}}(u,v) = \sum_{k\in\{L,M,H\}} \frac{\exp(\theta_{k,c})}{\sum_{j\in\{L,M,H\}}\exp(\theta_{j,c})} \cdot \exp\left[-\frac{(r(u,v)-\mu_k)^2}{2\sigma_k^2}\right], \tag{6}$$

where (u,v) is a frequency coordinate, $r(u,v)$ is its normalized radial frequency, $k \in \{L,M,H\}$ indexes the three bands, μ_k and σ_k are their centers and learnable bandwidths, $\theta_{k,c}$ is the channel-wise weight logit, and M_c^{eff} is the effective mask for channel c. The weighted spectrum is returned to the spatial domain:

$$(F_f)_c = \mathrm{IFFT}\big(\mathrm{FFT}(\mathrm{Conv}_{1\times1}(X))_c \odot M_c^{\mathrm{eff}}\big), \tag{7}$$

where $\mathrm{Conv}_{1\times1}$ is the input projection, FFT and IFFT are the two-dimensional Fourier transform and its inverse, and $(F_f)_c$ is channel c of the Fourier feature. The three bands mainly describe global morphology, regional transitions, and fine edges or noise, respectively.

Learnable Wavelet and Boundary Guidance To complement the limited spatial localization of Fourier modeling, the wavelet bank contains two-scale Difference-of-Gaussian, Mexican Hat, Laplacian-of-Gaussian, and two orthogonal Gabor kernels. Each kernel is initialized from its classical form and optimized end-to-end while extracting its response:

$$W_i = X *_{\mathrm{dw}} (K_i^0 + \Delta K_i), \tag{8}$$

where $i \in \{1,\dots,N\}$ indexes the N kernels, K_i^0 is the predefined initialization, ΔK_i is its learnable update, $*_{\mathrm{dw}}$ denotes depthwise convolution, and W_i is the corresponding response. The responses are fused as

$$W = \phi\big(\mathrm{GN}\big(\mathrm{Conv}_{1\times1}\big(\mathrm{Concat}_{i=1}^{N} W_i\big)\big)\big), \tag{9}$$

where ϕ denotes GELU and W is the fused wavelet feature.

Because wavelets respond to both boundaries and speckle, we construct an explicit boundary prior from the channel-averaged feature $\bar{X} = C^{-1}\sum_c X_c$:

$$E = \sqrt{(\bar{X} * K_x)^2 + (\bar{X} * K_y)^2}, \tag{10}$$

where K_x and K_y are horizontal and vertical Sobel kernels, $*$ denotes convolution, and E is the boundary prior. The gated wavelet feature is

$$\widetilde{W} = \sigma(\mathcal{F}_{\mathrm{gate}}(\mathrm{Concat}(S, F_f, W, E))) \odot W, \tag{11}$$

where $\mathcal{F}_{\mathrm{gate}}$ is a lightweight gating network, σ is the sigmoid function, and $\widetilde{W}$ is the boundary-focused wavelet feature.

Competitive Multi-Domain Fusion A lightweight network predicts pixel-wise competitive weights for the three domains:

$$(A_s, A_f, A_w) = \mathrm{Softmax}\Big(\mathcal{F}_{\mathrm{weight}}(\mathrm{Concat}(S, F_f, \widetilde{W}))\Big), \tag{12}$$

where $\mathcal{F}_{\mathrm{weight}}$ is the weight-prediction network and A_s, A_f, and A_w are the spatial, Fourier, and wavelet weights satisfying $A_s + A_f + A_w = 1$. Fusion and residual enhancement are performed jointly:

$$X_{\mathrm{out}} = X + \gamma \mathcal{F}_{\mathrm{proj}}\Big(A_s \odot S + A_f \odot F_f + A_w \odot \widetilde{W}\Big), \tag{13}$$

where $\mathcal{F}_{\mathrm{proj}}$ is a 1×1 convolution followed by GroupNorm, γ is a learnable scale initialized to a small value, and X_{out} is the BAMF output.

