Anatomy-aware Spatio-Temporal Modeling for Echocardiography Segmentation

Seok-Hwan Oh*
Barreleye Inc.
Seoul, Korea
shoh@barreleye.co.kr

Guil Jung*
Barreleye Inc.
Seoul, Korea
gijung@barreleye.co.kr

Myeong-Gee Kim
Barreleye Inc.
Seoul, Korea
mgkim@barreleye.co.kr

Young-Min Kim
Barreleye Inc.
Seoul, Korea
ymkim@barreleye.co.kr

Hyeon-Jik Lee
KAIST
Daejeon, Korea
dlguswlr0811@kaist.ac.kr

Sang-Yun Kim
KAIST
Daejeon, Korea
kmjmksy@kaist.ac.kr

Hyuksool Kwon
SNUBH
Seongnam, Korea
jinuking3g@snubh.org

Hyeon-Min Bae†
KAIST
Daejeon, Korea
hmbae@kaist.ac.kr

Abstract

Echocardiography segmentation plays a central role in quantitative cardiovascular assessment. However, automated echocardiography segmentation remains challenging due to speckle noise, poorly defined anatomical boundaries, acquisition variability, and the limited anatomical coverage of existing benchmarks. Conventional segmentation methods demonstrate limited accuracy under such low image quality and domain shifts. Inspired by the observation that expert cardiologists can maintain reliable anatomical delineation by leveraging their structural understanding of the cardiac anatomy, we propose that incorporating anatomical priors into echocardiography segmentation models can improve both precision and robustness. To this end, we propose AST-Seg, an anatomy-aware spatio-temporal framework for echocardiography segmentation. AST-Seg introduces an anatomy-aware feature encoder that regularizes transformer attention with anatomy-derived spatial-relation priors, and a Deformable Spatio-Temporal Mamba module that captures localized and periodic cardiac motion through adaptive spatial aggregation followed by temporal state-space modeling. In addition, we curate an Echocardiography Anatomy (EA) dataset with pixel-level annotations for 12 clinically relevant cardiac structures, enabling comprehensive multi-structure cardiac interpretation. Extensive experiments on in-distribution CAMUS and EA datasets, as well as an out-of-distribution point-of-care echocardiography dataset, demonstrate that AST-Seg provides domain-generalized segmentation with improved precision.

1 Introduction

Medical ultrasound is a widely used imaging modality in clinical practice owing to its non-ionizing nature, low cost, and real-time imaging capability. Among its various applications, echocardiography

*Equal contribution
†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

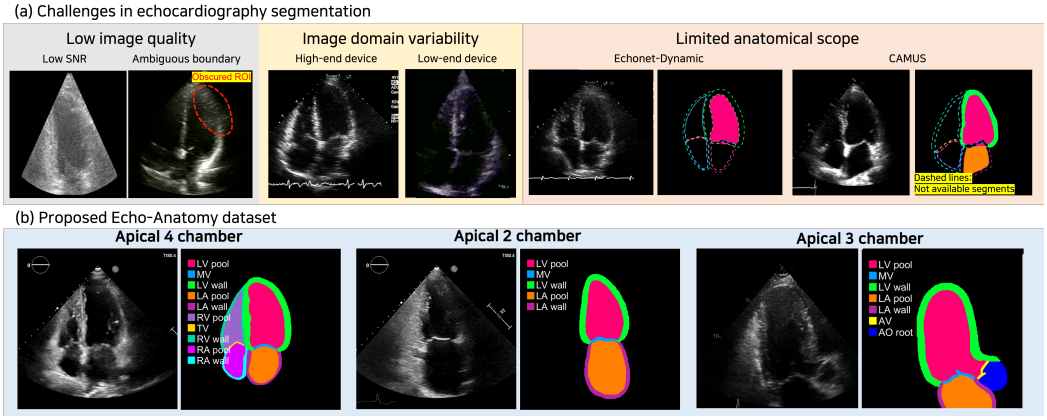

Figure 1: (a) Challenges in Echo segmentation, (b) Proposed EA dataset

(Echo) is essential for cardiac assessment, as it provides dynamic visualization of cardiac structures and function, playing a central role in diagnosing a broad spectrum of cardiovascular diseases [1, 2].

Precise analysis of cardiac morphology is critical for clinical decision-making. For example, the left ventricle (LV) blood pool (BP) volume is essential for assessing heart failure, and the morphology of the mitral valve, right ventricle (RV), and aorta is crucial for diagnosing valvular and structural heart diseases. These examples indicate that clinically reliable Echo interpretation is achieved through a structured understanding of multiple cardiac anatomies [3, 4, 5, 6].

Despite its broad clinical utility, Echo analysis remains labor-intensive and highly dependent on the operator's expertise. Clinically significant measurements rely on manual or semi-automated delineation of cardiac structures across multiple frames, which is time-consuming and susceptible to inter- and intra-observer variability [7, 8]. Consequently, there is a growing need for computational methods that can automate and standardize Echo assessment, thereby improving diagnostic efficiency and reproducibility.

Recent advances in deep learning have substantially improved automated medical image analysis, including cardiac segmentation [9, 10, 11]. However, Echo segmentation remains challenging due to several factors. First, ultrasound (US) images often exhibit a low signal-to-noise ratio and ambiguous anatomical boundaries, making reliable segmentation challenging [12, 13]. Second, real-world Echo data are highly heterogeneous across devices, acquisition protocols, and operators [14, 15]. Such variability leads to substantial domain shifts, whereas existing datasets are typically collected from a limited range of scanners and operators. As a result, models trained on such datasets suffer from limited generalization to unseen clinical environments. Third, existing Echo segmentation datasets remain anatomically limited. Widely used benchmarks such as EchoNet-Dynamic primarily focus on the LV BP, and CAMUS provides annotations for only a small subset of cardiac chambers (Figure 1-(a)) [16, 17]. However, comprehensive cardiac assessment requires analysis of a broader set of anatomies, including the RV, atria, valves, and aorta structures.

To address the limitations, we propose AST-Seg, an anatomy-aware spatio-temporal Echo segmentation framework. Our motivation is inspired by how experienced cardiologists interpret Echo. Expert cardiologists identify structures by integrating anatomical priors with temporal motion patterns across the cardiac cycle, enabling reliable interpretation even under severe image degradation, such as high speckle noise and poorly defined anatomical boundaries. The observation motivates the development of a neural network that explicitly encodes anatomical semantics and exploits spatio-temporal dynamics to achieve robust and precise cardiac segmentation.

Specifically, AST-Seg learns anatomy-aware representations by incorporating anatomical priors into feature extraction, enabling structure-aware feature representation beyond appearance cues. In addition, AST-Seg proposes a Deformable Spatio-Temporal Mamba module to effectively capture the Echo temporal context. The resulting anatomy-aware and spatio-temporal representation enables robust segmentation under challenging conditions, including noise, ambiguous boundaries, and domain shifts. Furthermore, we introduce a new large-scale Echo dataset with annotations for 12 clinically significant cardiac structures across diverse acquisition settings, substantially extending the

anatomical scope of existing benchmarks. Through extensive experiments on both in-distribution and out-of-distribution datasets, we show that anatomy-aware spatio-temporal modeling improves segmentation accuracy and enhances robustness under domain shift.

The main contributions of the paper are summarized as follows:

- We propose anatomy-aware representation learning for Echo segmentation, which regularizes transformer attention with anatomy-derived spatial-relation priors to learn structurally grounded cardiac representations beyond local appearance cues.
- We introduce a Deformable Spatio-Temporal Mamba module that effectively interprets localized and periodic cardiac motion through adaptive spatial aggregation followed by temporal state-space modeling.
- We curate an Echo anatomy dataset with pixel-level annotations for 12 clinically relevant cardiac structures, enabling comprehensive anatomy interpretation beyond existing benchmarks.

2 Related Work

2.1 Echocardiography segmentation

Automated Echo segmentation has been widely investigated to reduce the manual delineation burden and improve reproducibility in cardiac assessment [18]. Learning-based Echo segmentation has predominantly focused on a limited anatomical target, most commonly the LV BP or a small subset of cardiac chambers [17, 18, 19]. Public benchmarks such as EchoNet-Dynamic and CAMUS have played an important role in advancing automated Echo segmentation, however, anatomical coverage of these datasets remains restricted. As a result, existing schemes are primarily designed to support selected cardiac measurements, rather than comprehensive anatomical understanding [3, 20]. In clinical practice, however, Echo interpretation requires joint interpretation of multiple structures, including RV chambers, atria, valves, and aortic structures. Motivated by the limitation, we propose a multi-structure Echo segmentation framework with broader anatomical coverage.

