Echo-SAM: Zero-Shot Learning of Unseen Structures via Medical Knowledge Graph Grounding

Jiewen Yang **Xiaomeng Li***
The Hong Kong University of Science and Technology
jyangcu@connect.ust.hk eexmli@ust.hk

Abstract

Segmenting unseen anatomical structures in echocardiographic images is challenging because dense annotations are scarce and ultrasound often presents visually similar textures across different cardiac structures. Current foundation models remain limited in this setting because they treat anatomical knowledge as independent text prompts, overlooking the spatial topology that clinicians use to localize ambiguous structures from their visible neighbors. We propose **Echo-SAM**, a zero-shot ultrasound segmentation framework that grounds a structured cardiac knowledge graph in visual space through relation-conditioned geometric support inference. Echo-SAM localizes seen structures as anatomical anchors, propagates relational constraints through a knowledge-grounded Graph Neural Network, and enhances text prompts with image-grounded anatomical context. We further introduce a topology-to-geometry mapping that converts graph relations into geometry-aware support proposal scores, yielding posterior-guided support estimates for unseen structures. By treating relation-specific scale, thickness, and offset as latent geometric variables, Echo-SAM adapts graph-derived support proposals to image-specific anchors without relying on fixed coordinate templates. Extensive evaluations on zero-shot echocardiography segmentation benchmarks demonstrate that Echo-SAM substantially outperforms state-of-the-art foundation models, achieving up to 50.42% absolute improvement in Dice score. Code is available at: https://github.com/xmed-lab/Echo-SAM.

1 Introduction

Echocardiography is the primary imaging modality for cardiac assessment in clinical practice, providing real-time visualization of chamber morphology and myocardial dynamics at low cost and without radiation. Accurate delineation of cardiac structures [1] from echocardiographic images is essential for quantifying cardiac function, diagnosing disease, and guiding interventional planning [2]. However, the high intra- and inter-observer variability inherent to ultrasound acquisition, combined with the fine-grained nature of structures such as valve leaflets and septal boundaries, makes pixel-level annotation exceptionally laborious and demands specialist expertise rarely available at scale.

To address annotation scarcity, zero-shot segmentation methods in the natural image domain [3–6] often rely on massive pre-training corpora that implicitly encode unseen categories. Their generalization is further strengthened by vision-language alignment, such as CLIP [7], which enables semantic perception through soft labels. However, medical imaging fundamentally lacks access to such massive and diverse datasets, rendering these data-dependent strategies less directly applicable to clinical scenarios. Consequently, recent work has progressed along two complementary trajectories. Foundation models such as the Segment Anything Model (SAM) [8–10] and its medical adaptations [11–16]

*Corresponding Author

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

achieve broad anatomical coverage through prompt-based interaction, generalizing across modalities without task-specific fine-tuning. In parallel, knowledge-guided methods [16–18] inject domain expertise into vision-language frameworks by aligning textual or anatomical descriptions with visual features, enabling recognition of structures via semantic similarity rather than solely supervised patterns. Collectively, these approaches represent a meaningful shift from data-hungry pipelines toward more generalizable frameworks.

However, a fundamental limitation persists across all existing methods [8–18], where anatomical knowledge is treated as a flat, unstructured sequence of tokens, reducing rich relational structure to a bag of semantic labels. This design prevents the model from reasoning over anatomical topology, for instance, the fact that the mitral valve must geometrically reside between the left atrium and ventricle. Without explicitly propagating spatial constraints across structure boundaries, current visual-language-based methods [12, 13, 17, 19–21] are largely limited to correlating semantic embeddings with visual textures. Consequently, they struggle to infer the location of an unannotated structure from its annotated neighbors. Furthermore, their anatomical text representations are often weakly grounded in image-specific visual evidence, creating a persistent semantic gap between clinical knowledge and image-conditioned representations.

To address these limitations, we present **Echo-SAM**, a framework for anchor-supported target-level zero-shot echocardiographic segmentation. Echo-SAM first localizes routinely annotated structures as seen anatomical anchors, propagates their relational constraints through a cardiac knowledge graph, converts these constraints into geometry-aware support estimates for unseen targets, and finally uses the inferred support together with knowledge-enhanced prompts to guide SAM3 segmentation. Unlike methods that recognize novel structures mainly by embedding similarity, our model formulates unseen-structure localization as posterior-guided geometric support inference conditioned on anatomical topology and image-specific anchors. Concretely, Echo-SAM constructs a cardiac knowledge graph in which node embeddings encode anatomical concepts and edge embeddings encode typed spatial relations. A relation-conditioned graph grounding module produces both enhanced target prompts and relation relevance weights, allowing graph knowledge to affect decoder conditioning and topology-to-geometry support inference. Instead of treating graph relations as static semantic context, we instantiate them as image-conditioned support proposals around localized anatomical anchors. The resulting posterior-guided support score selects pseudo targets that provide spatial constraints for unseen structures without using their masks, boxes, points, or validation labels during training. Our main contributions are:

- A relation-conditioned medical knowledge graph with learnable node and edge embeddings, producing both knowledge-enhanced prompts and relation relevance weights for unseen target inference.
- A topology-to-geometry support inference framework that maps anatomical relations and seen anchors to geometry-aware support proposal scores, enabling posterior-guided estimation of unseen target support without target-specific localization prompts.
- Extensive experiments on echocardiographic benchmarks demonstrating that Echo-SAM significantly outperforms state-of-the-art foundation models in zero-shot settings, with up to 50.42% absolute Dice improvement.

2 Related Work

2.1 Segmentation in Cardiac Ultrasound Imaging

Automated segmentation of cardiac structures from echocardiographic sequences has been extensively studied as a prerequisite for clinical indices such as left ventricular ejection fraction and cardiac motion estimation [22–26]. Early convolutional approaches, including U-Net variants and their domain-adapted successors [27, 28], established strong baselines for LV endocardial and epicardial boundary delineation, while recurrent and spatio-temporal architectures [29] further exploited inter-frame consistency to improve segmentation stability across the cardiac cycle [22, 26, 30, 31]. Recent cardiac segmentation baselines include SimLVSeg [32], PKEchoNet [33], DSA [34], and GDKVM [35], alongside Vision LSTM [36]. More recent transformer-based models have incorporated long-range spatial dependencies to handle the high geometric variability characteristic of echocardiographic views [24, 26, 37].

