E^{3}SAM2: Entropy-Aware and Edge-Guided Adaptation of SAM2 for Echocardiography Video Segmentation

Long Zheng, Zhi Li *, Weidong Wang, Zhenyu Dai, Shuyun Li

State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, China
gs.lzheng23@gzu.edu.cn, zhili@gzu.edu.cn, gs.wdwang24@gzu.edu.cn, zydai@gzu.edu.cn, gs.shuyunli23@gzu.edu.cn

Abstract

Foundation segmentation models, such as SAM and its video-oriented variant SAM2, have achieved remarkable success in natural image and video segmentation. However, their direct application to echocardiography video is challenged by structural uncertainty arising from severe speckle noise and blurry anatomical boundaries. To address this, we propose E^{3}SAM2, a lightweight adaptation framework that introduces a novel entropy-based methodology to explicitly model and mitigate such uncertainty. Specifically, an entropy-guided attention mechanism is introduced to steer the model's focus toward structurally reliable features, particularly in speckle-dominated regions. Additionally, an entropy regularization loss is introduced to further enhance target-background discrimination. To better resolve indistinct anatomical contours, an edge-aware supervision module is incorporated to inject explicit boundary priors for sharper delineation. These components are efficiently integrated through a global-local feature adapter. Experiments on CAMUS and EchoNet-Dynamic datasets demonstrate that E^{3}SAM2 achieves state-of-the-art segmentation and clinical estimation performance, while maintaining high computational efficiency.

Code — https://github.com/ZhengLong777/E3SAM2

Introduction

Cardiovascular diseases account for nearly one-third of global deaths (Chen et al. 2020), highlighting the urgent need for precise and efficient diagnostic tools such as echocardiography. Echocardiography video analysis is essential for diagnosing and monitoring cardiovascular diseases, offering a non-invasive assessment of cardiac function and structure (Akkus et al. 2021). Accurate segmentation of cardiac chambers, particularly the left ventricle (LV), is essential for computing clinical metrics such as ejection fraction (EF) and volume estimates (Ouyang et al. 2020; Cameli et al. 2016). However, echocardiography videos present unique challenges (Figure 1) distinct from natural videos: (1) severe speckle noise that obscures reliable anatomical features, (2) low tissue contrast leading to blurry anatomical boundaries, (3) large deformations due to cardiac motion, and (4) limited annotated data, which necessitates effective sparse-supervision methods (Biradar, Dewal, and Rohit 2015). These challenges pose significant obstacles to accurate echocardiography video segmentation, requiring tailored solutions that handle structural uncertainty arising from severe speckle noise and blurry anatomical boundaries.

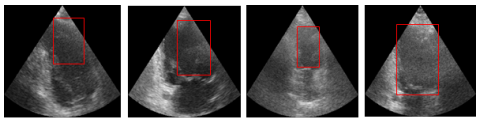

Figure 1: The challenges of echocardiography video segmentation: (a-b) speckle noise, (c-d) blurred contours.

Recent advances in foundation segmentation models, particularly the Segment Anything Model (SAM) (Kirillov et al. 2023) and its memory-augmented successor SAM2 (Ravi et al. 2024), have revolutionized segmentation tasks. SAM2, equipped with a streaming memory mechanism, is particularly suited for temporal consistency modeling in video segmentation. However, its direct application to echocardiography video yields suboptimal results due to its pre-training on natural images, which fails to account for the echocardiography video segmentation challenges of severe speckle noise and blurry anatomical boundaries. These limitations often result in inaccurate boundary delineation and structurally unreliable segmentations.

Recent efforts, such as MemSAM (Deng et al. 2024), have adapted SAM for echocardiography video segmentation by introducing a temporally aware prompting scheme with space-time memory to enhance time-space consistency. However, this approach faces two key limitations. First, it focuses on refining downstream prompting, which does not directly address the structural uncertainty inherent in echocardiography video frame. Second, the reliance on complex memory mechanisms incurs significant computational overhead. Notably, the authors of MemSAM themselves acknowledged this limitation and suggested the need for more lightweight solutions. These challenges underscore the importance of a framework that simultaneously addresses both issues by directly modeling and mitigating structural uncer-

*Corresponding author.

tainty within a computationally efficient architecture.

To this end, we propose E^{3}SAM2, a lightweight framework that adapts SAM2 for echocardiography video segmentation by introducing a synergistic entropy-aware and edge-guided method to reduce structural uncertainty and enhance anatomical boundary perception. The name **E**^{3} highlights our three core innovations: **E**ntropy-Guided Attention, **E**ntropy Regularization Loss, and **E**dge-Aware Supervision. To handle uncertainty induced by speckle noise, we introduce an entropy-guided attention mechanism that emphasizes structurally reliable regions, along with a contrastive-inspired entropy regularization loss that maximizes the entropy gap between background and foreground target regions to improve target-background separation. To further improve boundary accuracy, we propose an edge-aware supervision module that injects explicit boundary priors, guiding the model to better delineate blurry anatomical boundaries. These components are efficiently integrated into SAM2 via a global-local feature adapter, enabling efficient adaptation to echocardiography video. Extensive experiments on the CAMUS and EchoNet-Dynamic datasets demonstrate that E^{3}SAM2 achieves state-of-the-art segmentation performance and enables accurate clinical metric estimation, particularly under sparse supervision. Our contributions are summarized as follows:

- We propose E^{3}SAM2, a lightweight SAM2-based framework for echocardiography video segmentation, integrating a global-local adapter for accurate and efficient segmentation.
- We propose a novel entropy-guided attention mechanism and entropy regularization loss to directly mitigate structural uncertainty and improve target-background discrimination.
- We propose an edge-aware supervision module that injects explicit boundary priors to refine blurry anatomical boundaries.
- Extensive experiments on CAMUS and EchoNet-Dynamic demonstrate the effectiveness of E^{3}SAM2 in both segmentation accuracy and clinical metric estimation.