2.2 Foundation models for medical image segmentation

Foundation models have recently emerged as a prominent approach for visual representation learning. Promptable segmentation models such as SAM [21] have demonstrated strong generalization across natural image domains, motivating a growing body of work on adapting SAM-like networks to medical image segmentation. Representative methods such as MedSAM [22] and Medical SAM Adapter (MSA) [23] demonstrate that large-scale segmentation priors can be transferred and contribute to medical imaging. SAMUS [24] extends this line of research to medical US by optimizing SAM for US image segmentation. Despite these advances, existing SAM-based medical segmentation methods primarily focus on transferring pixel-level segmentation capability, with limited explicit modeling of anatomical context, which is crucial in medical imaging. In Echo, however, reliable segmentation depends not only on local image appearance but also on the structured understanding of cardiac anatomy. Motivated by the limitation, we propose an anatomy-aware representation learning framework designed for Echo video segmentation.

2.3 Spatio-temporal modeling for echocardiography video segmentation

Temporal modeling is essential for video-based medical image analysis, where clinically relevant structures are often better characterized through temporal context than by a single frame appearance. Early approaches employed recurrent neural networks [25] or LSTMs [26, 27] to capture temporal dependencies. However, such models demonstrate limitations in modeling long-range interactions and complex motion patterns. More recently, Transformers [28] have achieved strong performance in video understanding, however their quadratic complexity with respect to token length makes them computationally expensive for high-resolution medical videos. State space models, including Mamba [29], offer a scalable alternative for long-sequence modeling and have recently attracted attention for efficient video representation learning. Nevertheless, flattening all video patches into a single token sequence is sub-optimal for Echo, where motion is periodic, spatially localized, and anatomically constrained. Motivated by this observation, we develop a Deformable Spatio-Temporal

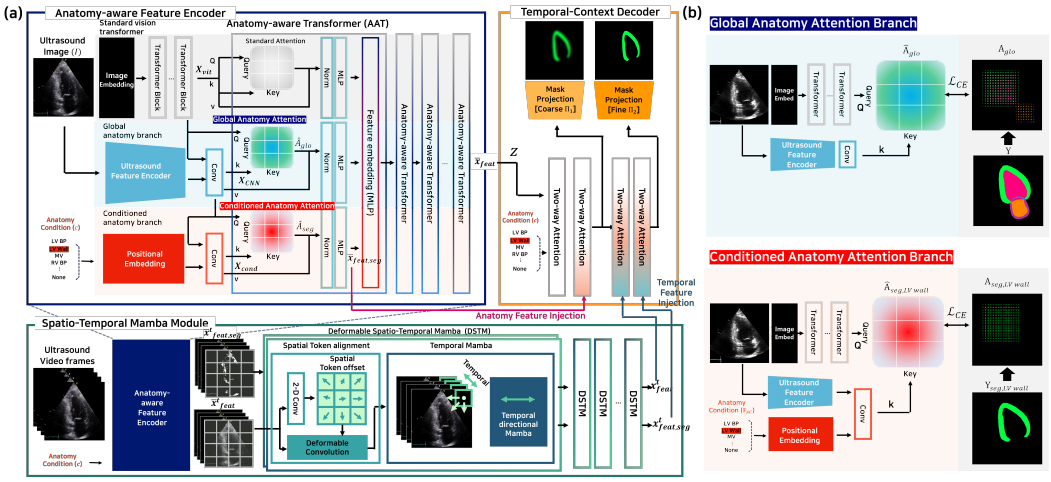

Figure 2: Overview of the (a) AST-Seg configuration and (b) anatomy representation learning.

Mamba module that integrates adaptive local anatomical aggregation with efficient state-space temporal modeling.

3 Method

We propose AST-Seg, an anatomy-aware spatio-temporal framework for Echo video segmentation. The method is designed to address two key challenges of Echo analysis: anatomically structured representation learning and effective temporal modeling of cardiac motion. To this end, AST-Seg consists of three components: (i) an Anatomy-aware Echo Feature Encoder (AEFE) that regularizes transformer attention with anatomical priors, (ii) a Deformable Spatio-Temporal Mamba (DSTM) module that captures cardiac motion effectively, and (iii) a Temporal-Context Decoder (TCD) that combines current-frame evidence with temporally refined anatomical features to produce precise segmentation. Section 3.1 introduces the Echo anatomy dataset that provides the anatomical supervision underlying this study, and Section 3.2 details the AST-Seg configuration.

3.1 Echocardiography anatomy dataset

We introduce a newly curated Echo Anatomy (EA) dataset with pixel-level annotations for 12 clinically relevant cardiac structures across three apical views. The dataset contains 21,535 annotated frames from 3,110 subjects. The dataset is split at the case level, with 19,125 frames used for training and 2,410 for testing. EA annotations were produced by three experienced sonographers using a standardized protocol and independently reviewed by an expert physician. As illustrated in Figure 1, the EA dataset substantially expands the anatomical coverage of existing Echo segmentation benchmarks, which typically focus on a limited subset of cardiac structures, thereby enabling comprehensive cardiac anatomy interpretation. To improve morphological diversity while reducing redundant annotations, one frame per Echo video is sampled for annotation, while the corresponding video provides temporal context. The Echo videos were acquired from three ultrasound vendors (GE Healthcare, Philips Healthcare, and Siemens) and five scanner models. Additional details are provided in Appendix A.

3.2 Anatomy-aware spatio-temporal echocardiography segmentation

We formulate Echo segmentation as an anatomy-conditioned video segmentation problem. For the EA dataset, each input clip $\mathcal{I} = \{I^t\}_{t=1}^{T}$ consists of $T = 10$ chronologically ordered frames uniformly sampled from a cardiac cycle. A single frame, indexed by τ, is annotated for the anatomies available in the corresponding view.

Figure 2 presents an overview of AST-Seg, which consists of an AEFE for anatomically structured representation learning, a DSTM module that captures temporal context through efficient local-to-

temporal aggregation, and a TCD that combines frame-level anatomical features with temporally refined features to generate the segmentation masks.

3.2.1 Anatomy-aware echocardiography feature encoder

Generic self-supervised image encoders mainly learn appearance-driven correlations, which are often insufficient for Echo due to speckle noise and ambiguous boundaries. Reliable cardiac segmentation requires representations that recognize the anatomical identity and spatial organization of the Echo. We therefore extend a standard vision transformer (ViT) [30] with two anatomy-aware attention branches: a global branch for whole cardiac anatomy organization and a conditioned branch for the target anatomy.

Given an Echo frame $I^t \in \mathbb{R}^{H\times W\times 3}$ and a one-hot anatomical condition vector $c \in \{0,1\}^K$, where $K = 12$, the encoder produces

$$(\bar{x}^t_{\text{feat}},\, \bar{x}^t_{\text{feat,seg}},\, \hat{A}^t_{\text{glo}},\, \hat{A}^t_{\text{seg}}) = \mathcal{E}_\theta(I^t, c). \tag{1}$$

Here, $\hat{A}^t_{\text{glo}}$ and $\hat{A}^t_{\text{seg}}$ are the generated global and conditioned anatomy attention maps. The feature $\bar{x}^t_{\text{feat,seg}}$ is the output of the conditioned anatomy branch, whereas $\bar{x}^t_{\text{feat}}$ denotes the output of the AEFE, which is a unified representation derived from the appearance-driven ViT representation, global anatomical context, and condition-specific anatomical feature. For clarity, frame indices are omitted in the following per-frame attention formulation.