3.4 Joint Optimization Objective

Segmentation, view prediction, and expert balancing are jointly optimized through

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_s(P, Y) + \frac{\lambda_v}{L}\sum_{l=1}^{L} \mathrm{CE}(q^{(l)}, y_s) + \lambda_b M \sum_{m=1}^{M} I_m L_m, \tag{14}$$

where Y is the ground-truth mask, $\mathcal{L}_s$ combines cross-entropy and Dice loss, L is the number of supervised encoder stages, $q^{(l)} = \mathrm{softmax}(\mathcal{P}^{(l)}(z^{(l)})/\tau_s)$ is the predicted view distribution at stage l, $\mathcal{P}^{(l)}$ and $z^{(l)}$ are the corresponding predictor and image token, CE denotes cross-entropy, $I_m = B^{-1}\sum_i \pi_{i,m}$ and $L_m = B^{-1}\sum_i \mathbb{I}(m \in \Omega_2^{(i)})$ are the mean importance and load of expert m, $\Omega_2^{(i)}$ is the selected expert set for sample i, $\mathbb{I}$ is the indicator function, and λ_v and λ_b weight view supervision and expert balancing.

4 Experiment

4.1 Experimental Setup

Datasets We evaluate on CAMUS (Leclerc et al., 2019), EchoNet-Dynamic, and an in-house clinical dataset. CAMUS contains 500 patients with 2CH/4CH images annotated for LV, LA, and MYO at ED and ES. EchoNet-Dynamic contains 10,030 4CH videos with LV annotations. We use the official splits of both public datasets. The in-house dataset covers five standard views, denoted by PSAX-AV, A5C, A4C, PSAX-MV, and PLAX, with 1,240/155/155 training/validation/test images annotated for the left and right atria, left and right ventricles, aorta, right ventricular outflow tract, and pulmonary artery. Together, these datasets enable evaluation across different scales, views, anatomical structures, and both public benchmark and real-world clinical settings across varied clinical contexts.

Evaluation Metrics We report Dice, Intersection over Union (IoU), True Positive Rate (TPR), 95% Hausdorff distance (HD95), Normalized Surface Dice (NSD), and Average Symmetric Surface Distance (ASSD). Dice and IoU measure overlap, TPR measures sensitivity, and HD95, NSD, and ASSD assess boundary agreement. Higher Dice, IoU, TPR, and NSD and lower HD95 and ASSD indicate better performance.

Implementation Details All models are trained on one H100 GPU for 100 epochs with batch size 4, 256×256 inputs, and AdamW (Loshchilov & Hutter, 2017) at a learning rate of 3×10^{-4}. Soft Dice and cross-entropy losses have weights 1.0 and 0.5. MoVE additionally uses view-classification and expert-balancing losses weighted by 0.05 and 0.01, respectively. During training, a cosine-decay-based view label teacher forcing strategy is adopted, with the teacher forcing probability gradually decreasing from 1.0 to 0.1.

4.2 Comparative Experiments

We compared MUSE with three task-specific methods (MCBTNet (Zhang et al., 2025a), BeU2-Net (Meng et al., 2025), and EchoONE (Hu et al., 2025)) and five relevant methods (U-Net++ (Zhou

Table 1: Comparison results on CAMUS and in-house dataset. Best results are in bold and second-best results are underlined.

Method	CAMUS					In-house dataset				
	mDice↑	mIoU↑	HD95↓	ASSD↓	NSD↑	mDice↑	mIoU↑	HD95↓	ASSD↓	NSD↑
Swin-UNet	0.8997	0.8224	7.6035	2.6075	0.4451	0.9126	0.8734	15.3939	13.3327	0.7740
Segformer	0.8899	0.8076	7.2709	2.7812	0.4073	0.9280	0.8859	7.8455	5.7543	0.7728
MCBTNet	0.8971	0.8118	7.0630	2.4829	0.4769	0.8665	0.8194	22.0935	19.6455	0.7256
VM-Unet	0.9067	0.8330	6.3163	2.3868	0.4459	0.9208	0.8799	11.4704	9.3721	0.7758
UNet++	0.9106	0.8395	6.6409	2.3264	0.4770	0.9297	0.8911	9.0744	6.9996	0.7989
DeepLabV3Plus	0.9089	0.8370	6.3426	2.3392	0.4612	0.9187	0.8758	8.4772	6.5854	0.7843
BeU2-Net	0.9106	0.8397	**6.1277**	2.2902	0.4700	0.9135	0.8699	11.0761	8.9023	0.7830
EchoONE	0.9081	–	–	–	–	–	–	–	–	–
MUSE	**0.9121**	**0.8420**	6.1347	**2.2577**	**0.4822**	**0.9392**	**0.9018**	**6.5540**	**4.5981**	**0.8116**

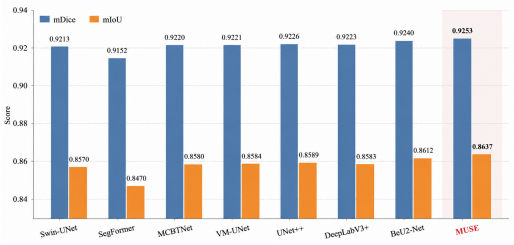


Figure 3: Experimental results on the EchoNet-Dynamic dataset.