Related Work

Echocardiography Segmentation

Deep learning has significantly advanced echocardiography segmentation, evolving from early CNN-based methods such as U-Net and its variants (Moradi et al. 2019; Leclerc et al. 2019a), to more recent Transformer-based architectures like TransUNet (Chen et al. 2021) and Swin-UNet (Cao et al. 2022). While these models have improved spatial understanding and global context modeling, they often struggle with structural uncertainty caused by severe speckle noise and blurry anatomical boundaries. To address this, temporal models have been introduced to exploit inter-frame consistency (Ahn et al. 2021; Wu et al. 2023; Wei et al. 2020; Wu et al. 2022). However, achieving both high boundary precision and temporal coherence remains challenging, particularly under sparse supervision.

SAM-Based Adaptation in Medical Imaging

The Segment Anything Model (SAM) (Kirillov et al. 2023) and its successor SAM2 (Ravi et al. 2024) have inspired various adaptations for medical imaging. For image segmentation, methods like MedSAM (Ma et al. 2024) fine-tuned SAM on medical datasets, while others enhanced its prompting strategies or boundary awareness (Gowda and Clifton 2024; Xu et al. 2024a; Lin et al. 2023; Ravishankar et al. 2023). For video segmentation, extensions such as MemSAM (Deng et al. 2024), MedSAM-2 (Zhu et al. 2024), and SAM2-UNet (Xiong et al. 2024) have been proposed to address temporal consistency. Notably, MemSAM introduced a memory mechanism tailored for echocardiography. However, most of these adaptations focus on refining the prompting process, which may not fully resolve the structural uncertainty inherent in echocardiography videos. This uncertainty arises from noise and boundary ambiguity, motivating our approach to directly address these fundamental challenges.

Uncertainty and Boundary Modeling

To address the core issues of noise and ambiguity, prior studies have explored uncertainty modeling and boundary-aware designs. Shannon entropy, as a classical measure of uncertainty, has been widely applied in tasks such as active learning (Gal and Ghahramani 2016) and quality assessment (Kendall and Gal 2017). In recent years, in fields such as Natural Language Processing, leveraging entropy to guide attention mechanisms and regularize feature learning has become an active area of research (Lin et al. 2019). These works often aim to manage attention head diversity or prevent phenomena like entropic overload in large models (Jha and Reagen 2025). However, these existing methods are not tailored to the unique challenges of echocardiography video segmentation, specifically the structural uncertainty induced by speckle noise. Similarly, boundary modeling has often employed explicit boundary priors derived from filters such as Sobel (Lin et al. 2025; Bui et al. 2024; Lu et al. 2025; Dong et al. 2022; Xu et al. 2024b), or has relied on implicit architectural designs (Sun et al. 2022; Gao et al. 2024). Furthermore, our approach to boundary enhancement marks a conceptual departure from prior work. Many methods apply boundary priors in the decoder stage or as post-processing. In contrast, our Edge-Aware Supervision (EAS) module operates within the encoder to guide feature extraction from the outset.Crucially, unlike previous work, the core idea of EAS is not simply to inject the output of traditional filters (such as Sobel) into the network as features, but to guide the encoder to learn edge-sensitive feature representations through loss function. Unlike MemSAM (Deng et al. 2024), which enhances temporal consistency through a space-time memory and temporal-aware prompting scheme, our method, E^{3}SAM2, directly addresses the uncertainty caused by noise and blurry boundaries in echocardiography videos.