Let $X_{\text{vit}} \in \mathbb{R}^{N\times C}$ denote the standard ViT token representation. $X_{\text{cnn}} \in \mathbb{R}^{N\times C}$ is a US image feature embedding extracted by a lightweight convolutional network (CNN). The global anatomy branch performs cross-attention employing X_{vit} as the query and X_{cnn} as the key and value:

$$Q_{\text{glo}} = X_{\text{vit}}W^Q_{\text{glo}}, \quad K_{\text{glo}} = X_{\text{cnn}}W^K_{\text{glo}}, \quad V_{\text{glo}} = X_{\text{cnn}}W^V_{\text{glo}}. \tag{2}$$

The resulting global anatomy attention and feature are computed as

$$\hat{A}_{\text{glo}} = \text{Softmax}\left(\frac{Q_{\text{glo}}K^\top_{\text{glo}}}{\sqrt{d}}\right), \qquad X_{\text{glo}} = \hat{A}_{\text{glo}}V_{\text{glo}}. \tag{3}$$

The proposed branch enables appearance-driven ViT tokens to incorporate US structural features guided by the global organization of cardiac anatomy. For the conditioned anatomy branch, the condition vector c is incorporated into $X_{\text{cond}} = \phi_{\text{cond}}(X_{\text{cnn}}, c)$. The branch then performs cross-attention using X_{vit} as the query and X_{cond} as the key and value.

$$Q_{\text{seg}} = X_{\text{vit}}W^Q_{\text{seg}}, \quad K_{\text{seg}} = X_{\text{cond}}W^K_{\text{seg}}, \quad V_{\text{seg}} = X_{\text{cond}}W^V_{\text{seg}}. \tag{4}$$

The conditioned attention and feature are obtained as

$$\hat{A}_{\text{seg}} = \text{Softmax}\left(\frac{Q_{\text{seg}}K^\top_{\text{seg}}}{\sqrt{d}}\right), \qquad \bar{x}_{\text{feat,seg}} = \hat{A}_{\text{seg}}V_{\text{seg}}. \tag{5}$$

Finally, the integrated feature representation is obtained by MLP projection:

$$\bar{x}_{\text{feat}} = \phi_{\text{fuse}}\left([X_{\text{vit}}, X_{\text{glo}}, \bar{x}_{\text{feat,seg}}]\right). \tag{6}$$

The $\bar{x}_{\text{feat}}$ representation provides the ViT-derived appearance feature while incorporating both global cardiac organization and target-specific anatomical context.

Anatomy representation learning. To regularize the learned attention with anatomically meaningful spatial relations, we construct anatomy-derived pairwise priors from segmentation masks. Let $M_q \in \{0,1\}^{h_p\times w_p}$ be the downsampled mask of anatomy q. We define the affinity between patch tokens i and j as

$$A_q(i,j) = \mathbf{1}[M_q(i) = 1]\mathbf{1}[M_q(j) = 1]. \tag{7}$$

The affinity encodes intra-structure token relationships by assigning a higher relational probability to tokens belonging to the same anatomical region. The global and conditioned priors are defined as

$$A_{\text{glo}} = \sum_{q=1}^{K} A_q, \qquad A_{\text{seg}} = \sum_{q=1}^{K} c_q A_q. \tag{8}$$

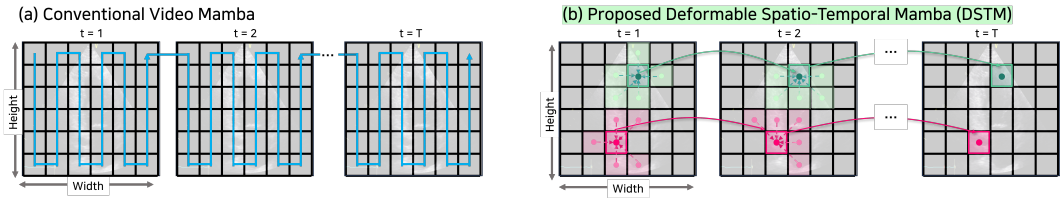

Figure 3: Illustration of (a) Conventional Video Mamba and (b) proposed DSTM

A_{glo} represents the overall relational structure across all cardiac anatomies, whereas A_{seg} encodes the relation prior associated with the queried target anatomy.

The target attention maps are used to supervise the attention distributions:

$$\mathcal{L}_{\text{attn}} = \sum_{b \in \{\text{glo,seg}\}} \mathcal{H}(\bar{A}_b, \hat{A}_b), \tag{9}$$

where $\bar{A}_b$ denotes the target attention map obtained by row-wise normalization of the nonzero rows of A_b. Rows containing only zeros are excluded from the anatomy-aware attention supervision. $\mathcal{H}$ denotes smoothed cross-entropy over the attention distribution. Proposed token-relation supervision encourages the encoder to capture anatomical relationships beyond local appearance similarity, improving robustness to Echo appearance variations.

3.2.2 Deformable Spatio-Temporal Mamba

Temporal modeling is particularly important in Echo, where the cardiac anatomies are more reliably characterized using the temporal context. Standard approaches employ temporal self-attention over all video tokens. However, the scheme incurs quadratic computational complexity, scaling as $\mathcal{O}((TN)^2)$ for T frames and $N = hw$ spatial tokens per frame. Although Mamba-based video models alleviate the problem by scanning a flattened 1D sequence of video tokens, such global serialization remains suboptimal for Echo, where motion is predominantly local, periodic, and anatomically constrained [31].

To better capture the Echo temporal characteristics, we propose a Deformable Spatio-Temporal Mamba (DSTM) module, which factorizes spatio-temporal modeling into two stages: adaptive local spatial aggregation and efficient temporal sequence modeling. Instead of serializing all spatio-temporal patches into a single sequence, DSTM first performs deformable aggregation of local anatomical features and then applies selective state-space modeling to capture temporal dynamics (Figure 3).

Let $\{\bar{x}^t_{\text{feat}}\}_{t=1}^T$ and $\{\bar{x}^t_{\text{feat,seg}}\}_{t=1}^T$ denote the frame-level encoder features before temporal modeling. DSTM is applied to each feature stream to obtain temporally refined memories:

$$\{x^t_{\text{feat}}\}_{t=1}^T = \text{DSTM}\left(\{\bar{x}^t_{\text{feat}}\}_{t=1}^T\right), \qquad \{x^t_{\text{feat,seg}}\}_{t=1}^T = \text{DSTM}\left(\{\bar{x}^t_{\text{feat,seg}}\}_{t=1}^T\right). \tag{10}$$

Here, $\bar{x}^t_{\text{feat}}$ and $\bar{x}^t_{\text{feat,seg}}$ provide frame-centric anatomical features, while x^t_{feat} and $x^t_{\text{feat,seg}}$ serve as temporal features that encode motion-aware temporal context for the decoder.

For a generic input stream $Z = \{z_t\}_{t=1}^T$, where $z_t \in \mathbb{R}^{C \times h \times w}$, DSTM first performs deformable local aggregation. For each frame t and spatial location p, we compute

$$\tilde{z}_t(p) = \sum_{n=1}^{K_d} m_{t,p,n}\, W_n\, z_t(p + p_n + \Delta p_{t,p,n}), \tag{11}$$

where K_d is the number of deformable sampling points, p_n are canonical local offsets, $\Delta p_{t,p,n}$ denote learned offsets, $m_{t,p,n}$ are modulation weights, and W_n are learnable projection matrices. The operation allows each token to adaptively aggregate spatial anatomical features instead of relying on a fixed spatial neighborhood, which is particularly suitable for Echo where structures undergo continuous deformation and motion throughout the cardiac cycle.

For each spatial location p, the aggregated features $\{\tilde{z}_t(p)\}_{t=1}^T$ form a temporal sequence, which is processed by a Mamba-based selective state-space model. In compact form, the recurrent update is

written as

$$h_{t,p} = A_t h_{t-1,p} + B_t \tilde{z}_t(p), \qquad y_t(p) = C_t h_{t,p}, \tag{12}$$

where $h_{t,p}$ denotes the latent temporal state, and A_t, B_t, and C_t are input-dependent state-space parameters conditioned on $\tilde{z}_t(p)$.

The final spatio-temporal feature is obtained through gated residual fusion,

$$z_t^\star(p) = \tilde{z}_t(p) + \gamma\, y_t(p), \tag{13}$$

where γ is a learnable gate. Unlike generic Video Mamba variants that flatten spatial and temporal tokens into a single scan order, DSTM adaptively captures the location-specific spatial context of each frame and performs selective state-space modeling along the temporal trajectory of each anatomy-informed location. The proposed DSTM dynamically aggregates local anatomical evidence and captures its temporal motion, which is better aligned with the localized and cyclic deformation patterns of Echo. Moreover, the computation scales linearly with the temporal length at each spatial location, providing a substantially more efficient alternative to spatio-temporal self-attention.