Table 2: Ablation study of MoVE and BAMF. Best results are in bold.

MoVE	BAMF	mDice↑	mIoU↑	HD95↓	ASSD↓	NSD↑
×	×	0.9009	0.8258	7.6878	2.6158	0.4512
✓	×	0.9095	0.8385	6.8056	2.3458	0.4742
×	✓	0.9095	0.8383	6.4918	2.3405	0.4712
✓	✓	**0.9121**	**0.8420**	**6.1347**	**2.2577**	**0.4822**

et al., 2018), Segformer (Xie et al., 2021), DeepLabV3Plus (Chen et al., 2018), Swin-UNet (Cao et al., 2022), and VM-UNet (Ruan et al., 2024)).

These previous studies have achieved state-of-the-art medical video segmentation by employing multiscale feature fusion, boundary-aware mechanisms, attention mechanisms, and efficient global context modeling. However, most of these methods still face challenges in handling cross-view and capturing complex visual information specific to each view. MUSE addresses these limitations by introducing the MoVE module into the encoder layer of the framework, thereby improving upon the performance of traditional echocardiography segmentation models, which typically struggle with multi-view, multi-object segmentation tasks. Table 1 presents the experimental results on the CAMUS and private datasets. As shown in the table, MUSE achieves the best overall performance among methods reporting all five metrics and surpasses EchoONE in terms of mDice. Although its HD95 on CAMUS is marginally higher than that of BeU2-Net by only 0.007, possibly because HD95 is sensitive to a small number of localized boundary deviations, MUSE achieves the best ASSD and NSD, demonstrating more accurate and consistent boundary delineation overall.

As shown in Figure 3, MUSE achieved the best mDice and mIoU on the single-view EchoNet-Dynamic dataset. Together with its strong multi-view performance, these results confirm its adaptability and generalization across different data formats.

4.3 Ablation Experiments

To thoroughly evaluate the effectiveness of our proposed MoVE module and the BAMF module, we conducted ablation experiments on a test set based on the CAMUS dataset. The results are shown in Table 2:

Effectiveness of MoVE When MoVE is used alone, the performance improves consistently across all evaluation metrics. Compared with the baseline, MoVE increases mDice, mIoU, and NSD from 0.9009, 0.8258, and 0.4512 to 0.9095, 0.8385, and 0.4742, corresponding to relative improvements of 0.95%, 1.54%, and 5.10%, respectively. Meanwhile, HD95 and ASSD decrease from 7.6878 and 2.6158 to 6.8056 and 2.3458, representing reductions of 11.47% and 10.32%, respectively. These results demonstrate that MoVE can effectively assign view-specific experts according to the

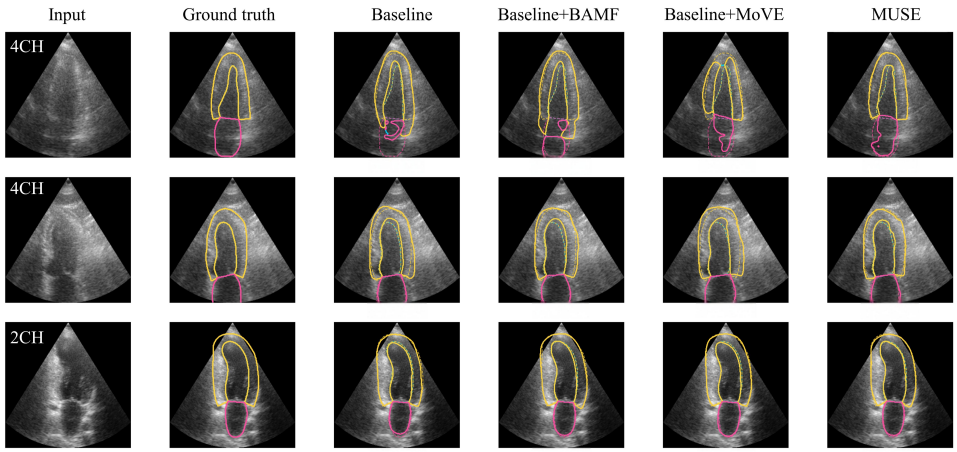


Figure 4: Qualitative visualization of ablation study results. Solid and dashed lines denote predicted and ground-truth contours, respectively. The red, blue, and yellow lines correspond to LA, LV, and MYO contours.

anatomical characteristics of different echocardiographic views, thereby enhancing both region-level segmentation accuracy and boundary localization. When BAMF is further incorporated, all five metrics are improved again, indicating that view-adaptive expert modeling and boundary-aware multi-domain feature fusion are complementary.