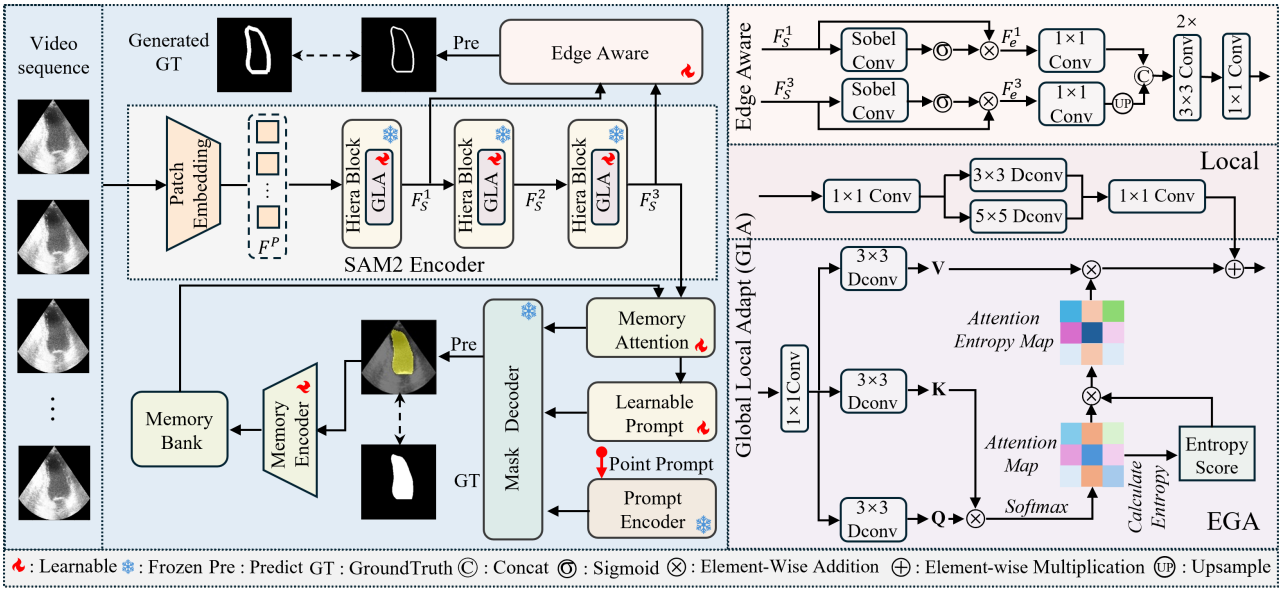

Figure 2: Overview of our E^{3}SAM2, an Entropy-Aware and Edge-Guided Enhancement of SAM2. The input video is encoded by the SAM2 encoder with a Global-Local Adapter and Edge-Aware Supervision, followed by prompt and memory integration, with the mask output from the SAM2 decoder.

Method

Overview of E^{3}SAM2

E^{3}SAM2 is a lightweight framework tailored for echocardiography video segmentation, targeting precise anatomical delineation and temporal coherence, as shown in Figure 2. It builds upon SAM2 by introducing a Global-Local Adapter (GLA) and three key components: Entropy-Guided Attention (EGA), Entropy Regularization Loss (L_{reg}), and Edge-Aware Supervision (EAS), to mitigate speckle noise and blur boundaries. The pipeline comprises six modules. Video encoding leverages a frozen SAM2 encoder and GLA to fuse local texture with global context, producing multi-scale features with reduced uncertainty via EGA and L_{reg}. EAS employs sobel-filtered edge maps as boundary priors, enhancing early boundary sensitivity and guiding attention toward reliable regions. Memory attention fusion integrates current features with historical memory to capture long-range temporal cues. Prompt generation initializes from a sparse point prompt and dynamically updates prompts based on memory. Mask prediction decodes fused features and prompts using the SAM2 decoder for frame-wise segmentation. Memory update compresses predicted masks and selectively stores them using a confidence-aware strategy in MedSAM-2 (Zhu et al. 2024). Through these components, E^{3}SAM2 achieves efficient, accurate, and boundary-aware video segmentation under challenging clinical conditions. The key components are detailed below.

Entropy-Guided Uncertainty Modeling

The cornerstone of our framework is a novel method for managing the structural uncertainty characteristic of echocardiography video. It combines an attention mechanism that implicitly guides focus based on feature-level certainty with a regularization loss that explicitly constrains attention distributions for improved structural consistency.

Entropy-Guided Attention Mechanism. To improve feature aggregation under uncertainty, we propose the Entropy-Guided Attention (EGA) mechanism that reweights spatial attention based on query-level confidence measured by Shannon entropy. Unlike standard attention that treats all regions uniformly, EGA prioritizes structurally reliable queries (low entropy) and suppress uncertain interactions (high entropy). This targeted aggregation improves the model's focus on structurally reliable region. As illustrated in Figure 2, the overall process proceeds as follows.

For each query position i, we compute the Shannon entropy of its attention distribution A_i over key positions j.

$$H(A_i) = -\sum_j A_{i,j} \log A_{i,j}, \tag{1}$$

this entropy value $H(A_i)$ indicates the certainty of the attention focus. A lower value reflects a more concentrated distribution and higher reliability. To transform the entropy into a confidence measure, we normalize it to derive a certainty score:

$$\gamma_i = 1 - \frac{H(A_i)}{\log N}, \tag{2}$$

here, N represents the total number of key positions, and $\gamma_i \in [0, 1]$ denotes the structural certainty at position i, with higher values indicating greater reliability. Using the certainty score, we reweight the attention map to emphasize confident interactions:

$$A_i' = \frac{\gamma_i \cdot A_i}{\sum_j (\gamma_i \cdot A_i) + \varepsilon}, \tag{3}$$

where ε is a small constant added for numerical stability. The reweighted attention is then applied to the value features to produce the final output:

$$\text{Output} = A' \cdot V. \tag{4}$$

Unlike other sparse attention mechanisms (Cheng et al. 2025; Chen et al. 2023) that perform hard token selection, our approach applies a soft, continuous reweighting. This guides the model to focus on low-entropy regions corresponding to reliable anatomical structures while suppressing structural uncertainty areas. This mechanism provides a robust feature foundation for subsequent decoding, and its effectiveness is validated in our experiments (see Table 4).

Entropy Regularization Loss. To further improve target-background discrimination, we introduce an Entropy Regularization Loss (L_{reg}). Our hypothesis is that a well-trained model's attention should be highly focused (low entropy) within the foreground target, but can remain diffuse (high entropy) over the irrelevant background. To encourage this pattern, we propose a contrastive-inspired loss that maximizes the entropy gap between background and foreground target regions. Specifically, L_{reg} is defined as the negative of this entropy gap:

$$\begin{aligned} L_{\text{reg}} &= -\left(H'_b - H'_f\right) \\ &= -\left(\frac{\sum_{k \in M_b} H(A)_k}{|M_b|} - \frac{\sum_{k \in M_f} H(A)_k}{|M_f|}\right) \end{aligned} \tag{5}$$

where k denotes the spatial position, and $|M_f|$ and $|M_b|$ represent the number of pixels in the foreground and background regions,respectively,computed from the ground-truth segmentation masks.