3.2.3 Temporal-Context Decoder

To generate the segmentation mask, we propose a Temporal-Context Decoder that performs coarse-to-fine decoding by integrating current-frame evidence with temporally refined anatomical context. For frame t, the decoder takes as input the frame-centric features $(\bar{x}^t_{\text{feat}}, \bar{x}^t_{\text{feat,seg}})$, the temporally refined features $(x^t_{\text{feat}}, x^t_{\text{feat,seg}})$, and the anatomy condition c.

In TCD, the decoder first forms a coarse mask representation using the current-frame features,

$$S_t^{(1)} = \mathcal{T}_{\text{cur}}(\bar{x}^t_{\text{feat}}, \bar{x}^t_{\text{feat,seg}}, c), \tag{14}$$

and then refines the segmentation mask using temporal context features through the transformer blocks,

$$S_t^{(2)} = \mathcal{T}_{\text{tmp}}(S_t^{(1)}, x^t_{\text{feat}}, x^t_{\text{feat,seg}}). \tag{15}$$

The intermediate and final mask predictions are obtained as

$$\hat{Y}^{(k)}_{t,c} = \Pi_k(S_t^{(k)}), \qquad k \in \{1, 2\}, \tag{16}$$

where Π_1 and Π_2 denote the coarse and fine mask projection heads, respectively.

The design explicitly employs frame-faithful mask formation and temporal-context refinement: the first stage anchors the prediction to the instantaneous appearance of the current frame, whereas the second stage improves anatomy segmentation by incorporating the temporal context of the Echo.

3.2.4 Implementation details

The encoder is pretrained using the masked image modeling (MIM) scheme [32] on the curated 7.9M frame annotation-free Echo Video dataset (EV dataset, described in Appendix A). After the initial MIM pretraining, we further adapt the encoder through anatomy-aware representation learning. Specifically, self-supervised MIM learning on the annotation-free EV dataset and anatomy-prior-guided optimization of the AEFE branches on the EA dataset are interleaved, enabling the encoder to retain generic Echo visual representations while progressively incorporating explicit anatomical priors into the encoder.

AST-Seg is optimized for segmentation using cross-entropy loss and Dice loss, following the MedSAM [22] training scheme. Since the decoder produces both an intermediate mask and a final refined mask, the segmentation loss is applied to both predictions. Specifically,

$$\mathcal{L}_{\text{seg}} = \sum_{k\in\{1,2\}} \lambda^{(k)} \left[\mathcal{L}_{\text{CE}}\left(\hat{Y}^{(k)}_{\tau,c}, Y_{\tau,c}\right) + \mathcal{L}_{\text{Dice}}\left(\hat{Y}^{(k)}_{\tau,c}, Y_{\tau,c}\right) \right], \tag{17}$$

where the coefficients $\lambda^{(k)}$ weight segmentation losses and $Y_{\tau,c}$ denotes the ground-truth mask.

Table 1: Quantitative comparison on the ID CAMUS and the OOD PoC-Echo datasets. AST-Seg results are reported as mean ± standard deviation.

Category	Method	CAMUS (ID)									PoC-Echo (OOD)		
		LV BP			LV wall			LA BP			LV BP		
		Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓
Conventional NN	U-Net	92.22	85.75	7.53	83.87	72.59	9.09	85.99	76.59	10.78	80.90	71.28	25.66
	SwinUNet	92.61	86.40	6.74	86.45	76.38	7.29	88.51	80.23	9.10	82.91	73.24	27.31
	H2Former	92.91	86.98	6.61	86.15	75.93	6.82	88.02	79.41	8.95	83.10	72.92	21.37
SAM-based	MedSAM	88.87	82.66	8.45	81.30	71.59	8.32	88.28	80.59	10.38	85.46	78.01	16.62
	MSA	91.12	84.54	8.17	84.55	73.62	7.39	87.91	79.19	8.73	86.19	77.57	21.27
	SAMed	91.79	84.98	7.76	84.77	73.91	8.13	88.06	79.61	10.23	86.37	79.22	19.83
	SAMUS	91.22	85.12	6.72	84.39	75.02	6.85	88.64	80.35	8.22	86.14	79.02	13.50
	MemSAM	93.24	87.53	4.79	87.74	78.41	7.19	88.08	79.22	9.50	87.48	80.05	12.83
	MediViSTA	90.94	83.65	7.73	83.19	71.66	7.71	87.10	78.16	9.39	83.65	73.82	23.91
Proposed	AST-Seg	**95.03** ±1.70	**90.57** ±2.74	**4.75** ±4.15	**89.00** ±2.80	**80.29** ±4.41	**5.89** ±3.12	**89.52** ±5.77	**81.53** ±8.42	**6.86** ±4.50	**90.83** ±5.45	**83.59** ±8.47	**11.21** ±8.56

Table 2: Quantitative assessments on the EA dataset.

Method	Blood pool								Wall								Valve						Aorta		Average	
	LV BP		LA BP		RV BP		RA BP		LV wall		LA wall		RV wall		RA wall		MV		TV		AV		AO root		Mean	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net	90.25	83.18	80.09	69.82	84.48	74.88	81.23	70.66	69.97	54.17	61.61	46.45	58.52	44.12	55.58	41.01	55.57	40.33	48.75	35.81	53.25	40.87	73.95	62.12	67.77	55.29
SwinUNet	90.95	83.79	79.63	68.69	84.47	74.67	81.26	70.69	68.31	52.71	61.88	46.65	57.79	43.87	56.12	42.34	56.13	43.01	48.41	35.75	56.12	42.05	76.22	64.45	68.11	55.72
H2Former	91.11	84.93	81.55	70.49	84.91	75.45	81.81	70.71	68.46	52.91	59.94	44.27	59.15	44.68	55.65	41.16	57.22	43.36	49.22	36.51	57.37	44.01	73.76	61.04	68.35	55.79
MedSAM	90.07	82.08	86.11	76.04	78.87	68.26	82.65	71.05	71.82	58.13	62.41	47.13	58.99	44.65	57.25	42.10	60.96	45.96	46.09	33.28	57.50	44.44	76.88	63.08	69.13	56.35
MSA	88.90	80.38	79.46	69.33	79.52	68.32	80.06	68.95	68.81	54.77	60.09	44.15	59.47	44.91	55.31	41.15	55.45	42.19	46.91	33.59	53.25	40.03	76.68	64.11	66.99	54.32
SAMed	90.84	83.51	83.87	73.72	78.90	66.64	79.21	67.15	71.07	56.82	59.41	43.98	59.19	45.01	53.89	40.40	58.84	45.21	48.81	35.98	54.71	41.17	77.44	66.02	68.02	55.47
SAMUS	89.47	81.98	82.03	72.56	75.28	63.46	79.59	67.83	69.89	56.19	63.01	47.33	60.01	45.13	57.29	41.70	60.34	45.20	49.03	36.13	55.54	40.71	76.38	64.02	68.16	55.19
MemSAM	91.34	85.02	86.28	76.90	84.12	73.85	86.06	76.66	73.81	58.85	63.53	48.55	61.08	45.81	59.43	43.82	60.18	45.17	**52.75**	38.86	55.12	42.21	77.98	65.52	70.97	58.44
MediViSTA	89.98	82.21	82.80	72.41	82.67	72.15	84.11	73.48	70.78	56.24	63.15	48.31	**61.76**	46.45	57.52	41.75	59.69	44.81	51.56	37.55	54.35	40.52	76.52	65.32	69.57	56.77
Proposed	**92.26**	**85.78**	**87.94**	**79.32**	**86.14**	**75.79**	**87.82**	**78.81**	**75.10**	**60.78**	**66.18**	**51.48**	61.10	**46.73**	**60.49**	**45.58**	**62.77**	**47.51**	52.00	**39.07**	**59.37**	**45.01**	**80.87**	**69.49**	**72.67**	**60.45**

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate AST-Seg under both in-distribution (ID) and out-of-distribution (OOD) environments. For ID evaluation, the public CAMUS benchmark and the curated EA dataset are employed. To assess robustness under domain shift, we additionally curate a point-of-care (PoC) Echo dataset that consists of 259 echocardiograms acquired using portable US devices. The PoC-Echo dataset is used only for evaluation. The dataset demonstrates substantial distribution shifts from conventional Echo, including lower image quality, reduced signal-to-noise ratio, and ambiguous boundaries. The Echo frames are resized to 256×256.