Effectiveness of BAMF When BAMF is used alone without MoVE, mDice, mIoU, and NSD improve from 0.9009, 0.8258, and 0.4512 to 0.9095, 0.8383, and 0.4712, yielding relative improvements of 0.95%, 1.51%, and 4.43%, respectively. More substantial improvements are observed for the boundary-related metrics: HD95 decreases from 7.6878 to 6.4918, while ASSD decreases from 2.6158 to 2.3405, corresponding to reductions of 15.56% and 10.52%, respectively. This confirms that BAMF effectively enhances boundary-sensitive representations by integrating spatial, frequency-domain, and boundary information at the bottleneck. Such feature enhancement enables the decoder to recover more accurate anatomical contours, particularly in regions affected by speckle noise and ambiguous boundaries.

Visual Analysis of Ablation Experiments We provide qualitative comparisons of the ablation experiments, as shown in Figure 4. Three representative cases from different echocardiographic views with varying segmentation difficulties are presented. Qualitative results show that adding MoVE or BAMF individually can already reduce the deviation between the predicted and ground-truth contours to some extent, whereas the predicted boundaries of the complete model are generally more consistent with the annotated contours.

The quantitative and qualitative results above collectively demonstrate that MoVE and BAMF improve feature representation at two distinct levels—view-adaptive modeling and multi-domain boundary enhancement, respectively—and that their synergistic effect enhances the model's segmentation accuracy and stability in complex ultrasound scenarios.

4.4 Model Explainability Analysis

To verify the working mechanisms of the proposed expert routing mechanism and the BAMF module, we conduct explainability experiments for expert routing and BAMF. In all experiments, teacher forcing for view labels was disabled during the testing phase, allowing the model to autonomously perform view recognition and expert assignment based solely on the input images.

Expert Routing Interpretability Analysis To further analyze whether the view-adaptive expert routing module can learn the feature differences between different views, this paper calculated the expert routing weights corresponding to different views based on samples from the test set and plotted an expert routing heatmap, as shown in Figure 5. The horizontal axis of the figure represents different expert branches, while the vertical axis represents different view numbers. The shade of

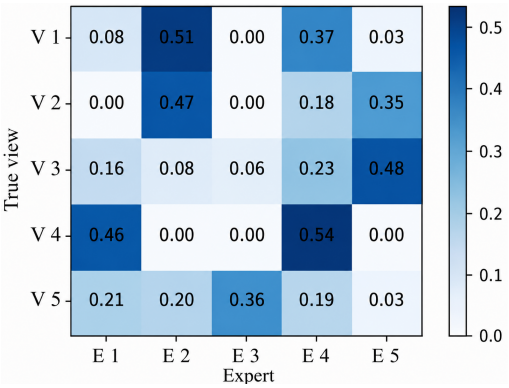


Figure 5: Expert routing heatmap visualization on the private dataset.

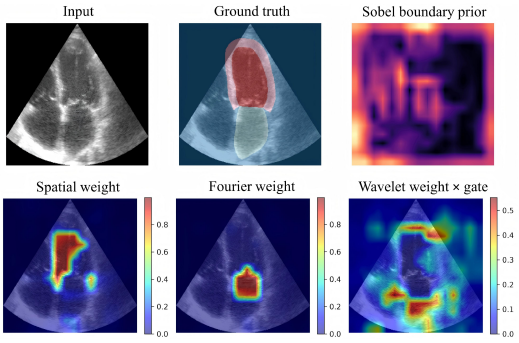


Figure 6: Qualitative visualization of boundary gate responses.

color indicates the average sparse routing weight of samples from that view assigned to the corresponding expert; the darker the color, the higher the weight assigned to that expert for that view.