By minimizing L_{reg}, the model is encouraged to increase background entropy while reducing foreground target entropy. This improves target-background discrimination and enhances robustness against complex, noisy backgrounds. This regularization is applied within each GLA, ensuring attention patterns are optimized across all feature levels, with its effectiveness confirmed in our ablation study (see Table 3).

Edge-Aware Supervision

To specifically address the challenge of blurry anatomical boundaries, we introduce an Edge-Aware Supervision (EAS) module, as illustrated in Figure 2. Unlike previous work, the core idea of EAS is not simply to inject the output of traditional filters (such as Sobel) into the network as features, but to guide the encoder to learn edge-sensitive feature representations through loss function. This module leverages hierarchical features, combining fine-grained textures from low-level features F_S^1 with semantic context from high-level features F_S^3 to generate a detailed edge probability map. Given F_S^1 and F_S^3, we first apply Sobel filters along the x and y directions to extract spatial gradients:

$$G_x = Sobel_x(F), \tag{6}$$

$$G_y = Sobel_y(F), \tag{7}$$

where F represents each feature map F_S^1 and F_S^3. The resulting gradients are concatenated and projected into a unified feature space to obtain edge-enhanced features F_e for each scale. To integrate multi-scale edge cues, the high-level edge feature F_e^3 is upsampled by a factor of 4 to match the resolution of F_e^1. The aligned features are concatenated and progressively refined to produce a precise edge probability map that emphasizes boundaries. The EAS module significantly enhances the model's capability to localize anatomical contours in noisy and low-contrast scenarios by injecting high-frequency boundary priors into the encoder's early stages. This improvement is confirmed by our ablation study (see Table 3).

Global-Local Adapter

The aforementioned innovations are combined within the Global-Local Feature Adapter (GLA), a parameter-efficient module. As shown in Figure 2, each GLA comprises two parallel branches: a Local Branch that captures multi-scale texture details through depth-wise convolutions, and a Global Branch that processes features via our Entropy-Guided Attention module. The final output fuses these branches, offering a compact and effective mechanism to integrate local and global information.

Overall Training Objective

The E^{3}SAM2 framework is trained end-to-end by optimizing a composite loss function, L_{total}, which is a weighted sum of three components as shown in Equation (8):

$$L_{\text{total}} = \alpha L_{\text{seg}} + \beta L_{\text{reg}} + \delta L_{\text{edge}} \tag{8}$$

where α, β, and δ are balancing hyper-parameters.

Segmentation Loss (L_{seg}): A standard combination of Dice Loss (Milletari, Navab, and Ahmadi 2016) and Binary Cross-Entropy Loss to ensure both regional accuracy and pixel-wise correctness.

Entropy Regularization Loss (L_{reg}): As defined in Section 3.2, this loss enhances target-background discrimination.

Edge Supervision Loss (L_{edge}): A Binary Cross-Entropy loss between the predicted edge map from the EAS module and a ground-truth edge map generated via a sobel filter.

Experiment

Datasets and Evaluation Metrics

We evaluated our method on two widely used echocardiography video segmentation benchmarks: CAMUS (Leclerc et al. 2019b) and EchoNet-Dynamic (Ouyang et al. 2019). The CAMUS dataset contains 500 clinical cases, each including 2D apical two-chamber and four-chamber view videos. It provides annotations across all frames. In contrast, the EchoNet-Dynamic dataset consists of 10,030 2D two-chamber view videos, but only label end-diastolic (ED) and end-systolic (ES) phases.

To comprehensively evaluate model performance under different supervision regimes, we constructed two training

Method	CAMUS-Semi				EchoNet-Dyn.			
	Dice	IoU	HD95	ASSD	Dice	IoU	HD95	ASSD
UNet (Ronneberger et al. 2015)	90.13	82.36	5.77	2.35	91.36	83.27	4.98	3.01
SwinUNet (Cao et al. 2022)	88.84	80.33	6.10	2.60	87.79	80.14	6.61	5.71
H2Former (He et al. 2023)	91.31	84.30	5.27	2.05	90.21	82.46	5.12	3.78
MedSAM (Ma et al. 2024)	85.42	75.14	8.42	3.34	86.47	79.19	7.97	4.88
MSA (Wu et al. 2025)	88.03	78.98	7.53	2.85	87.91	78.34	6.67	4.34
SAMed (Zhang and Liu 2023)	87.45	78.14	9.17	3.10	86.35	78.96	7.12	4.59
SonoSAM (Ravishankar et al. 2023)	89.80	81.79	6.60	2.45	89.61	82.33	6.58	3.80
SAMUS (Lin et al. 2023)	91.11	83.94	5.08	2.07	91.79	84.32	5.35	3.22
MemSAM (Deng et al. 2024)	93.31±3.04	87.61±5.12	3.82±1.80	1.57±0.72	**92.78**±3.38	85.89±5.12	**4.57**±2.34	2.71±0.78
E^{3}SAM2-Tiny (ours)	93.40±2.95	87.75±3.71	3.71±1.53	1.55±0.69	92.36±3.73	86.02±2.62	6.20±1.13	2.57±1.33
E^{3}SAM2-Small (ours)	**93.42**±2.93	**87.79**±3.70	**3.70**±1.54	**1.54**±0.65	92.41±3.75	**86.10**±2.55	6.13±1.55	**2.55**±1.11