Evaluation metrics and baselines. The Dice similarity coefficient (Dice), mean Intersection-over-Union (mIoU), and the 95th-percentile Hausdorff distance (HD95) are employed as the quantitative metrics. AST-Seg is compared with representative conventional segmentation networks, including U-Net [33], SwinUNet [34], and H2Former [35], as well as SAM-based and video segmentation baselines, including MedSAM [22], MSA [23], SAMed [36], SAMUS [24], MemSAM [12], and MediViSTA [37]. The networks are trained on the CAMUS and EA datasets. The video segmentation models, including MemSAM, MediViSTA and AST-Seg, employ 10 frames of input clip.

4.2 Comparative study

Tables 1 and 2 present quantitative assessments of AST-Seg with conventional, SAM-based, and video segmentation baselines on in-distribution CAMUS and EA benchmarks and the out-of-distribution PoC-Echo dataset. On CAMUS, AST-Seg consistently achieves enhanced segmentation of the LV BP, LV wall, and left atrial (LA) BP, validating the effectiveness of anatomy-aware spatio-temporal modeling under standard Echo acquisition settings. Under portable US domain shift, conventional baselines degrade substantially and SAM-based methods show limited generalization, whereas AST-Seg shows robust performance on PoC-Echo. Figure 4-(a) shows that AST-Seg provides precise segmentation even in cases with ambiguous anatomical boundaries, supporting its robustness under challenging Echo imaging conditions. The results demonstrate that incorporating anatomical representation learning and the DSTM contributes to domain generalization in Echo. On the EA

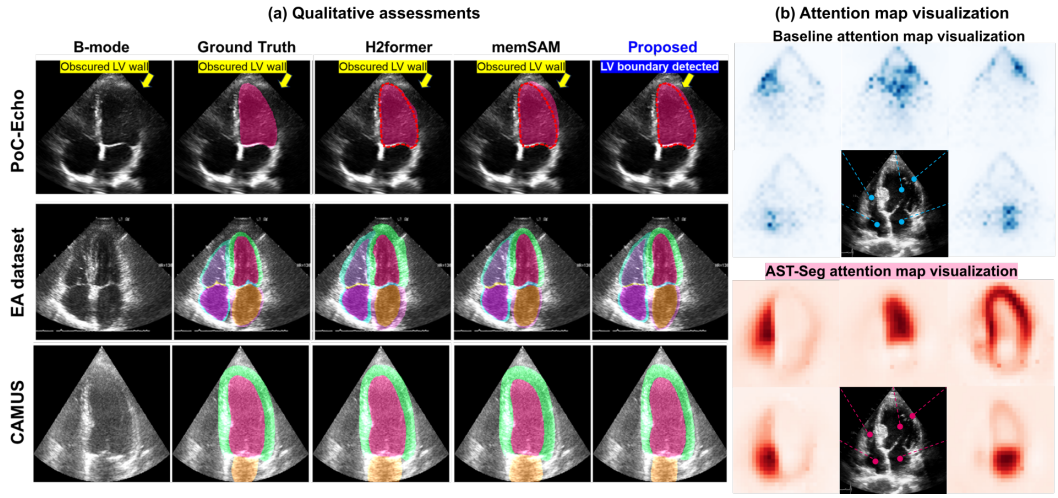

Figure 4: Qualitative results and attention map visualization. In (b), marked points denote query locations, and each attention map shows the corresponding attention weights over the spatial positions.

dataset, AST-Seg achieves the highest average Dice and mIoU, including challenging atrium and aortic structures, demonstrating its ability to provide comprehensive Echo anatomical understanding.

4.3 Ablation Study

Table 3 shows the ablation study to evaluate the contributions of the AEFE, DSTM module, and TCD. AEFE improves the baseline by employing anatomical relation priors, while DSTM further provides precise segmentation by exploiting the temporal context of Echo. The TCD achieves additional enhancement by refining current-frame predictions with temporal anatomical context.

Attention visualization. Figure 4 presents attention maps visualization of AST-Seg compared with DINOv3 [38]. While DINOv3 primarily attends to regions with similar local appearance, AST-Seg provides anatomically structured attention that captures the token-level relationships among spatial tokens belonging to the same cardiac structure. The qualitative assessment demonstrates that the proposed anatomy-aware representation learning enables token-level anatomical relation recognition beyond low-level visual similarity.

4.4 Analysis of Anatomy-aware Representation Learning

We assess the effect of anatomy-aware representation learning by comparing AST-Seg with comparative representation learning methods, including MAE [32], DISCOVR [39], EchoCardMAE [40], and USFM [41]. As shown in Table 4-(a), the anatomy-aware feature learning consistently outperforms comparative schemes. While large-scale self-supervised encoders provide useful appearance representations, they do not explicitly encode the anatomical context required for precise Echo segmentation. In contrast, AST-Seg supervises token-level anatomical relations during representation learning, leading to improved segmentation performance.

4.5 Analysis of Spatio-Temporal Modeling

The effectiveness of DSTM is assessed with competitive temporal modeling methods, including Video Mamba and a Video Transformer. As shown in Table 4-(b), the temporal schemes generally

Table 3: Ablation study of AST-Seg components on CAMUS.

Variant	Components			LV BP			LV wall			LA BP			Params (M)	FLOPs/frame (G)
	AEFE	DSTM	TCD	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓		
Baseline	–	–	–	90.01	82.21	5.49	84.03	72.78	6.45	87.85	78.85	8.45	97.54	23.53
+ Anatomy-attn.	✓	–	–	93.32	87.56	5.04	86.77	76.82	6.21	88.64	80.09	7.84	108.80	26.11
+ DSTM	✓	✓	–	94.76	90.13	4.88	88.55	79.51	5.98	89.19	81.09	7.37	110.09	26.74
+ Temporal Decoder	✓	✓	✓	**95.03**	**90.57**	**4.75**	**89.00**	**80.29**	**5.89**	**89.52**	**81.53**	**6.86**	115.47	26.91

Table 4: Analysis of representation learning and the temporal modeling scheme on CAMUS.

(a) Representation learning.

Encoder	LV BP			LV wall			LA BP		
	Dice	mIoU	HD95	Dice	mIoU	HD95	Dice	mIoU	HD95
MAE	89.28	80.91	7.37	81.86	69.78	6.93	85.40	75.61	9.09
DISCOVR	92.23	85.77	6.89	84.86	74.02	7.37	86.67	77.49	9.05
EchoCardMAE	91.98	85.30	6.84	84.61	73.58	7.10	88.37	80.05	8.54
USFM	90.03	82.21	6.41	82.28	70.09	6.42	86.64	77.62	8.50
Ours w.o. DSTM	**93.32**	**87.56**	**5.04**	**86.77**	**76.82**	**6.21**	**88.64**	**80.09**	**7.84**

(b) Temporal modeling.

Temporal module	LV BP			LV wall			LA BP			Module FLOPs (G)
	Dice	mIoU	HD95	Dice	mIoU	HD95	Dice	mIoU	HD95	
None	93.32	87.56	5.04	86.77	76.82	6.21	88.64	80.09	7.84	0.00
Video Mamba	94.08	88.89	5.85	87.43	77.81	6.01	89.18	81.01	7.80	0.54
Video Trans.	94.63	89.86	5.74	87.49	77.87	5.93	88.40	79.88	7.69	1.07
DSTM	**95.03**	**90.57**	**4.75**	**89.00**	**80.29**	**5.89**	**89.52**	**81.53**	**6.86**	0.63

improve over the image-based baseline, confirming the effectiveness of employing temporal context in Echo. However, directly modeling video tokens with generic temporal architectures is not optimal for Echo. Video Mamba serializes spatial patches into a one-dimensional sequence, while Video Transformer relies on global token interactions that are less aligned with the localized and periodic nature of cardiac motion. In contrast, DSTM adaptively aggregates spatial features and interprets the temporal sequence at each spatial location. The DSTM reduces the computational cost compared with the Transformer, while achieving enhanced segmentation performance under a computational resource comparable to the Video Mamba.