Based on the routing results from the private dataset, it can be seen that there are significant differences in expert selection across different views. The dominant pairs for views 1–5 are E1/E3 (0.51/0.37), E1/E4 (0.47/0.35), E3/E4 (0.23/0.48), E0/E3 (0.46/0.54), and E1/E3 (0.21/0.36), respectively. In summary, the expert routing heatmap validates the design rationality of the cross-view adaptive expert routing module from the perspective of interpretability. The model does not apply a fixed feature transformation to all samples but instead adaptively assigns different expert branches based on the cross-view attributes of the input images. The differentiated responses of different experts across different cross-views indicate that the expert branches have learned feature compensation patterns with cross-view preferences, thereby alleviating the difficulty a unified encoder faces in adapting to distribution differences across multiple cross-views.

Explainability Analysis of the BAMF Module Figure 6 visualizes spatial, Fourier, and wavelet fusion weights across cases and views. Their distinct spatial distributions and limited overlap of high-response regions suggest region-adaptive feature selection rather than fixed-ratio fusion.

Notably, the response of the wavelet gating module in regions with strong edges is lower than in non-edge regions, suggesting that this module does not unconditionally enhance all high-frequency components but rather suppresses speckle noise and false edges while selecting wavelet features.

Based on these results, it can be concluded that BAMF improves the accuracy and robustness of echocardiographic segmentation through the complementary fusion of spatial localization, global frequency modeling, and multiscale boundary description.

5 Conclusion

In this study, we propose a multi-view expert-based echocardiography segmentation framework to address the issue of inaccurate multi-view segmentation: the classification results of each view are used to guide experts in selecting the appropriate view experts. Additionally, we designed the BAMF module, which enhances feature representations in bottleneck layers by combining boundary-aware gating with multi-domain fusion techniques, thereby further improving segmentation accuracy. Furthermore, we constructed a clinically oriented multi-view, multi-structure echocardiography dataset with broader view and anatomical coverage than existing public benchmarks. Experiments conducted on two datasets fully demonstrate the superiority of our method, particularly its advantages in segmenting multi-view, multi-structure echocardiographic segmentation, effectively bridging the gap between existing echocardiogram segmentation models and actual clinical needs.

AI USE STATEMENT

Large Language Models (LLMs) were used solely to assist with polishing the writing of this paper, such as improving readability and clarity of exposition. All ideas, experimental designs, implementations, analyses, and conclusions are original contributions of the authors.

REFERENCES

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in neural information processing systems*, 35:32897–32912, 2022.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. 2024.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Erna Chen, Zemin Cai, and Jian-huang Lai. Weakly supervised semantic segmentation of echocardiography videos via multi-level features selection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 388–400. Springer, 2022.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Jiongtong Hu, Wufeng Xue, Jun Cheng, Yingying Liu, Wei Zhuo, and Dong Ni. Echoone: segmenting multiple echocardiography planes in one model. In *Proceedings of the computer vision and pattern recognition conference*, pp. 5207–5216, 2025.

RA Jacobs, MI Jordan, SJ Nowlan, and GE Hinton. Adaptive mixtures of local experts, Neural Computation, 3(1):79–87. *Neural Computation*, 3(1):79–87, 1991.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Sarah Leclerc, Erik Smistad, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Mourad Belhamissi, Sardor Israilov, Thomas Grenier, et al. Lu-net: a multistage attention network to improve the robustness of segmentation of left ventricular structures in 2-d echocardiography. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(12):2519–2530, 2020.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. In *European Conference on Computer Vision*, pp. 273–302. Springer, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Ming Li, Shizhou Dong, Zhifan Gao, Cheng Feng, Huahua Xiong, Wei Zheng, Dhanjoo Ghista, Heye Zhang, and Victor Hugo C de Albuquerque. Unified model for interpreting multi-view echocardiographic sequences without temporal information. *Applied Soft Computing*, 88:106049, 2020.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International conference on medical image computing and computer-assisted intervention*, pp. 485–495. Springer, 2022.

Fei Liu, Kun Wang, Dan Liu, Xin Yang, and Jie Tian. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Medical image analysis*, 67:101873, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Fadillah Maani, Asim Ukaye, Nada Saadi, Numan Saeed, and Mohammad Yaqub. Simlvseg: simplifying left ventricular segmentation in 2-d+ time echocardiograms with self-and weakly supervised learning. *Ultrasound in Medicine & Biology*, 50(12):1945–1954, 2024.

Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: the m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3302–3313, 2023.

Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

Yuanqin Meng, Shengjie Chai, Haoyu Xiao, Zhaohui Meng, Qingwang Wang, and Tao Shen. BeU2-Net: Boundary-enhanced u2-net for simultaneous four-chamber segmentation in transthoracic echocardiography. *IEEE Journal of Biomedical and Health Informatics*, 29(6):4227–4240, 2025.