Table 1: Segmentation performance of the proposed method and state-of-the-art baselines on the CAMUS-Semi and EchoNet-Dynamic datasets. Dice/IoU: higher is better; HD95/ASSD: lower is better. HD95 and ASSD are measured in millimeters (mm) on CAMUS-Semi and in pixels on EchoNet-Dynamic. Results are reported as mean ± standard deviation.

protocols on CAMUS: CAMUS-Full, which uses dense annotations on all frames, and CAMUS-Semi, which leverages only ED and ES annotations. For both settings, full-frame annotations were used during testing. Each video sequence was uniformly sampled to 10 frames, with ED and ES set as the first and last frames, respectively. All frames were resized to 256×256 pixels. For fair comparison with prior work, for the CAMUS dataset, we divided it into training, validation, and test sets in a ratio of 7:1:2, while we used the original split for the EchoNet-Dynamic dataset.

To assess segmentation quality, we used four standard evaluation metrics: the mean Dice coefficient (**mDice**), mean Intersection over Union (**mIoU**), Hausdorff Distance-95% (**HD95**), and Average Symmetric Surface Distance (**ASSD**). For each metric, we also report the standard deviation to evaluate prediction stability. In addition, we also report three statistical metrics of Left Ventricular Ejection Fraction (LV_{EF}). Following the Simpson's biplane method of disks (SMOD), which integrates apical two-chamber and four-chamber views for more accurate volumetric assessment, we calculated the LV_{EF} based on the predicted end-diastolic (ED) and end-systolic (ES) segmentations. The predicted LV_{EF} values were evaluated against the ground truth using Pearson correlation coefficient (**Corr**), mean bias (**Bias**), and standard deviation (**Std**). It is worth noting that LV_{EF} estimation is highly sensitive to segmentation accuracy at key time frames, and SMOD provides greater clinical reliability compared to single-view approaches.

Implementation Details

Our method, E^{3}SAM2, is built upon the SAM2 framework. We developed two versions, E^{3}SAM2-Tiny and E^{3}SAM2-Small, based on the corresponding SAM2 backbones. We employ a parameter-efficient fine-tuning strategy, freezing the SAM2 backbone and only optimizing our novel modules.We trained for 200 epochs on CAMUS dataset and 100 epochs on EchoNet-Dynamic. Models were trained using the AdamW optimizer (Loshchilov and Hutter 2017) with a base learning rate of 1×10^{-4}, using a batch size of 4 and

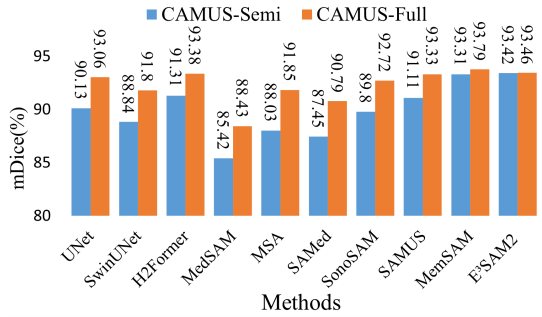

Figure 3: Segmentation performance of the proposed method (E^{3}SAM2-Small) with state-of-the-art methods on the CAMUS-Semi and CAMUS-Full datasets.

resizing all input frames to 256×256. The weights for our composite loss function ($L_{\text{total}} = \alpha L_{\text{seg}} + \beta L_{\text{reg}} + \delta L_{\text{edge}}$) were empirically determined via validation experiments, with $\alpha = 0.8$, $\beta = 0.1$, and $\delta = 0.2$.

Comparison with State-of-the-art Methods

We compare E^{3}SAM2-Tiny and E^{3}SAM2-Small with a diverse set of state-of-the-art segmentation approaches, including traditional architectures such as UNet(Ronneberger, Fischer, and Brox 2015) (CNN-based), SwinUNet(Cao et al. 2022) (Transformer-based), and H2Former(He et al. 2023) (CNN–Transformer hybrid), as well as SAM-adapted models like MedSAM(Ma et al. 2024), MSA(Wu et al. 2025), SAMed(Zhang and Liu 2023), SonoSAM(Ravishankar et al. 2023), SAMUS(Lin et al. 2023), and MemSAM(Deng et al. 2024). Among these, SonoSAM, SAMUS, and MemSAM are tailored for ultrasound, with MemSAM additionally incorporating a temporal memory mechanism.

Quantitative comparison. Our quantitative evaluation demonstrates that E^{3}SAM2 achieves state-of-the-art performance, particularly under challenging, data-limited conditions. As shown in Table 1, on CAMUS-Semi dataset, both