5 Conclusion

In this paper, we present AST-Seg, an anatomy-aware spatio-temporal framework for domain-generalized echocardiography segmentation. By employing anatomical supervision with deformable spatio-temporal state-space modeling, AST-Seg incorporates the anatomical and temporal priors that underlie echocardiographic interpretation. Together with the newly curated EA dataset covering 12 clinically relevant cardiac anatomies, AST-Seg achieves precise segmentation compared with conventional, SAM-based, and video segmentation baselines under both in-distribution and out-of-distribution environments. The results demonstrate that anatomy-aware representation learning and the proposed temporal modeling offer an effective framework for robust echocardiography video segmentation.

Acknowledgments and Disclosure of Funding

This work received no third-party funding. The authors declare no competing interests.

References

[1] Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015.

[2] Roberto M Lang, Michelle Bierig, Richard B Devereux, Frank A Flachskampf, Elyse Foster, Patricia A Pellikka, Michael H Picard, Mary J Roman, James Seward, Jack Shanewise, et al. Recommendations for chamber quantification. *European journal of echocardiography*, 7(2):79–108, 2006.

[3] Helmut Baumgartner, Volkmar Falk, Jeroen J Bax, Michele De Bonis, Christian Hamm, Per Johan Holm, Bernard Iung, Patrizio Lancellotti, Emmanuel Lansac, Daniel Rodriguez Muñoz, et al. 2017 esc/eacts guidelines for the management of valvular heart disease. *Polish Heart Journal (Kardiologia Polska)*, 76(1):1–62, 2018.

[4] Sridhar Venkatachalam, Geru Wu, and Masood Ahmad. Echocardiographic assessment of the right ventricle in the current era: Application in clinical practice. *Echocardiography*, 34(12):1930–1947, 2017.

[5] Lawrence G Rudski, Wyman W Lai, Jonathan Afilalo, Lanqi Hua, Mark D Handschumacher, Krishnaswamy Chandrasekaran, Scott D Solomon, Eric K Louie, and Nelson B Schiller. Guidelines for the echocardiographic assessment of the right heart in adults: a report from the american society of echocardiography: endorsed by the european association of echocardiography, a registered branch of the european society of cardiology, and the canadian society of echocardiography. *Journal of the American society of echocardiography*, 23(7):685–713, 2010.

[6] Steven A Goldstein, Arturo Evangelista, Suhny Abbara, Andrew Arai, Federico M Asch, Luigi P Badano, Michael A Bolen, Heidi M Connolly, Hug Cuéllar-Calàbria, Martin Czerny, et al. Multimodality imaging of diseases of the thoracic aorta in adults: from the american society of echocardiography and the european association of cardiovascular imaging: endorsed by the society of cardiovascular computed tomography and society for cardiovascular magnetic resonance. *Journal of the American Society of Echocardiography*, 28(2):119–182, 2015.

[7] Paul A Vignola, Antoine Bloch, Alan D Kaplan, Harriet J Walker, Philip N Chiotellis, and Gordon S Myers. Interobserver variability in echocardiography. *Journal of Clinical Ultrasound*, 5(4):238–242, 1977.

[8] Victor Mor-Avi, Bijoy Khandheria, Robert Klempfner, Juan I Cotella, Merav Moreno, Denise Ignatowski, Brittney Guile, Hailee J Hayes, Kyle Hipke, Abigail Kaminski, et al. Real-time artificial intelligence–based guidance of echocardiographic imaging by novices: image quality and suitability for diagnostic interpretation and quantitative analysis. *Circulation: Cardiovascular Imaging*, 16(11):e015569, 2023.

[9] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:508599, 2020.

[10] Maxime Sermesant, Hervé Delingette, Hubert Cochet, Pierre Jaïs, and Nicholas Ayache. Applications of artificial intelligence in cardiovascular imaging. *Nature Reviews Cardiology*, 18(8):600–609, 2021.

[11] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. Deep learning interpretation of echocardiograms. *NPJ digital medicine*, 3(1):10, 2020.

[12] Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. Memsam: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9622–9631, 2024.

[13] J Alison Noble and Djamal Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on medical imaging*, 25(8):987–1010, 2006.

[14] Seok-Hwan Oh, Guil Jung, Sang-Yun Kim, Myeong-Gee Kim, Young-Min Kim, Hyeon-Jik Lee, Hyuk-Sool Kwon, and Hyeon-Min Bae. Uncertainty-aware meta-weighted optimization framework for domain-generalized medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 775–785. Springer, 2024.

[15] Roberto Vega, Masood Dehghan, Arun Nagdev, Brian Buchanan, Jeevesh Kapur, Jacob L Jaremko, and Dornoosh Zonoobi. Overcoming barriers in the use of artificial intelligence in point of care ultrasound. *NPJ Digital Medicine*, 8(1):213, 2025.

[16] 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, page 2, 2019.

[17] 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.

[18] Samaneh Mazaheri, Puteri Suhaiza Binti Sulaiman, Rahmita Wirza, Fatimah Khalid, Suhaini Kadiman, Mohd Zamrin Dimon, and Rohollah Moosavi Tayebi. Echocardiography image segmentation: A survey. In *2013 international conference on advanced computer science applications and technologies*, pages 327–332. IEEE, 2013.

[19] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021.

[20] William J. McKenna, Angelika Kleinebenne, Petros Nihoyannopoulos, and Rodney Foale. Echocardiographic measurement of right ventricular wall thickness in hypertrophic cardiomyopathy: Relation to clinical and prognostic features. *Journal of the American College of Cardiology*, 11(2):351–358, 1988.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[22] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[23] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[24] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2024.

[25] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[27] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024.

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[33] 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*, pages 234–241. Springer, 2015.

[34] 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*, pages 205–218. Springer, 2022.

[35] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9):2763–2775, 2023.

[36] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[37] Sekeun Kim, Pengfei Jin, Cheng Chen, Kyungsang Kim, Zhiliang Lyu, Hui Ren, Sunghwan Kim, Zhengliang Liu, Aoxiao Zhong, Tianming Liu, et al. Medivista: Medical video segmentation via temporal fusion sam adaptation for echocardiography. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[38] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[39] Divyanshu Mishra, Mohammadreza Salehi, Pramit Saha, Olga Patey, Aris T Papageorghiou, Yuki M Asano, and J Alison Noble. Self-supervised learning of echocardiographic video representations via online cluster distillation. *arXiv preprint arXiv:2506.11777*, 2025.

[40] Xuan Yang, Rui Xu, Xinchen Ye, Zhihui Wang, Miao Zhang, Yi Wang, Xin Fan, Hongkai Wang, Qingxiong Yue, Xiangjian He, et al. Echocardmae: Video masked auto-encoders customized for echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 171–180. Springer, 2025.

[41] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.

A Details of the Curated Datasets

In this section, details on the datasets curated for representation learning, anatomy-aware segmentation, and out-of-distribution evaluation are described. The study employs three complementary echocardiography datasets: a large-scale annotation-free video dataset for self-supervised pretraining, an echocardiography anatomy dataset with annotations for cardiac segmentation, and a point-of-care echocardiography dataset for evaluating domain generalization.

A.1 Echocardiography Video Dataset

For self-supervised representation learning, we curate a large-scale annotation-free echocardiography dataset consisting of 7,901,515 image frames extracted from 203,352 Echo videos. The dataset is employed to pretrain a transformer encoder. The pre-training is performed to learn echocardiography visual representations from large-scale clinical data.

The dataset includes 14 standard echocardiography views, which are Apical 2-Chamber (A2C), Apical 3-Chamber (A3C), Apical 4-Chamber (A4C), Parasternal Long-Axis (PLAX), Parasternal Short-Axis at the Apex (PSAX APEX), Parasternal Short-Axis at the Aortic Valve Level (PSAX AV), Parasternal Short-Axis at the Mitral Valve Level (PSAX MV), Parasternal Short-Axis at the Papillary Muscle Level (PSAX PM), RV Apical 4-Chamber (RVA4C), RV Inflow View (RVIV), RV Outflow View (RVOV), Subcostal 4-Chamber View (SC4V), Subcostal Long-Axis View (SCLA), and Suprasternal Notch View (SSN). The dataset covers diverse anatomical structures and Echo views, which is significant for learning generalizable echocardiography representations.