David Ouyang, Bryan He, Amirata Ghorbani, Matt P Lungren, Euan A Ashley, David H Liang, and James Y Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, volume 5, pp. 2, 2019.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pp. 28435–28445, 2024.

Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pp. 18332–18346. PMLR, 2022.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

Mohamed Saeed, Rand Muhtaseb, and Mohammad Yaqub. Contrastive pretraining for echocardiography segmentation with limited data. In *Annual conference on medical image understanding and analysis*, pp. 680–691. Springer, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2024.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11329–11344, 2023.

Sarina Thomas, Andrew Gilbert, and Guy Ben-Yosef. Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 380–390. Springer, 2022.

Hongrong Wei, Junqiang Ma, Yongjin Zhou, Wufeng Xue, and Dong Ni. Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences. *Medical Image Analysis*, 84:102686, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Boheng Zhang, Zelin Zheng, Yanqi Zhao, Yi Shen, and Mingjian Sun. Mcbtnet: Multi-feature fusion cnn and bi-level routing attention transformer-based medical image segmentation network. *IEEE Journal of Biomedical and Health Informatics*, 2025a.

Wei Zhang, Wenxin Xu, Tong Chen, Collin Sakal, and Xinyue Li. Integrating images and genomics for multi-modal cancer survival analysis via mixture of experts. *Information Fusion*, pp. 103521, 2025b.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pp. 3–11. Springer, 2018.

Xun Zhu, Ying Hu, Fanbin Mo, Miao Li, and Ji Wu. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-moe. *Advances in Neural Information Processing Systems*, 37:81225–81256, 2024.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37:103305–103333, 2024.

A Appendix

A.1 Datasets

We evaluate MUSE on two public benchmarks and one in-house clinical dataset. CAMUS contains 2D echocardiograms from 500 patients. For each patient, apical two-chamber (2CH) and apical four-chamber (4CH) images are provided at end diastole (ED) and end systole (ES), together with pixel-level annotations of the left ventricular cavity (LV), myocardium (MYO), and left atrium (LA). We follow the official patient-level split of 400/50/50 patients for training, validation, and testing, corresponding to 1,600/200/200 annotated images, respectively.

EchoNet-Dynamic contains 10,030 apical four-chamber echocardiography videos from independent subjects and provides expert contour annotations for left ventricular ejection fraction, volume measurements, and key frames for end-diastole (ED) and end-systole (ES). We follow the official video-level split, extract the ED and ES frames from each valid video, and rasterize the expert contours to obtain binary left-ventricular segmentation masks.

In-house dataset. We additionally collected and annotated an in-house dataset of real-world clinical echocardiographic examinations. It contains 1,550 2D frames spanning five echocardiographic views and provides pixel-level annotations for seven foreground structures: the left atrium (LA), left ventricle (LV), right atrium (RA), right ventricle (RV), aorta (Ao), right-ventricular outflow tract (RVOT), and pulmonary artery (PA). Because visible anatomy varies across imaging planes, only structures observable in each frame were annotated. The dataset was split into 1,240/155/155 images for training, validation, and testing, respectively.

Preprocessing and data augmentation. For all three datasets, images were converted to single-channel grayscale and resized to 256×256 pixels; linear interpolation was used for images and nearest-neighbor interpolation for discrete segmentation masks. CAMUS and in-house dataset images were normalized using per-image min–max normalization, while EchoNet-Dynamic images were normalized to the range $[0, 1]$ by dividing by 255. During training, data augmentation was applied only to the in-house training set, including random affine transformations, grayscale intensity perturbations, Gaussian noise, and Gaussian blurring; no augmentation was applied during validation.

A.2 Explainability Analysis of the BAMF Module

To examine the behavior of the frequency–wavelet–spatial fusion module, we analyzed its branch fusion weights and frequency-band responses on 200 CAMUS test samples, as shown in Figure 7. The average fusion weights of the spatial, Fourier, and wavelet branches were 0.074, 0.038, and 0.888, respectively, indicating that bottleneck enhancement is driven primarily by the wavelet branch, with the spatial and Fourier branches providing complementary information. The high weight of the wavelet branch suggests that local grayscale variations, tissue texture, and structural edges in echocardiograms are particularly important for fine-grained segmentation; the spatial branch primarily supplements information on the main cardiac chambers and local semantic details, while the Fourier branch aids in modeling the overall morphology and frequency structure. The standard errors of the weights for all three branches were less than 0.001, indicating good consistency of this response pattern across different test samples. Before fusion, however, wavelet features are modulated by a boundary gate whose mean response is 0.149. Thus, the high wavelet fusion weight does not imply uniform enhancement across spatial locations; instead, the boundary gate selectively activates informative regions while suppressing spurious high-frequency responses caused by ultrasound speckle noise.