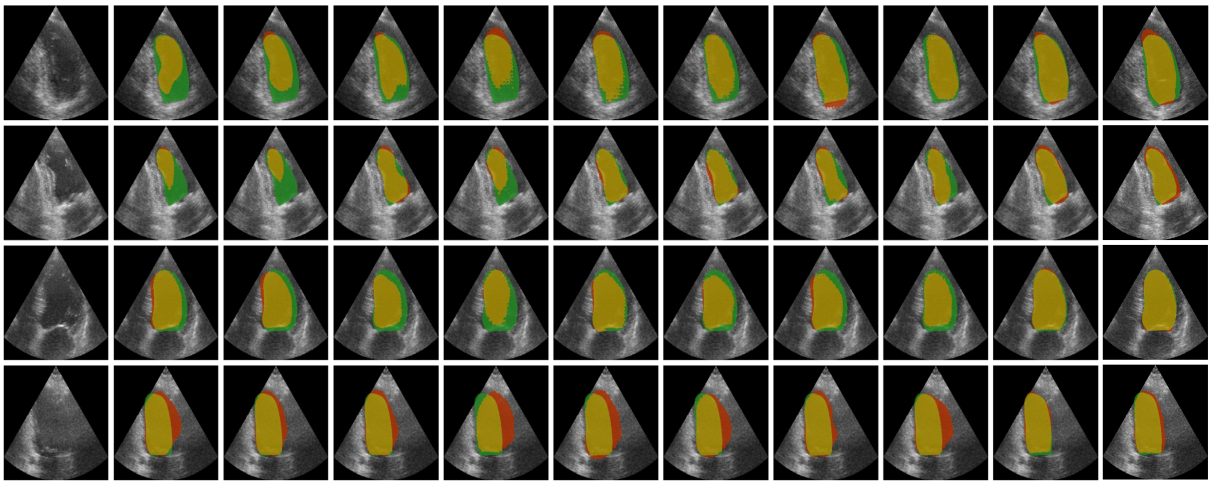

Figure 4: Visual comparison with state-of-the-art methods on the CAMUS-Semi test set. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

Method	corr (%)	bias	std
UNet (Ronneberger et al. 2015)	67.15	11.65	9.39
SwinUNet (Cao et al. 2022)	59.41	6.90	<u>9.06</u>
H2Former (He et al. 2023)	58.61	**0.69**	**7.49**
MedSAM (Ma et al. 2024)	41.63	11.22	11.19
MSA (Wu et al. 2025)	31.00	13.25	14.96
SAMed (Zhang and Liu 2023)	28.22	13.34	12.24
SonoSAM (Ravishankar et al. 2023)	56.18	11.83	9.12
SAMUS (Lin et al. 2023)	67.55	7.02	9.16
MemSAM (Deng et al. 2024)	78.92	<u>4.86</u>	11.10
E^{3}SAM2-Tiny (ours)	<u>79.55</u>	6.70	10.88
E^{3}SAM2-Small (ours)	**81.16**	5.51	11.48

Table 2: Clinical metrics comparison with state-of-the-art methods on the CAMUS-Semi dataset. Corr (%): higher is better; bias and std: lower are better.

E^{3}SAM2-Tiny and E^{3}SAM2-Small outperform all competing methods across all metrics. These results validate the effectiveness of our core design, which integrates entropy-aware (EGA, L_{reg}) with edge guidance (EAS) to address the structural uncertainty inherent in echocardiography videos. On EchoNet-Dynamic dataset, our models remain highly competitive. While MemSAM excels in certain boundary-sensitive metrics, our models achieve superior mIoU and ASSD scores, reflecting more accurate overall segmentation.

To further evaluate our method, we compared it under the same setting on CAMUS-Semi and CAMUS-Full datasets, as shown in Figure 3. While most methods exhibit a notable performance decline with data-limited supervision, E^{3}SAM2 remains remarkably stable, with its mDice score decreasing by only 0.04%. This highlights the model's ability to extract stable anatomical features, enabling deployment in clinical settings with sparse annotations.

The clinical effectiveness of our method is demonstrated through LV_{EF} estimation on the CAMUS-Semi dataset (see Table 2). Existing approaches often struggle under sparse annotations, where degraded segmentation quality can compromise the reliability of clinical assessments. In contrast, E^{3}SAM2 delivers reliable performance under limited annotations, achieving the highest correlation score (81.16%, 79.55%) for LV_{EF} estimation. This highlights that the enhanced segmentation precision of E^{3}SAM2 directly results in more accurate and reliable clinical quantification.

Setting	EGA	EAS	L_{reg}	Dice	IoU	HD95	ASSD
Baseline	×	×	×	93.24	87.49	3.84	1.58
+EGA	✓	×	×	93.27	87.53	3.77	1.57
+EAS	✓	✓	×	93.34	87.65	3.71	1.56
Full model	✓	✓	✓	**93.42**	**87.79**	**3.70**	**1.54**

Table 3: Ablation study of E^{3}SAM2 components on the CAMUS-Semi dataset. Dice/IoU: higher is better; HD95/ASSD: lower is better.

Qualitative comparison. Figure 4 presents a visual comparison of challenging cases from the CAMUS-Semi test set, highlighting the different behaviors under visual scenarios. In heavily speckled images (rows 1–2), most baseline methods produce fragmented masks with large error regions. In contrast, both MemSAM and our E^{3}SAM2 produce more coherent and anatomically plausible masks. This is attributed to E^{3}SAM2's Entropy-Guided Attention (EGA), which suppresses uncertainty caused by noise to preserve key anatomical structures. In blurred-boundary cases (rows 3–4), baseline models frequently struggle to capture complete ventricle contours. By prioritizing structural integrity, E^{3}SAM2 achieves better boundary-aware metrics (HD95, ASSD; Table 1) and the highest LV_{EF} correlation (81.16%; Table 2), confirming its clinical reliability.