Despite these advances, the field has converged on a limited anatomical scope. Most frameworks are designed and benchmarked primarily on LV segmentation, while myocardium, atria, valves, and septal structures are less consistently modeled or benchmarked. This bias reflects the annotation scarcity in echocardiographic datasets, where LV labels are most abundant and directly tied to actionable clinical metrics. Consequently, even SAM-based cardiac adaptations [11, 14, 15] that inherit the universal segmentation capacity of foundation models are typically fine-tuned on LV-centric data, effectively hindering open-set generalization. As a result, these models remain fragile when encountering structures absent from training, rendering them unsuitable for the comprehensive anatomical coverage required in clinical workflows.

2.2 Prompt-Guided Segmentation in Medical Imaging

The Segment Anything Model [8–10] introduced a prompt-based interaction paradigm that decouples segmentation from category-specific supervision, and its successors [8, 10] have extended this to video and high-dimensional inputs. Medical adaptations such as MedSAM [15], MedicoSAM [16], MemSAM [11], and Medical SAM Adapter (Med-SA) [14] improve medical-domain precision through curated medical corpora or prompt propagation across echocardiographic frames. However, these approaches often rely on bounding box or point prompts extracted from ground-truth annotations, leaking target-location information unavailable during zero-shot inference. In parallel, MedCLIP-SAM [12, 13], MedROV-SAM [17], and related vision-language methods [18] align free-text anatomical descriptions with image embeddings, enabling structure recognition driven by semantic similarity rather than pixel-level supervision.

However, significant limitations persist in these prompt-guided approaches. They primarily rely on language as a soft semantic anchor rather than structured spatial reasoning. Given the differences between medical imaging and natural photography, zero-shot descriptions, even with domain adaptation, do not explicitly encode the relational topology exploited by clinicians. Text embeddings are also often computed offline and fixed throughout training, limiting feedback between visual evidence and medical knowledge. Consequently, while these models can identify structures based on appearance similarity, they have limited ability to localize an unannotated structure relative to its annotated neighbors, where anatomical layout provides essential spatial cues [38, 39], especially for thin or weakly contrasted cardiac structures.

3 Methodology

We present **Echo-SAM**, a framework for zero-shot echocardiographic segmentation that grounds structured medical knowledge graphs in visual representations. Unlike conventional data-driven approaches (see section 2), Echo-SAM leverages explicit anatomical relationships to infer the spatial locations of unseen structures. The overall pipeline is illustrated in Figure 1.

Problem Formulation. Let $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ be an echocardiographic image defined on the 2D image domain Ω. We consider an anatomical ontology $\mathcal{S} = \mathcal{S}_{\text{seen}} \cup \mathcal{S}_{\text{unseen}}$, where $\mathcal{S}_{\text{seen}} \cap \mathcal{S}_{\text{unseen}} = \emptyset$. The training set $\mathcal{D}_{\text{train}}$ provides masks and boxes only for structures in $\mathcal{S}_{\text{seen}}$; ground-truth masks, boxes, points, and validation labels for structures in $\mathcal{S}_{\text{unseen}}$ are excluded from training and model selection. At inference, given a target structure $s_n \in \mathcal{S}_{\text{unseen}}$, the goal is to predict its latent support $Y_n = (M_n, \mathbf{b}_n)$, where $M_n : \Omega \to \{0, 1\}$ is the target mask and $\mathbf{b}_n \in [0, 1]^4$ is its normalized bounding box. To bridge this annotation gap, we employ a medical knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to align visual features with structured anatomical priors over the full ontology $\mathcal{S}$, enabling knowledge-guided generalization from seen anchors to unseen structures.

3.1 Medical Knowledge Graph Construction and Grounding

Existing vision-language models treat anatomical knowledge as unstructured text and do not explicitly capture the relational topology clinicians rely on (e.g., inferring a valve location from adjacent ventricles). To bridge this semantic gap, we decompose medical knowledge into complementary node-level (morphology) and edge-level (topology) components, transforming static anatomical knowledge into dynamic, learnable visual representations. This decomposition enables explicit neural reasoning over anatomical structures, allowing the model to aggregate contextual information from related neighbors to enhance segmentation of unseen targets.

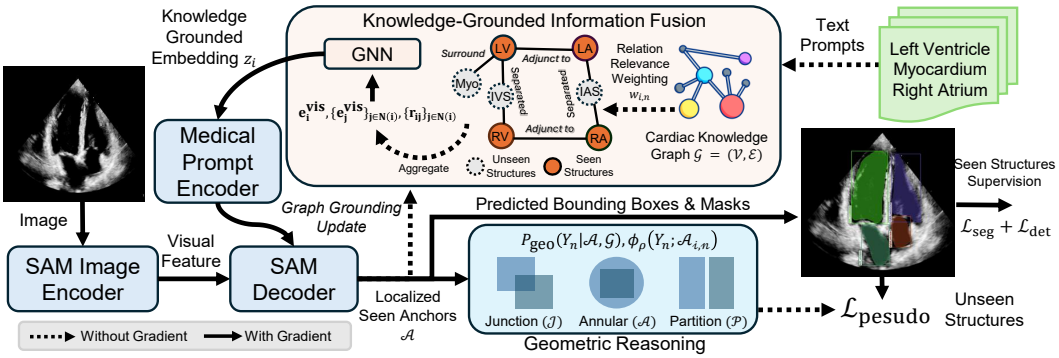


Figure 1: Overview of the Echo-SAM framework. The Medical Knowledge Graph grounds anatomical morphology and topology in visual representations, producing relation-conditioned prompts for unseen targets. The Geometric Support Inference Module converts graph relations into geometry-aware support estimates for anatomically plausible unseen structures.

Knowledge Graph Definition. Formally, we construct $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from standardized cardiac anatomy references and echocardiographic view guidelines, followed by expert review. This knowledge-driven construction does not require pixel-level annotations for unseen structures, preserving its zero-shot scalability. Each node $v_i \in \mathcal{V}$ corresponds to an anatomical structure $s_i \in \mathcal{S}$. Each edge $e_{ij} = (v_i, \rho_{ij}, v_j) \in \mathcal{E}$ represents a typed anatomical relation, where ρ_{ij} belongs to a finite relation vocabulary $\mathcal{R}$, including adjacency, separation, continuity, and enclosure. Each node maintains a morphology-aware representation grounded by visual evidence and a semantic representation derived from anatomical descriptions. Each edge is associated with a learnable relation embedding $\mathbf{r}_{ij}$ that represents the geometric semantics of relation type ρ_{ij}. The graph therefore provides both concept-level semantics and relation-level structure for unseen target inference. The formulation is category-independent: in principle, any cardiac structure can be represented by a node and localized through valid relations to available anchors. Chambers, walls, septa, valves, and great vessels share the same inference mechanism; the acquisition view determines which structures and relations are observable. Detailed graph construction and relation definitions are provided in Appendix V.1.