To characterize the frequency preferences of the Fourier branch, we averaged its learnable low-, mid-, and high-frequency weights across all feature channels, as shown in Figure 8. The average weights for the low-, mid-, and high-frequency bands were 0.320, 0.331, and 0.349, respectively. The three bands receive broadly balanced weights, with a slight preference for high-frequency information. Low-frequency components primarily describe the overall morphology and large-scale structure of the cardiac chambers; mid-frequency components reflect tissue contours and transitional information between regions; and high-frequency components are closely related to myocardial boundaries,

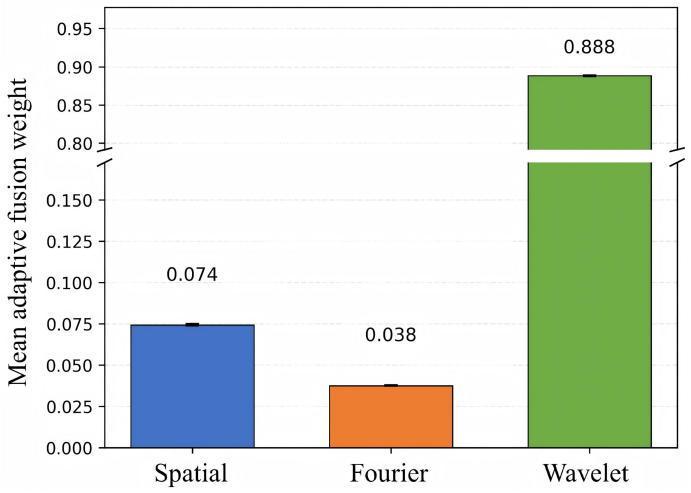


Figure 7: Average fusion weights of the spatial, Fourier, and wavelet branches.

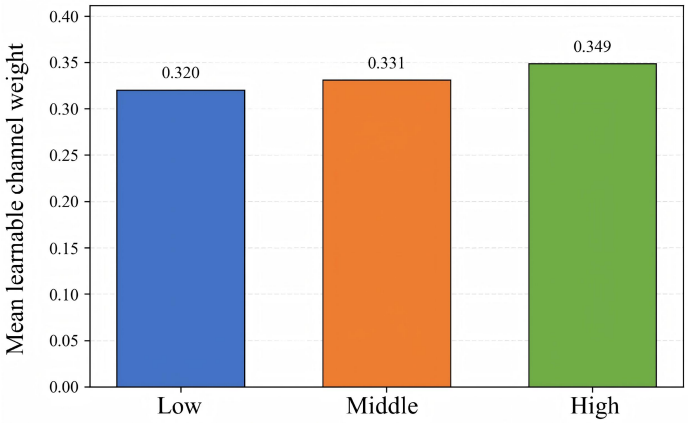


Figure 8: Average learnable channel weights for the low-, mid-, and high-frequency bands in the Fourier branch.

local texture, and subtle structural changes. These results show that the Fourier branch integrates complementary frequency components rather than favoring a single band, thereby providing a more complete structural representation for segmentation.

A.3 Additional Qualitative Results

To further assess MUSE on multi-view, multi-class echocardiographic segmentation, we visualize six representative samples from the in-house test set in Figure 9. Each row corresponds to one test image, while the columns show the original image, manual annotations, MUSE predictions, and predictions from the competing models. The selected samples cover five different cardiac views and encompass challenging scenarios involving multi-chamber structures, slender structures, small targets, and blurred boundaries.