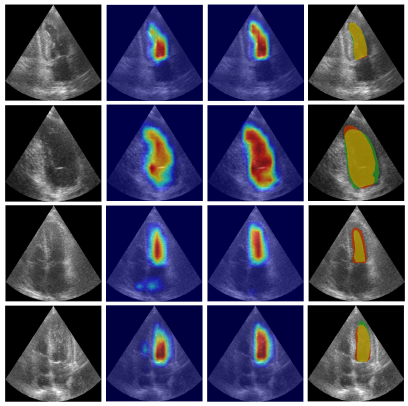

Figure 5: Visualization of attention maps with and without Entropy-Guided Attention(EGA). With entropy guidance, the model's attention is more precisely focused on the ventricular area.

Setting	Dice	IoU	HD95	ASSD
Self-attention	93.24	87.49	3.84	1.58
Top-k Attn. (Chen et al. 2023)	93.33	87.63	3.72	1.56
Mask Attn. (Cheng et al. 2025)	93.28	87.56	3.79	1.57
EGA (ours)	**93.34**	**87.66**	**3.69**	**1.55**

Table 4: Performance comparison of different attention mechanisms on the CAMUS-Semi dataset. Dice/IoU: higher is better; HD95/ASSD: lower is better.

Ablation Studies

To validate the effectiveness of each component, we performed a series of ablation studies on the CAMUS-Semi dataset.

Effectiveness of Each Component. We evaluated the contribution of our core modules: Entropy-Guided Attention (EGA), Entropy Regularization (L_{reg}), and Edge-Aware Supervision (EAS). As shown in Table 3, the full model achieves the best performance, while ablations lead to specific degradations. Disabling EAS notably impairs boundary accuracy (e.g., HD95 increases from 3.70 to 3.77), confirming its role in precise anatomical contours. Removing L_{reg} reduces region overlap performance (mDice decreases from 93.42% to 93.34%), demonstrating its effectiveness in improving target-background discrimination. The best results are achieved when all components are combined, highlighting the synergistic benefits of our integrated design.

Effectiveness of EGA. To evaluate the effectiveness of EGA, we compared it with standard self-attention mechanisms and other sparse approaches. As shown in Table 4, while variants such as Top-k attention outperform the baseline, our EGA achieves the highest performance across all metrics. Figure 5 further illustrates this superiority: unlike the diffuse activations produced by standard attention, EGA generates anatomical contour-consistent attention maps that align closely with ventricular structures. These results provide direct visual and quantitative evidence that the EGA mechanism effectively suppresses irrelevant noise and enhances focus on anatomically relevant regions.

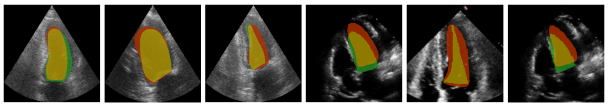

Figure 6: Failure cases on the CAMUS (a-c) and EchoNet-Dynamic (d-f) test sets.

Model	FLOPs	Total params	Trainable params	FPS
MemSAM (Deng et al. 2024)	1770G	151M	58M	41
E^{3}SAM2-Tiny (ours)	**159G**	**42M**	**19M**	**78**
E^{3}SAM2-Small (ours)	190G	49M	**19M**	77

Table 5: Comparison of model complexity and efficiency. Lower FLOPs and parameter counts and higher FPS are better.

Model Efficiency. In addition to accuracy, we also demonstrated the advantages of our proposed method in efficiency by calculating the number of parameters (Params) and floating point operations (FLOPs), and frames per second (FPS). As shown in Table 5, the Small and Tiny variants require only 190G and 159G FLOPs respectively, which are substantially lower than MemSAM's 1770G FLOPs. Both variants contain just 19M trainable parameters, approximately one third of MemSAM's 58M, resulting in reduced training time, lower memory consumption, and improved adaptability in resource-constrained settings. These architectural efficiencies enable real-time inference speeds of 77 to 78 FPS, nearly twice that of MemSAM at 41 FPS. In summary, E^{3}SAM2 offers a highly accurate, lightweight, and efficient solution for practical deployment in clinical echocardiography video analysis.

Conclusion

In this paper, we propose a novel and efficient semi-supervised video segmentation method for echocardiography video segmentation, aiming to effectively extend SAM2 to the echocardiography video domain. Our method achieves state-of-the-art results on both the CAMUS and EchoNet-Dynamic test sets with efficient inference speed.

Despite its strong performance, E^{3}SAM2 still has some limitations. As shown in Figure 6, it may still struggle in cases of extremely poor image quality where anatomical structures are indiscernible even to clinical experts. The model's high boundary sensitivity, a core feature driven by the EAS module, can sometimes lead to over-segmentation in these challenging scenarios. This highlights a key avenue for future work: developing supervision mechanisms that dynamically balance boundary sensitivity with noise robustness.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grants No.62062023.

References

Ahn, S. S.; Ta, K.; Thorn, S.; Langdon, J.; Sinusas, A. J.; and Duncan, J. S. 2021. Multi-frame attention network for left ventricle segmentation in 3D echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 348–357. Springer.

Akkus, Z.; Aly, Y. H.; Attia, I. Z.; Lopez-Jimenez, F.; Arruda-Olson, A. M.; Pellikka, P. A.; Pislaru, S. V.; Kane, G. C.; Friedman, P. A.; and Oh, J. K. 2021. Artificial intelligence (AI)-empowered echocardiography interpretation: a state-of-the-art review. *Journal of clinical medicine*, 10(7): 1391.

Biradar, N.; Dewal, M. L.; and Rohit, M. K. 2015. Speckle noise reduction in B-mode echocardiographic images: A comparison. *IETE Technical Review*, 32(6): 435–453.