3.2 Knowledge-Grounded Information Fusion

To bridge the semantic gap between medical knowledge and visual features, we employ a Graph Neural Network (GNN) for relation-conditioned prompt grounding. The GNN does not directly generate masks or coordinates. Instead, it has two explicit roles: it converts neighboring anatomical evidence into a target-specific prompt embedding, and it estimates relation relevance scores that weight the geometric support scores used in the support inference below. Thus, learnable edge embeddings affect both decoder conditioning and topology-to-geometry support inference, rather than serving only as auxiliary prompt fusion features.

For each graph node v_i, we use the structures directly connected to it in the anatomy graph as contextual evidence. Specifically, $\mathcal{N}(v_i) = \{v_j \mid e_{ij} \in \mathcal{E}\}$ denotes the set of anatomical neighbors of v_i. The node feature $\mathbf{e}_i^{\text{vis}}$ represents the image-conditioned evidence for structure s_i: for seen structures it is extracted from the localized anchor region, while for unseen structures it is initialized from the semantic prompt and grounded through messages from related seen nodes. The graph fusion function $F_{\text{GNN}}(\cdot)$ then combines the current node feature, neighboring node features, and relation embeddings into a target-specific representation:

$$\mathbf{z}_i = F_{\text{GNN}}\left(\mathbf{e}_i^{\text{vis}}, \{\mathbf{e}_j^{\text{vis}}\}_{v_j \in \mathcal{N}(v_i)}, \{\mathbf{r}_{ij}\}_{v_j \in \mathcal{N}(v_i)}\right). \tag{1}$$

This representation preserves local morphology while incorporating contextual evidence from anatomically related structures, without using unseen target annotations.

For an unseen target s_n, let $\mathcal{S}_{\text{avail}}(n)$ denote structures whose support is available before estimating s_n: the seen structures and, when required by a cascaded relation, intermediate supports inferred without target annotations. For a candidate dependency $s_i \in \mathcal{D}(s_n)$, where $\mathcal{D}(s_n) \subseteq \mathcal{S}_{\text{avail}}(n)$ is

the set of available structures connected to s_n through valid graph relations, we further compute a relation relevance score from the grounded node embeddings and the corresponding edge embedding. The scoring function $g_{\text{rel}}(\cdot)$ maps this triplet to a scalar compatibility score:

$$a_{i,n} = g_{\text{rel}}(\mathbf{z}_i, \mathbf{z}_n, \mathbf{r}_{i,n}), \qquad w_{i,n} = \frac{\exp(a_{i,n})}{\sum_{s_k \in \mathcal{D}(s_n)} \exp(a_{k,n})}. \tag{2}$$

The set $\mathcal{S}_{\text{unseen}}$ remains fixed: estimating an intermediate support only makes it available in $\mathcal{S}_{\text{avail}}(n)$ and does not make its category seen. The normalized score $w_{i,n}$ is used to weight dependency-specific geometric support scores in Eq. 4. In this way, the GNN determines not only the semantic representation of the target prompt, but also which anatomical relations should dominate the inferred support for that target in a given view.

Prompt Enhancement. The knowledge-grounded embedding is integrated into the text prompt features through a residual fusion $\mathbf{p}_i^{\text{enhanced}} = (1-\alpha)\mathbf{p}_i^{\text{text}} + \alpha \mathbf{z}_i$, where $\mathbf{p}_i^{\text{text}}$ is the original text prompt embedding and $\alpha \in [0, 1]$ controls the contribution of graph-grounded evidence. This formulation preserves the target semantics provided by the language encoder while injecting relation-conditioned visual context from the graph. The enhanced prompt conditions the decoder query, whereas the relation weights in Eq. 2 condition the geometric support score.

Graph Grounding Update. The graph embeddings are updated during training so that stable anatomical knowledge is gradually grounded by visual evidence from seen structures while preserving semantic consistency from text. This update changes the graph from a static ontology into an image-grounded anatomical representation. Implementation details of the graph fusion and momentum update are provided in Appendix V.7.

3.3 Geometric Support Inference for Zero-Shot Structure Localization

We formalize the localization of unseen anatomical structures as posterior-guided geometric support inference. Let $\mathcal{A}_0 = \{(\hat{M}_i, \hat{\mathbf{b}}_i)\}_{s_i \in \mathcal{S}_{\text{seen}}}$ denote the image-specific seen anchors localized by the model, where $\hat{M}_i$ and $\hat{\mathbf{b}}_i$ are the predicted mask and normalized box of seen structure s_i, retaining only available localizations. We write $\mathcal{A}$ for the available support set used for a target, initialized by $\mathcal{A}_0$ and augmented only with model-inferred intermediate supports when a graph relation requires them. No element of $\mathcal{A}$ uses ground-truth masks or boxes for unseen structures. Given $\mathcal{A}$ and the knowledge graph $\mathcal{G}$, our objective is to score plausible latent target supports $Y_n = (M_n, \mathbf{b}_n)$ for an unseen structure s_n. Rather than learning a deterministic coordinate mapping, we define a geometric support score as:

$$S_{\text{geo}}(Y_n | \mathcal{A}, \mathcal{G}) = \Psi_{\text{anchor}}(Y_n; \mathcal{A}, \mathcal{G}) + \Psi_{\text{prior}}(Y_n; \mathcal{A}, \mathcal{G}), \tag{3}$$

Here S_{geo} is a scalar score over candidate supports, not a set of structures; Ψ_{prior} encodes morphology and topology from the knowledge graph, and Ψ_{anchor} enforces spatial consistency between the unseen target and the observed anchors. A normalized support distribution $P_{\text{geo}}(Y_n | \mathcal{A}, \mathcal{G})$ is obtained from this score over sampled support proposals.