As shown for Sets 1 and 6, these images contain multiple adjacent anatomical structures with small grayscale differences between tissues, and some boundaries are blurred due to ultrasonic speckle noise. Several competing models exhibit local contour shifts; notably, MCBTNet shows substantial regional deformation and confusion between adjacent classes in Sets 1 and 6. In contrast, MUSE accurately recovers the positions and morphologies of various structures; its predicted contours are closer to the manual annotations, and it maintains clear separation between different categories. This result demonstrates MUSE's ability to distinguish adjacent tissues under low-contrast condi-

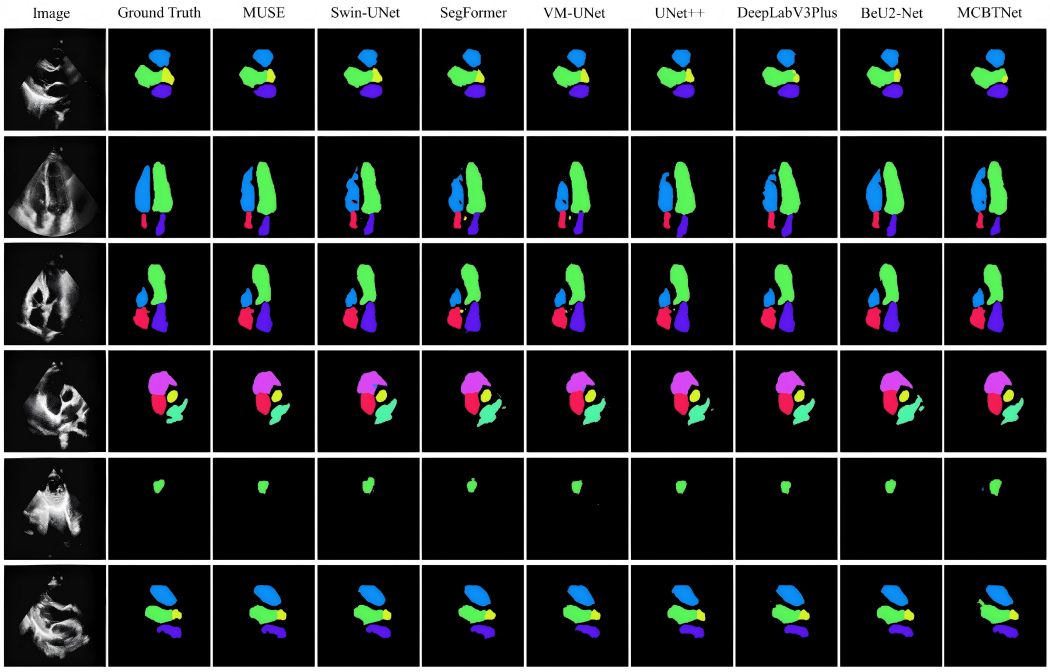


Figure 9: Visual comparison of multi-class cardiac structure segmentation results on the in-house dataset. Black denotes the background, purple the left atrium (LA), green the left ventricle (LV), red the right atrium (RA), blue the right ventricle (RV), yellow the aorta (Ao), magenta the right ventricular outflow tract (RVOT), and cyan-green the pulmonary artery (PA).

tions. Sets 2 and 3 highlight the performance of the proposed method in multi-chamber structures and fine-grained boundary restoration. In Set 2, the target structures are elongated; some comparison models exhibit internal voids, boundary indentations, and localized segmentation omissions. MUSE, however, maintains the continuity of major cardiac chambers and more completely preserves elongated structures. In Set 3, all models accurately locate the larger cardiac chambers overall, but some models still exhibit contour shifts at the junctions of adjacent structures. MUSE produces more complete segmentation regions, and the relative positions and contour shapes across all categories are highly consistent with the manual annotations, indicating strong structural modeling capabilities in complex multi-chamber scenarios.

Sets 4 and 5 further highlight the differences among models in scenarios involving irregular structures and small target segmentation. Set 4 contains structures with irregular shapes and relatively tortuous boundaries; some competing models produce scattered regions, local discontinuities, or boundary spillover to varying degrees around the targets. In contrast, MUSE's predictions are more continuous, reducing isolated pixels and regional fragmentation while maintaining category consistency across different anatomical structures. The targets in Set 5 are small in area and prone to interference from background noise. Although most models identify the primary targets, VM-UNet and MCBTNet produce additional false-positive regions, whereas MUSE localizes small targets with high accuracy and effectively suppresses erroneous responses in the background, demonstrating greater robustness to small-scale structures and complex backgrounds.

Overall, the qualitative results show that MUSE performs consistently in terms of boundary continuity, small-target localization, structural integrity, and differentiation between adjacent classes. These results suggest that slice-adaptive expert routing dynamically adjusts feature representations to account for anatomical differences across slices, while the boundary-aware multi-domain fusion mechanism enhances blurred edges and local details. Consequently, MUSE achieves a strong balance in terms of regional overlap accuracy, boundary alignment, and cross-slice stability.