To improve acquisition diversity, the dataset is collected from multiple ultrasound vendors. The dataset exhibits substantial variability in image appearance, speckle statistics, spatial resolution, and vendor-specific post-processing. Such variability is significant for learning representations that are robust to real-world clinical heterogeneity.

A.2 Echocardiography Anatomy Dataset

We introduce a newly curated echocardiography anatomy (EA) dataset with pixel-level annotations for 12 clinically relevant cardiac structures across three apical views. The dataset contains 21,535 annotated frames from 3,110 subjects and is designed to provide comprehensive anatomical supervision for echocardiography segmentation. The dataset is annotated with the following four-stage protocol: (1) annotator training, (2) initial annotation, (3) expert review, and (4) revision. Prior to annotation, the sonographers were trained with a standardized manual in accordance with the recommendations of the American Society of Echocardiography and the European Association of Cardiovascular Imaging. After training, each study case was annotated by a sonographer and independently reviewed by the expert physician.

Unlike existing echocardiography segmentation datasets that commonly focus on a small number of structures, the EA dataset provides dense annotations for 12 anatomical regions, which are LV BP, LV wall, LA BP, LA wall, Mitral Valve (MV), Right Ventricle (RV) BP, RV wall, Right Atrium (RA) BP, RA wall, Tricuspid Valve (TV), Aortic Root (AO root), and Aortic Valve (AV). In order to improve morphological diversity and reduce redundant annotations, one representative frame is selected from each echocardiography video. The sampling is performed to avoid repeatedly annotating adjacent frames with highly similar anatomical configurations while increasing the diversity of subjects, views, and image appearances included in the dataset.

The dataset includes images acquired using EPIQ 7C and EPIQ CVx from Philips Healthcare, Vivid E9 and Vivid IQ from GE Healthcare, and ACUSON SC2000 from Siemens. The diversity of the dataset contributes to robustness to vendor-dependent image characteristics, including differences in speckle pattern, contrast, gain, resolution, and post-processing. The dataset is split at the case level, with 19,125 annotated frames used for training and 2,410 frames used for testing.

A.3 Point-of-Care Echocardiography Dataset

To evaluate out-of-distribution generalization, we curate the point-of-care echocardiography (PoC-Echo) dataset. The dataset is acquired using portable ultrasound systems, including Vscan from GE Healthcare, Butterfly iQ from Butterfly Network, and Lumify from Philips Healthcare. Compared

with high-end console-based echocardiography systems, point-of-care devices produce images with lower spatial resolution, reduced signal-to-noise ratio, and stronger artifacts.

The dataset contains 259 samples with pixel-level annotations of the LV BP. The annotations were performed by one experienced sonographer and reviewed by an expert physician. The point-of-care dataset serves as a clinically meaningful robustness benchmark. Since portable echocardiography is increasingly used in bedside, emergency, and resource-limited settings, strong performance on the dataset suggests improved potential for real-world deployment.

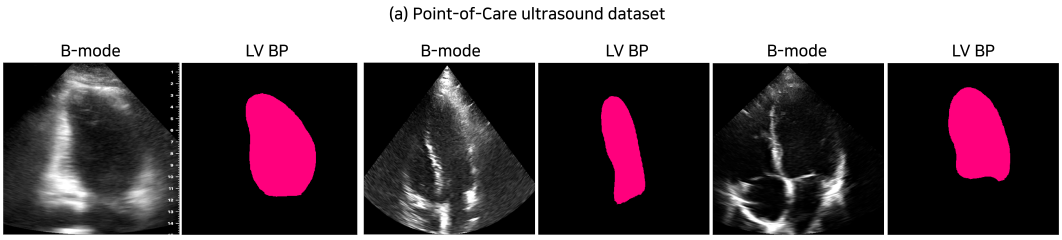

Figure 5: Point-of-Care Echocardiography dataset cases.

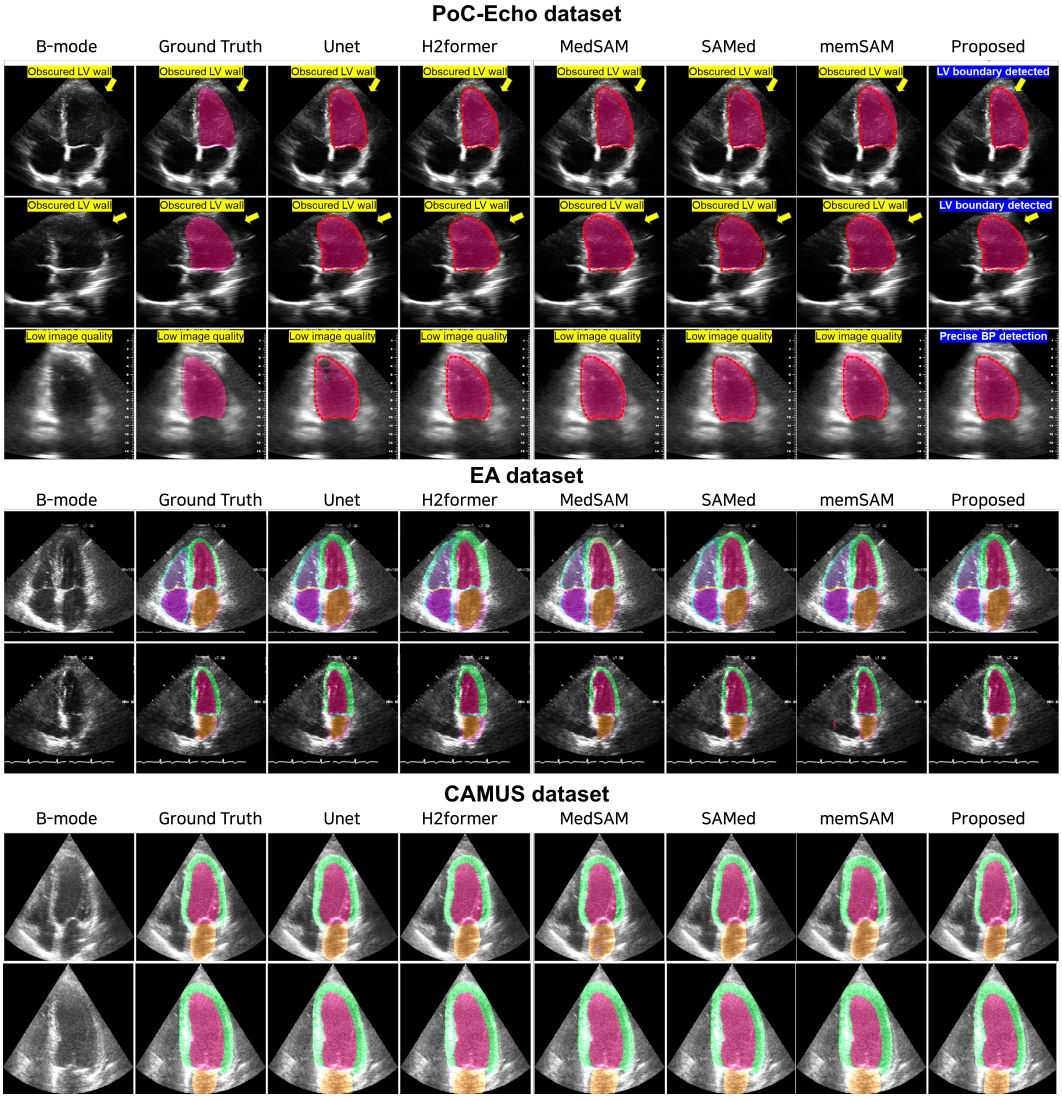

Figure 6: Qualitative segmentation results

B AST-Seg Configuration Details

In this section, we describe the details of AST-Seg configuration. We employed ViT-Base as the backbone encoder with a patch size of 16×16. The temporal input clip length is set to $T = 10$. Given an input clip, AST-Seg predicts the segmentation mask of the target frame, while the remaining frames provide temporal context for motion-aware anatomical refinement.

For Echo segmentation, the model is trained for 200 epochs using AdamW with a batch size of 16, a learning rate of 10^{-4} and weight decay of 0.01. We employ a cosine annealing learning rate scheduler. To improve generalization across echocardiography acquisition conditions, we apply color jittering, resizing, and random cropping during training. Frame-level augmentations are applied consistently within each clip to avoid disrupting temporal correspondence. At inference, AST-Seg predicts a segmentation mask for each queried anatomical condition. Experiments were performed using an NVIDIA H200 GPU with 141 GB memory.