Topological-to-Geometric Mapping via Support Proposals. Distinct from prior approaches that rely on soft attention mechanisms to implicitly correlate features, we explicitly represent anatomical relations as geometry-aware support proposal scores. We posit that the relation types in $\mathcal{G}$ induce specific constraints on the support region of s_n. We define a family of conditional support scores $\phi_\rho(Y_n; \mathcal{A})$ indexed by relation type $\rho \in \mathcal{R}$: (1) for structures connected by a tissue bridge, the junction score ϕ_{conn} concentrates support between related anchors along the principal anatomical axis; (2) for surrounding structures, the annular score ϕ_{sur} follows a dilation of the central anchor, capturing concentric cardiac topology; and (3) for structures separated by boundaries, the partition score ϕ_{sep} is defined near the separating interface between anchors, enforcing disjoint support constraints.

Formally, the graph-derived support score is approximated as a mixture of these geometry-aware support scores weighted by relation relevance from the grounded graph:

$$\Psi_{\text{prior}}(Y_n | \mathcal{A}, \mathcal{G}) = \sum\nolimits_{s_i \in \mathcal{D}(s_n)} w_{i,n} \cdot \phi_{\rho_{i,n}}(Y_n; \mathcal{A}_{i,n}), \tag{4}$$

where $\mathcal{A}_{i,n} \subseteq \mathcal{A}$ denotes the relation-specific anchor subset used by the support score, $\rho_{i,n}$ selects the relation-specific support proposal family, and $w_{i,n}$ is the normalized relation relevance score

defined in Eq. 2. This formulation enables the model to infer the support of s_n through explicit spatial reasoning while allowing the learned graph representation to modulate which anatomical dependencies are most informative.

Latent Modeling of Relation Geometry. A critical challenge in zero-shot localization is that the same anatomical relation must be instantiated at different scales, thicknesses, and offsets across images. To avoid turning a graph relation into a fixed coordinate rule, we treat the geometric parameters inside each support proposal as latent random variables rather than fixed constants. Specifically, for relation type ρ, let τ_ρ denote the latent scale, thickness, or offset parameter used by ϕ_ρ, and let $P(\tau_\rho)$ be its prior distribution, such as a uniform or truncated Gaussian distribution centered at a clinically plausible reference value τ_ρ^0.

The final support estimate is approximated by averaging or sampling over this latent variability, yielding a normalized support distribution:

$$P_{\text{geo}}(Y_n|\mathcal{A},\mathcal{G}) = \int \text{Normalize}(S_{\text{geo}}(Y_n|\mathcal{A},\tau_\rho,\mathcal{G}))\, P(\tau_\rho)\, d\tau_\rho. \tag{5}$$

This posterior-guided support scoring approximates a distribution over plausible image-specific supports, enabling the model to account for uncertainty in relation geometry without committing to a fixed coordinate template. Consequently, our method uses anatomical topology as a structured support prior while leaving final mask prediction to the image-conditioned decoder. Implementation details for posterior-guided support scoring are provided in Appendix V.6.

3.4 Pseudo-Label Supervision Objective

For each unseen structure s_n, the posterior-guided geometric score yields a pseudo target $\tilde{Y}_n = (\tilde{M}_n, \tilde{\mathbf{b}}_n)$ only when sufficient available dependencies are present. This pseudo target supervises the decoder by encouraging a prompt-matched query to produce a mask and box that are consistent with the inferred anatomical support. Thus, $\tilde{Y}_n$ serves as a graph-derived support constraint rather than ground-truth supervision for the unseen target. If no valid dependency configuration is available, neither a pseudo target nor a pseudo loss is produced for s_n; the knowledge-enhanced prompt can still condition the decoder.

The total objective combines standard supervision on seen structures with posterior-guided pseudo supervision on unseen structures $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seen}} + \lambda_{\text{pseudo}}\mathcal{L}_{\text{pseudo}}$. Here $\mathcal{L}_{\text{seen}}$ denotes the standard detection and segmentation supervision for annotated structures, $\mathcal{L}_{\text{pseudo}}$ is applied only to target structures satisfying the graph dependency condition, and λ_{pseudo} controls the contribution of pseudo supervision. Detailed query assignment, losses, and weighting coefficients are provided in Appendix II.

4 Results

4.1 Datasets

CardiacUDA [26] is sourced from two medical centers and comprises 314 echocardiogram videos in apical four-chamber (A4C) view. We use this dataset for training and validation, with five annotated frames per video covering the left/right ventricles and atria. **CAMUS [23]** offers pixel-wise annotations for the left ventricle/atrium and myocardium in A4C and A2C views, encompassing 500 individuals. Both 450 training cases and 50 testing cases have annotations for two frames at the end-diastole (ED) and end-systole (ES) phases. **HMC-QU [25]** is designed for myocardial infarction analysis and includes 109 A4C frames with myocardium segmentation annotations. **EchoNet-Dynamic [22]** supports cardiac function evaluation and comprises 10,030 A4C echocardiography videos with LV segmentation annotations at ED and ES phases. **EchoNet-Pediatric [40]** supports the study of cardiac motion and chamber size in pediatric echocardiography, consisting of 7,643 videos from A4C and parasternal short-axis views with LV segmentation annotations at ED and ES phases. For abbreviation details, please see Appendix IV.

4.2 Implementation Details

We build Echo-SAM on top of SAM3 with a ViT-L/14 visual backbone and a cardiac knowledge graph containing 15 structures and 42 anatomical edges. Architecture and optimization details are

Table 1: Comparison of fully-supervised methods. All methods were trained on the full dataset. Closed-set methods segment predefined anatomical classes using fixed prediction heads, while prompt-conditioned methods produce masks according to text or fine-grained box prompts.