Bui, N.-T.; Hoang, D.-H.; Nguyen, Q.-T.; Tran, M.-T.; and Le, N. 2024. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7985–7994.

Cameli, M.; Mondillo, S.; Solari, M.; Righini, F. M.; Andrei, V.; Contaldi, C.; De Marco, E.; Di Mauro, M.; Esposito, R.; Gallina, S.; et al. 2016. Echocardiographic assessment of left ventricular systolic function: from ejection fraction to torsion. *Heart failure reviews*, 21(1): 77–94.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; and Rueckert, D. 2020. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7: 25.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, X.; Li, H.; Li, M.; and Pan, J. 2023. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5896–5905.

Cheng, A.; Yin, C.; Chang, Y.; Ping, H.; Li, S.; Nazarian, S.; and Bogdan, P. 2025. MaskAttn-UNet: A Mask Attention-Driven Framework for Universal Low-Resolution Image Segmentation. *arXiv preprint arXiv:2503.10686*.

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

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gao, S.; Zhang, P.; Yan, T.; and Lu, H. 2024. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9894–9903.

Gowda, S. N.; and Clifton, D. A. 2024. Cc-sam: Sam with cross-feature attention and context for ultrasound image segmentation. In *European Conference on Computer Vision*, 108–124. Springer.

He, A.; Wang, K.; Li, T.; Du, C.; Xia, S.; and Fu, H. 2023. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9): 2763–2775.

Jha, N. K.; and Reagen, B. 2025. Entropy-Guided Attention for Private LLMs. *arXiv preprint arXiv:2501.03489*.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Leclerc, S.; Smistad, E.; Grenier, T.; Lartizien, C.; Ostvik, A.; Cervenansky, F.; Espinosa, F.; Espeland, T.; Berg, E. A. R.; Jodoin, P.-M.; et al. 2019a. RU-Net: A refining segmentation network for 2D echocardiography. In *2019 IEEE International Ultrasonics Symposium (IUS)*, 1160–1163. IEEE.

Leclerc, S.; Smistad, E.; Pedrosa, J.; Østvik, A.; Cervenansky, F.; Espinosa, F.; Espeland, T.; Berg, E. A. R.; Jodoin, P.-M.; Grenier, T.; et al. 2019b. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE transactions on medical imaging*, 38(9): 2198–2210.

Lin, K.-Y.; Hsu, C.-C.; Chen, Y.-N.; and Ku, L.-W. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. *arXiv preprint arXiv:1908.08191*.

Lin, X.; Xiang, Y.; Zhang, L.; Yang, X.; Yan, Z.; and Yu, L. 2023. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. *arXiv preprint arXiv:2309.06824*, 4(11).

Lin, Y.; Zhang, D.; Fang, X.; Chen, Y.; Cheng, K.-T.; and Chen, H. 2025. Rethinking boundary detection in deep learning-based medical image segmentation. *Medical Image Analysis*, 103615.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, W.; Chen, S.-B.; Li, H.-D.; Shu, Q.-L.; Ding, C. H.; Tang, J.; and Luo, B. 2025. Legnet: Lightweight edge-Gaussian driven network for low-quality remote sensing image object detection. *arXiv preprint arXiv:2503.14012*.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Moradi, S.; Oghli, M. G.; Alizadehasl, A.; Shiri, I.; Oveisi, N.; Oveisi, M.; Maleki, M.; and Dhooge, J. 2019. MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67: 58–69.

Ouyang, D.; He, B.; Ghorbani, A.; Lungren, M. P.; Ashley, E. A.; Liang, D. H.; and Zou, J. Y. 2019. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, volume 5.

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

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ravishankar, H.; Patil, R.; Melapudi, V.; and Annangi, P. 2023. Sonosam-segment anything on ultrasound images. In *International workshop on advances in simplifying medical ultrasound*, 23–33. Springer.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Sun, Y.; Wang, S.; Chen, C.; and Xiang, T.-Z. 2022. Boundary-guided camouflaged object detection. *arXiv preprint arXiv:2207.00794*.

Wei, H.; Cao, H.; Cao, Y.; Zhou, Y.; Xue, W.; Ni, D.; and Li, S. 2020. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 623–632. Springer.

Wu, H.; Lin, J.; Xie, W.; and Qin, J. 2023. Super-efficient echocardiography video segmentation via proxy-and kernel-based semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2803–2811.

Wu, H.; Liu, J.; Xiao, F.; Wen, Z.; Cheng, L.; and Qin, J. 2022. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78: 102397.

Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; and Jin, Y. 2025. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102: 103547.

Xiong, X.; Wu, Z.; Tan, S.; Li, W.; Tang, F.; Chen, Y.; Li, S.; Ma, J.; and Li, G. 2024. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *arXiv preprint arXiv:2408.08870*.

Xu, Q.; Li, J.; He, X.; Liu, Z.; Chen, Z.; Duan, W.; Li, C.; He, M. M.; Tesema, F. B.; Cheah, W. P.; et al. 2024a. Espmedsam: Efficient self-prompting sam for universal domain-generalized medical image segmentation. *arXiv preprint arXiv:2407.14153*.

Xu, R.; Xu, C.; Li, Z.; Zheng, T.; Yu, W.; and Yang, C. 2024b. Boundary guidance network for medical image segmentation. *Scientific Reports*, 14(1): 17345.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhu, J.; Hamdi, A.; Qi, Y.; Jin, Y.; and Wu, J. 2024. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*.