C Extended experiments

C.1 Qualitative assessments

Figure 6 shows qualitative segmentation results of AST-Seg and comparative baselines on PoC-Echo, EA, and CAMUS datasets. Under the out-of-distribution PoC-Echo dataset, where image quality is degraded and anatomical boundaries are ambiguous, AST-Seg achieves precise segmentations compared with the competitive methods. On the EA dataset, AST-Seg accurately reconstructs diverse cardiac structures, including challenging anatomies such as the mitral valve and right ventricle, demonstrating its ability to support comprehensive cardiac anatomy interpretation. On CAMUS dataset, AST-Seg provides accurate delineation of the LV BP, LV wall, and LA BP, confirming its effectiveness under conventional echocardiography benchmarks.

C.2 Effect of anatomy supervision

To isolate the contribution of anatomy supervision, we compare two AST-Seg variants that retain the anatomy attention branches and are trained with and without the attention-supervision loss. As shown in Table 5, anatomy supervision improves mean Dice score by 1.46 across the three CAMUS structures, with consistent improvements in mIoU and HD95.

Table 5: Effect of anatomy supervision on CAMUS. Both AST-Seg variants retain the anatomy attention branches and exclude DSTM and temporal refinement.

Model	Anatomy supervision	LV BP			LV wall			LA BP		
		Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓
AST-Seg w.o. DSTM	–	91.33	84.34	5.39	85.09	74.36	6.37	87.94	79.01	8.19
	✓	**93.32**	**87.56**	**5.04**	**86.77**	**76.82**	**6.21**	**88.64**	**80.09**	**7.84**

C.3 Effect of encoder initialization

We evaluate the full AST-Seg model under three encoder initialization schemes, which are random initialization, MIM pretraining on the 7.9M-frame annotation-free Echo dataset, and the proposed anatomy-aware representation learning (ARL), which further adapts the MIM-pretrained encoder using anatomical supervision. As shown in Table 6, MIM pretraining improves performance over random initialization, while the proposed representation learning yields further improvements.

C.4 Effect of the EA dataset for echocardiography domain generalization

We evaluate the contribution of the EA dataset to domain-generalized echocardiography segmentation by training representative baselines under two data configurations, which are CAMUS only and CAMUS+EA dataset. The networks are evaluated on the out-of-distribution PoC-Echo dataset, which comprises portable ultrasound images and exhibits substantial domain shift relative to conventional

Table 6: Effect of encoder initialization on CAMUS dataset.

Encoder initialization	LV BP			LV wall			LA BP		
	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓
Random initialization	93.29	87.55	5.03	87.31	77.59	6.13	88.43	79.81	7.81
MIM pretraining	94.02	88.81	4.92	88.03	78.73	6.05	88.96	80.63	7.47
Proposed ARL	**95.03**	**90.57**	**4.75**	**89.00**	**80.29**	**5.89**	**89.52**	**81.53**	**6.86**

echocardiography. Table 7 presents the quantitative results for conventional neural networks and SAM-based baselines. Integrating the EA dataset into the training set consistently improves segmentation performance across all baselines, yielding higher Dice and mIoU as well as lower HD95 on the OOD PoC-Echo benchmark. The results indicate that the EA dataset provides more generalized echocardiography segmentation by introducing echocardiography images obtained with diversified ultrasound vendors and models and acquisition environments.

Table 7: Effect of the EA dataset on out-of-distribution echocardiography segmentation.

Category	Method	Training dataset	PoC-Echo (OOD) – LV BP		
			Dice ↑	mIoU ↑	HD95 ↓
Conventional NN	U-Net	CAMUS	61.49	49.83	47.47
		CAMUS + EA	80.90 (+19.41)	71.28 (+21.45)	25.66 (−21.81)
	SwinUNet	CAMUS	66.15	53.04	51.93
		CAMUS + EA	82.91 (+16.76)	73.24 (+20.20)	27.31 (−24.62)
SAM-based	MSA	CAMUS	76.80	64.71	31.24
		CAMUS + EA	86.19 (+9.39)	77.57 (+12.86)	21.27 (−9.97)
	SAMed	CAMUS	78.40	66.35	28.21
		CAMUS + EA	86.37 (+7.97)	79.22 (+12.87)	19.83 (−8.38)
	SAMUS	CAMUS	82.71	74.47	17.96
		CAMUS + EA	86.14 (+3.43)	79.02 (+4.55)	13.50 (−4.46)
Proposed	AST-Seg	CAMUS + EA	**90.83**	**83.59**	**11.21**

C.5 Computational cost assessment

We measured the parameter counts and inference latency of AST-Seg and the comparative baselines. Among the video-based methods, AST-Seg has the fewest parameters and achieves the lowest per-frame latency. The computation cost of AST-Seg is also comparable to that of single-frame SAM-based methods such as MSA and SAMed while achieving higher segmentation accuracy. Conventional architectures, including U-Net, SwinUNet, and H2Former, are more computationally efficient. However, the network presents consistently lower performance in both in-distribution and out-of-distribution evaluation.

Table 8: Computational cost comparison. Inference latency is measured on a single NVIDIA RTX 4090 GPU.

Method	Params (M)	Latency / frame (ms)
U-Net	31.03	5.38
SwinUNet	27.16	13.0
H2Former	33.69	8.54
MedSAM	90.74	11.5
MSA	104.3	15.0
SAMed	90.60	12.8
SAMUS	133.3	20.5
MemSAM	152.0	19.4
MediViSTA	171.9	16.7
AST-Seg	**115.5**	**15.1**

C.6 Assessment of left ventricular ejection fraction

We assess the clinical relevance of segmentation performance through left ventricular ejection fraction (EF) estimation on CAMUS. EF is calculated from end-diastolic and end-systolic volumes obtained using Simpson's biplane method. As shown in Table 9, AST-Seg achieves the highest Pearson correlation with reference EF ($r = 0.8711$) and the lowest MAE (3.36%) among the evaluated methods, demonstrating improved estimation of cardiac function alongside its segmentation gains.

Table 9: Assessment of left ventricular ejection fraction on CAMUS.

Method	Pearson coefficient ↑	MAE (%) ↓
U-Net	0.7945	6.31
H2Former	0.8010	5.50
MedSAM	0.6145	7.23
MSA	0.8058	5.45
SAMed	0.8115	4.53
MemSAM	0.8279	4.26
MediViSTA	0.7261	7.37
AST-Seg	**0.8711**	**3.36**

D Limitation

In this paper, we propose a framework for extensive echocardiography anatomy segmentation. Although the framework expands anatomical coverage compared with existing echocardiography segmentation benchmarks, the current framework and curated dataset are limited to apical views. Extending echocardiography interpretation to parasternal, subcostal, and suprasternal views would further broaden the applicability of anatomy-aware segmentation to more comprehensive echocardiographic interpretation. AST-Seg explicitly supervises intra-structure token relations, whereas inter-structure interactions are modeled implicitly through feature aggregation in the global anatomy branch. Explicitly modeling these interactions could more faithfully reflect clinical anatomical reasoning and represents a promising direction for future research. In addition, the OOD evaluation is limited to LV blood-pool segmentation under the portable-device domain shift represented by PoC-Echo. Further investigation across additional institutions, operators, acquisition conditions, and anatomical structures would support a more rigorous assessment of the framework's generalizability across domains.

E Ethics statement

All patient data used in this study were obtained under Institutional Review Board (IRB) approval and were fully anonymized in compliance with the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper clearly states the main contributions and scope of the work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper describes the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not introduce theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides implementation details to reproduce the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be publicly released at https://github.com/joseph9337/AST-Seg. The datasets can be accessed upon request, subject to IRB approval.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the implementation and experimental details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have reported the quantitative results with appropriate statistics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper describes the computational resource of the proposed scheme.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and conducted the study in accordance with the principles.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the potential clinical impact of the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The study uses de-identified clinical data in accordance with HIPAA and GDPR guidance.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets and methods used in the paper are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The details of newly curated data are described.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study was conducted under IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The LLM is used only for writing, editing, or formatting.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.