Method	Venue & Year	Prompt & Type	CAMUS				EchoNet-Dynamic			
			DSC ↑	mIoU ↑	HD ↓	ASSD ↓	DSC ↑	mIoU ↑	HD ↓	ASSD ↓
Closed-set Segmentation										
Vision LSTM [36]	ICLR'25	-	92.14	89.11	3.79	3.39	90.24	89.14	2.65	1.69
SimLVSeg [32]	UMB'24	-	92.54	89.71	3.65	3.12	91.91	89.08	2.65	1.65
PKEchoNet [33]	AAAI'23	-	93.49	90.95	3.42	2.93	92.60	89.89	2.53	1.48
DSA [34]	TMI'24	-	94.25	91.80	3.27	2.37	92.91	90.26	2.46	1.44
GDKVM [35]	ICCV'25	-	95.11	92.97	3.05	1.98	93.46	90.86	2.38	1.36
Prompt-conditioned Segmentation										
MedCLIP-SAM [12]	MIA'25	Text	83.50	74.28	9.54	4.39	84.29	77.82	8.53	5.12
MedROV-SAM [17]	WACV'26	Text	84.29	77.18	8.99	4.52	83.96	76.57	8.94	5.33
MedSAM [15]	Nat.Com'24	Box	85.42	75.14	8.42	3.34	86.47	79.19	7.97	4.88
SAM3 [8]	-	Text	87.85	83.92	4.26	2.79	89.61	83.09	4.70	3.98
Med-SA [14]	MIA'25	Box	88.03	78.98	7.53	2.85	87.91	78.34	6.67	4.34
MemSAM [11]	CVPR'24	Box	93.63	90.97	3.47	2.60	92.71	89.90	2.56	1.51
Echo-SAM (Our)	-	Text	**95.41**	**93.25**	**2.98**	**1.86**	**94.02**	**91.57**	**2.22**	**1.30**

Table 2: Comparison on the unseen Myo structure under the zero-shot protocol.

Method	Venue & Year	Prompt & Type	CAMUS (Myo) [23]				HMC-QU (Myo) [25]			
			DSC ↑	mIoU ↑	HD ↓	ASSD ↓	DSC ↑	mIoU ↑	HD ↓	ASSD ↓
GDKVM [35]	ICCV'25	-	0.00	0.00	∞	∞	0.00	0.00	∞	∞
MemSAM [11]	CVPR'24	Box	0.00	0.00	∞	∞	0.00	0.00	∞	∞
MedSAM [15]	Nat.Com'24	Box	16.95	9.33	172.36	41.18	6.19	3.21	236.90	54.40
SAM3 [8]	-	Text	13.51	7.62	156.77	43.82	14.46	8.90	195.72	50.06
Med-SA [14]	MIA'25	Box	6.17	3.29	69.95	22.65	1.40	0.73	62.14	23.36
MedCLIP-SAM [12]	MIA'25	Text	0.01	0.01	247.12	126.51	0.03	0.16	237.85	53.62
MedROV-SAM [17]	WACV'26	Text	4.45	3.39	194.13	92.57	2.20	1.31	147.23	72.93
Echo-SAM (Our)	-	Text	**60.24**	**51.43**	**7.65**	**3.23**	**57.09**	**45.47**	**5.94**	**1.63**

Table 3: Comparison on the unseen LV structure under the zero-shot protocol.

Method	Prompt & Type	CardiacUDA [26]				CAMUS [23]				EchoNet-Dynamic [22]				EchoNet-Pediatric [40]			
		DSC ↑	mIoU ↑	HD ↓	ASSD ↓	DSC ↑	mIoU ↑	HD ↓	ASSD ↓	DSC ↑	mIoU ↑	HD ↓	ASSD ↓	DSC ↑	mIoU ↑	HD ↓	ASSD ↓
GDKVM [35]	-	0	0	∞	∞	0	0	∞	∞	0	0	∞	∞	0	0	∞	∞
MemSAM [11]	Box	0	0	∞	∞	0	0	∞	∞	0	0	∞	∞	0	0	∞	∞
SAM3 [8]	Text	8.56	6.11	286.52	76.01	15.04	10.18	169.32	72.55	31.22	18.74	53.96	14.35	25.84	15.66	47.99	16.81
MedSAM [15]	Box	15.30	9.19	334.35	84.07	33.60	20.27	184.35	52.43	29.57	17.52	42.70	12.02	29.34	17.37	44.16	12.24
Med-SA [14]	Box	0.34	0.18	62.39	31.34	33.44	23.96	57.01	22.02	33.25	24.69	60.13	23.79	34.36	25.16	60.78	23.10
MedCLIP-SAM [12]	Text	0.00	0.00	409.63	229.75	0.01	0.01	234.16	67.28	0.08	0.05	56.91	24.36	0.05	0.03	57.90	19.93
MedROV-SAM [17]	Text	0.00	0.00	∞	∞	0.00	0.00	∞	∞	0.00	0.00	∞	∞	0.00	0.00	∞	∞
Echo-SAM (Our)	Text	**65.72**	**45.28**	**11.71**	**5.35**	**49.39**	**31.64**	**20.52**	**9.80**	**57.40**	**40.57**	**16.73**	**7.35**	**58.99**	**41.35**	**15.26**	**6.96**

provided in Appendix V.7. Since public adult echocardiography benchmarks with dense multi-structure labels are predominantly available in A4C, one of the most widely used clinical views, the main comparisons use A4C for reproducible evaluation; this is an evaluation-scope constraint rather than an architectural restriction (Appendix VI). The graph covers major cardiac structure groups, but dense quantitative evaluation is restricted by the structures and labels available in public datasets; missing labels do not establish failure on an unmeasured target. **Zero-shot Protocol.** In zero-shot experiments, models are trained on CardiacUDA using only LA, RA, and RV annotations. Unseen LV and Myo ground-truth masks are excluded from training, pseudo-label construction, checkpoint selection, and hyperparameter tuning; they are used only for offline metric computation. A detailed protocol is provided in Appendix VI. In the main zero-shot comparisons, no method is given target-specific ground-truth boxes, points, or masks, since these would reveal the unseen target location. For bbox/point prompt-dependent baselines, we use a full-image box as a label-free, non-localizing prompt; Echo-SAM instead uses knowledge-enhanced text prompts and clinical graph priors. Shared anatomy-prior and oracle-box diagnostics, anchor robustness, run-level variability, and ASD evaluation are reported only in Appendix X.

Evaluation Metrics We employ four metrics to rigorously quantify segmentation efficacy: the Dice Similarity Coefficient (DSC) and mean Intersection over Union (mIoU) for region overlap assessment,

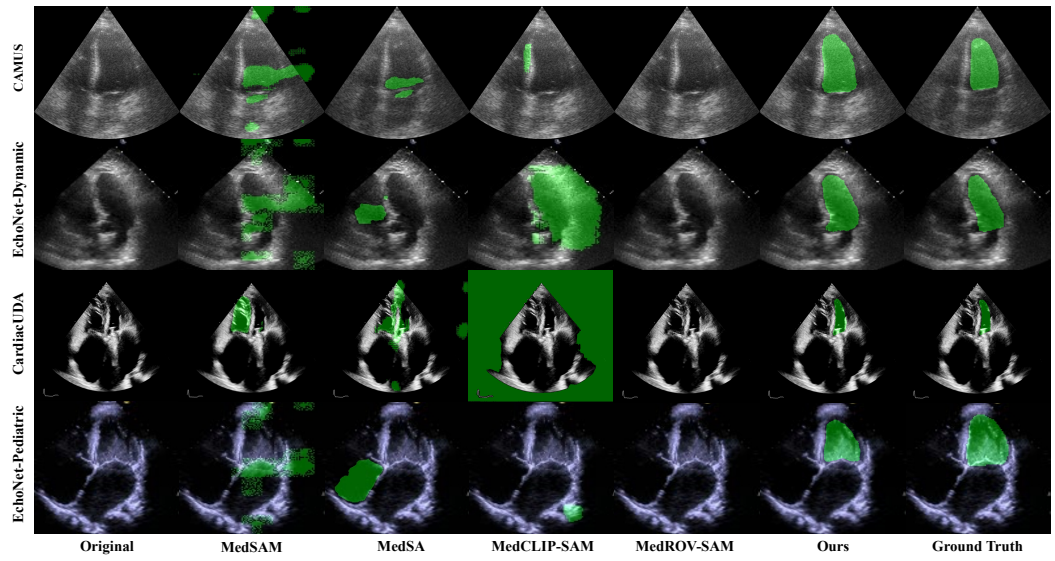


Figure 2: The segmentation results of the left ventricle across different methods and datasets.

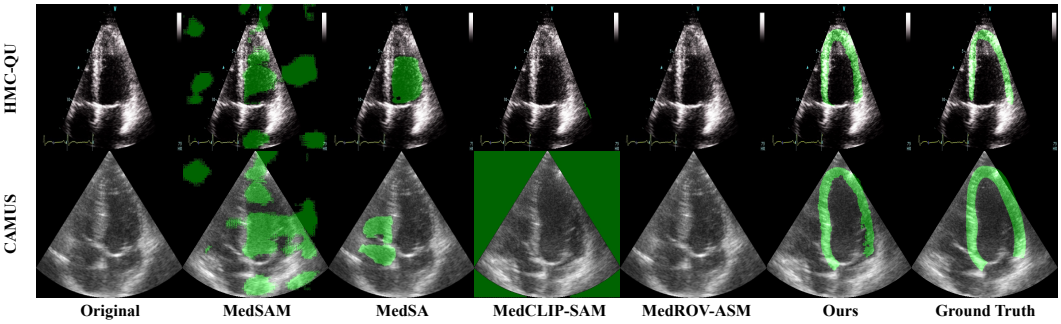


Figure 3: The segmentation results of myocardium across different methods and datasets.

alongside the 95th percentile Hausdorff Distance (HD95) and Average Symmetric Surface Distance (ASSD) for boundary fidelity analysis. For boundary metrics, all masks are evaluated on the image grid and distances are reported in pixel units with unit spacing. Comprehensive mathematical formulations and the specific instance-matching protocols are detailed in Appendix III.

4.3 Quantitative Analysis

Fully Supervised Segmentation. As shown in Table 1, Echo-SAM achieves the best performance across all overlap and boundary metrics on CAMUS and EchoNet-Dynamic. It obtains 95.41% DSC and 93.25% mIoU on CAMUS, improving over the strongest fully supervised competitor GDKVM [35] by 0.30% DSC and 0.28% mIoU while reducing HD95 and ASSD by 0.07 and 0.12 pixels. On EchoNet-Dynamic, Echo-SAM reaches 94.02% DSC and 91.57% mIoU, with gains of 0.56% DSC and 0.71% mIoU over GDKVM. The margin is more pronounced against prompt-based foundation models, where Echo-SAM exceeds SAM3 [8] by 7.56% DSC on CAMUS and 4.41% DSC on EchoNet-Dynamic. These consistent gains indicate that graph-enhanced prompts provide useful anatomical context even under full supervision, constraining ambiguous ultrasound appearance through neighboring cardiac structures and improving both region overlap and boundary localization.

Zero-Shot Segmentation on Unseen Structures. Tables 2 and 3 evaluate the setting where models are trained only with LA, RA, and RV annotations from CardiacUDA, while Myo and LV are never annotated during training. For unseen Myo, Echo-SAM achieves 60.24% DSC on CAMUS and 57.09% DSC on HMC-QU, exceeding the best competing methods by 43.29% and 42.63%, respectively. The boundary metrics show the same pattern: HD95 is reduced to 7.65 and 5.94 pixels,

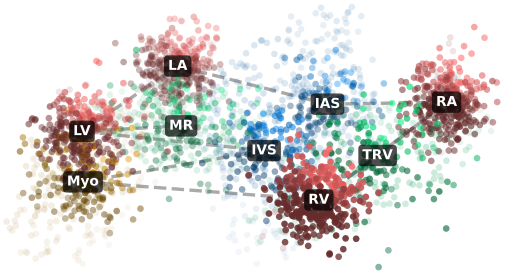


Figure 4: The visualization of the learned Medical Knowledge Graph in apical four-chamber view, which includes chambers (red), valves (green), muscle (blue), and myocardium (orange) structures.

Table 4: Ablation of different modules. Experiments are conducted on CAMUS and HMC-QU datasets (Dice %). Graph. and Geo. denote the knowledge graph and geometric support inference modules, respectively.

Graph.	Geo.	CAMUS (Myo)	CAMUS (LV)	HMC-QU (Myo)
✗	✗	13.51	15.04	14.46
✓	✗	39.74	32.93	35.98
✗	✓	47.35	40.23	46.41
✓	✓	**60.24**	**49.39**	**57.09**

Table 5: Ablation of graph fusion strategies, experiment on CAMUS and HMC-QU (Dice%).

Visual	Text	CAMUS (Myo)	CAMUS (LV)	HMC-QU (Myo)
✗	✗	47.35	40.23	46.41
✗	✓	52.20	41.68	50.05
✓	✗	57.66	46.97	54.18
✓	✓	**60.24**	**49.39**	**57.09**

whereas the best finite baseline HD95 remains 69.95 pixels on CAMUS and 62.14 pixels on HMC-QU. This result is clinically meaningful because Myo is a thin annular structure with weak ultrasound contrast; accurate segmentation requires localizing the myocardial band relative to the ventricular cavity rather than matching texture alone.

For unseen LV, Echo-SAM obtains 65.72%, 49.39%, 57.40%, and 58.99% DSC on CardiacUDA, CAMUS, EchoNet-Dynamic, and EchoNet-Pediatric, with absolute gains of 50.42%, 15.79%, 24.15%, and 24.63% over the strongest baseline on each dataset. These improvements indicate that the model transfers beyond the training domain and remains effective across adult, pediatric, and cross-institutional data. Existing foundation models degrade sharply because the zero-shot protocol removes oracle location prompts: box-based methods receive only weak full-image cues, and text-based methods cannot infer precise cardiac positions from semantic labels alone. In contrast, Echo-SAM treats localized LA/RA/RV structures as anatomical anchors, propagates their spatial constraints through the cardiac knowledge graph, and converts these relations into geometry-aware support estimates. Because support candidates are scored around image-specific anchors rather than predefined coordinates, the selected pseudo targets can localize structures that were treated as background during training.

4.4 Qualitative Analysis

Figures 2 and 3 present qualitative comparisons of segmentation results for the LV and Myo across different methods and datasets. As illustrated in Figure 2, existing approaches often produce incomplete or misaligned LV masks under the zero-shot protocol. In contrast, Echo-SAM localizes the LV without annotated LV training data, indicating stronger generalization to unseen anatomical regions. The segmentation of myocardium presents a greater challenge due to its thin, annular morphology surrounding the LV. As shown in Figure 3, competing methods often miss parts of the myocardial boundary, whereas our approach produces masks that more consistently envelop the LV structure. These cross-method and cross-dataset visual results support the quantitative findings and suggest that integrating anatomical knowledge graphs with geometric support inference is useful for zero-shot segmentation of complex, fine-grained structures.

4.5 Ablation Study

Ablation of Modules. As shown in Table 4, disabling all modules reduces Echo-SAM to the original SAM3 with Dice scores of 13.51, 15.04, and 14.46 on CAMUS (Myo, LV) and HMC-QU (Myo), respectively. The full model achieves optimal performance with improvements of 46.73, 34.35, and 42.63 on these benchmarks. As illustrated in Figure 4, our Medical Knowledge Graph captures anatomical topology across different cardiac structures, enabling the model to reason over spatial relationships and morphological characteristics. The geometric support inference component alone yields Dice scores of 47.35, 40.23, and 46.41 on CAMUS (Myo, LV) and HMC-QU (Myo), demonstrating that explicit topological constraints provide strong spatial support for localizing unseen

structures while adapting relation-specific proposals to the observed anchors. Adding graph-grounded prompt fusion on top of the geometric support inference module further improves Dice by 12.89, 9.16, and 10.68 on the same benchmarks, respectively. This gain indicates that the GNN contribution is not limited to accompanying the geometric support score: relation-conditioned node and edge embeddings help the decoder select target-specific mask hypotheses and weight the anatomical dependencies used for support inference. Conversely, the graph-only variant improves over vanilla SAM3 but remains below the geometry-only variant, suggesting that neural graph grounding and explicit geometric support inference provide complementary benefits.

Effect of the Medical Knowledge Graph Update Strategies. Table 5 evaluates the impact of different modalities for updating the Medical Knowledge Graph, with the geometric support inference module fixed across all settings. The baseline without graph updates achieves a Dice score of 47.35 on CAMUS (Myo). Incorporating textual knowledge alone improves this to 52.20, demonstrating that semantic descriptions provide valuable context. However, leveraging visual features for graph propagation yields a larger gain (57.66), suggesting that image-derived embeddings capture finer anatomical details crucial for segmentation. Crucially, employing both visual and textual updates achieves the best performance of 60.24 on CAMUS and 57.09 on HMC-QU. This result highlights that text offers high-level semantic constraints, and visual features provide image-specific adaptation. The GNN fusion effectively integrates these heterogeneous sources, refining node representations to bridge the gap between abstract anatomical knowledge and image-conditioned support inference, thereby maximizing the zero-shot segmentation capability.

5 Conclusion

In this work, we introduced Echo-SAM, a framework for zero-shot echocardiographic segmentation that integrates explicit anatomical topology with geometry-aware support inference. Our approach constructs a cardiac knowledge graph where structural relationships are encoded as learnable edge embeddings. Through relation-conditioned graph grounding, the model bridges anatomical priors and image-grounded visual features while producing both enhanced prompts and relation weights for unseen target inference. The topology-to-geometry module models relation-specific scale, thickness, and offset as latent variables, adapting support estimates to image-specific anchors.

The shared relation-based formulation covers the major cardiac structure groups and supports extension to additional cardiac targets through valid, observable anatomical dependencies. Experiments on echocardiographic benchmarks demonstrate the effectiveness of knowledge-driven geometric support inference for zero-shot segmentation. Future work will explore temporal anatomical reasoning and pathology-conditioned relations, and extend evaluation across acquisition views and cardiac structures (Appendix V.5).

Acknowledgment

The work described in this paper was supported by grants from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (Project Nos. R6005-24, N_HKUST654/24, and C6088-25Y). It was also partially supported by RGC research grants under Project Nos. AoE/E-601/24-N and T45-401/22N and a research grant from the Innovation and Technology Commission (ITC) (Project No. GHP/124/22).

Competing Interests

All authors declare no competing interests.

References

[1] Carol Mitchell, Peter S Rahko, Lori A Blauwet, Barry Canaday, Joshua A Finstuen, Michael C Foster, Kenneth Horton, Kofo O Ogunyankin, Richard A Palma, and Eric J Velazquez. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 32(1):1–64, 2019.

[2] Julie De Backer, Kristina H Haugaa, Nina Eide Hasselberg, Michèle De Hosson, Margarita Brida, Silvia Castelletti, Matthew Cauldwell, Elisabetta Cerbai, Lia Crotti, Natasja MS de Groot, et al. 2025 esc guidelines for the management of cardiovascular disease and pregnancy: Developed by the task force on the management of cardiovascular disease and pregnancy of the european society of cardiology (esc) endorsed by the european society of gynecology (esg). *European heart journal*, 46(43):4462–4568, 2025.

[3] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11592, 2022.

[5] Xinyu Liu, Beiwen Tian, Zhen Wang, Rui Wang, Kehua Sheng, Bo Zhang, Hao Zhao, and Guyue Zhou. Delving into shape-aware zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2999–3009, 2023.

[6] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2593–2602, 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[8] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[9] 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.

[10] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[11] 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.

[12] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Analysis*, page 103749, 2025.

[13] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 643–653. Springer, 2024.

[14] 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.

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

[16] Anwai Archit, Luca Freckmann, and Constantin Pape. Medicosam: Robust improvement of sam for medical imaging. *IEEE Transactions on Medical Imaging*, 2025.

[17] Tooba Tehreem Sheikh, Jean Lahoud, Rao Muhammad Anwer, Fahad Shahbaz Khan, Salman Khan, and Hisham Cholakkal. Medrov: Towards real-time open-vocabulary detection across diverse medical imaging modalities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8628–8638, 2026.

[18] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14766–14776, 2025.

[19] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023.

[20] Shuchang Ye, Usman Naseem, Mingyuan Meng, and Jinman Kim. Alleviating textual reliance in medical language-guided segmentation via prototype-driven semantic approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22316–22326, 2025.

[21] Qingtao Pan, Zhengrong Li, Guang Yang, Qing Yang, and Bing Ji. Evivlm: When evidential learning meets vision language model for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

[22] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[23] 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.

[24] Jiewen Yang, Yiqun Lin, Bin Pu, and Xiaomeng Li. Bidirectional recurrence for cardiac motion tracking with gaussian process latent coding. *Advances in Neural Information Processing Systems*, 37:34800–34823, 2024.

[25] Aysen Degerli, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. Early myocardial infarction detection over multi-view echocardiography. *Biomedical Signal Processing and Control*, 87:105448, 2024.

[26] Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, and Xiaomeng Li. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11878–11887, 2023.

[27] 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.

[28] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[29] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14043–14053. IEEE, 2022.

[30] Nathan Painchaud, Nicolas Duchateau, Olivier Bernard, and Pierre-Marc Jodoin. Echocardiography segmentation with enforced temporal consistency. *IEEE transactions on medical imaging*, 41(10):2867–2878, 2022.

[31] Ziyang Zheng, Jiewen Yang, Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. GL-Fusion: Global-local fusion network for multi-view echocardiogram video segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 78–88. Springer Nature Switzerland, 2023.

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

[33] Huisi Wu, Jingyin Lin, Wende Xie, and Jing Qin. Super-efficient echocardiography video segmentation via proxy-and kernel-based semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2803–2811, 2023.

[34] Jingyin Lin, Wende Xie, Li Kang, and Huisi Wu. Dynamic-guided spatiotemporal attention for echocardiography video segmentation. *IEEE Transactions on Medical Imaging*, 43(11):3843–3855, 2024.

[35] Rui Wang, Yimu Sun, Jingxing Guo, Huisi Wu, and Jing Qin. Gdkvm: Echocardiography video segmentation via spatiotemporal key-value memory with gated delta rule. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12191–12200, 2025.

[36] Benedikt Alkin, Maximilian Beck, Korbinian Pöppel, Sepp Hochreiter, and Johannes Brandstetter. Vision-LSTM: xLSTM as generic vision backbone. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer, 2021.

[38] Bin Pu, Xingguo Lv, Jiewen Yang, Guannan He, Xingbo Dong, Yiqun Lin, Shengli Li, Ying Tan, Fei Liu, Ming Chen, Zhe Jin, Kenli Li, and Xiaomeng Li. Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41204–41220. PMLR, 2024.

[39] Bin Pu, Jiewen Yang, Xingguo Lv, Xingbo Dong, Lei Zhao, Shengli Li, Kenli Li, and Xiaomeng Li. ToMo-UDA++: Unsupervised domain adaptation for anatomical structure detection using enhanced topology and morphology knowledge. *International Journal of Computer Vision*, 134(5):230, 2026.

[40] Charitha D Reddy, Leo Lopez, David Ouyang, James Y Zou, and Bryan He. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *Journal of the American Society of Echocardiography*, 36(5):482–489, 2023.

[41] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[42] Jiewen Yang, Yiqun Lin, Bin Pu, Jiarong Guo, Xiaowei Xu, and Xiaomeng Li. CardiacNet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In *Computer Vision – ECCV 2024*, pages 293–311. Springer Nature Switzerland, 2024.

Technical Appendices and Supplementary Material

I Broader Impact

The development of **Echo-SAM** offers a possible direction for future medical image segmentation. By demonstrating effective zero-shot segmentation based on structured anatomical reasoning and anchor-conditioned support inference, this work encourages a shift from data-heavy supervised learning to knowledge-driven frameworks. Structural priors and geometric support inference provide a way to reduce dependence on target annotations.

Specifically, Echo-SAM moves beyond static text-based alignment to anchor-conditioned geometric support inference. Clinically, this provides a route toward segmenting sparsely annotated cardiac structures without requiring target-specific retraining. This capability addresses the challenge of data scarcity and supports the application of AI diagnostics in settings with limited resources.

Beyond adult echocardiography, our method provides a potential framework for ultrasound scenarios where anatomical layout is constrained but dense labels remain scarce. For example, fetal ultrasound, carotid ultrasound, and abdominal ultrasound all contain recurring spatial relations among visible landmarks and target structures. By constructing task-specific ultrasound knowledge graphs and validating their spatial relations with clinical expertise, the same principle could support label-efficient segmentation in these settings. Modeling relation-specific geometry as a distribution supports adaptation of anatomical priors across subjects, developmental stages, and acquisition views. Potential clinical harms include false-positive support for absent structures, missed boundaries under poor image quality, and overconfidence under acquisition or population shifts. Clinical deployment requires external validation, target-presence calibration, and expert oversight.

II Loss Configuration

This appendix provides a comprehensive elaboration on the loss functions and optimization strategies employed in our framework. While the main text outlines the overall objective, this section details the specific mathematical formulations, hyperparameter settings, and balancing mechanisms for model convergence. We explicitly define the composite detection loss, which integrates focal-modulated classification, object presence prediction, and hybrid bounding box regression (L1 and GIoU). Furthermore, we describe the segmentation objective that combines per-pixel focal loss with soft Dice loss to address severe foreground–background imbalance. Finally, we specify the cost matrix and weighting coefficients used in the Hungarian matching algorithm to establish optimal one-to-one correspondence between predictions and ground-truth instances. A summary of all key hyperparameters is provided in Table I.

II.1 Detection Loss Details

The detection objective adheres to the DETR framework [41] and is composed of three distinct terms: classification, bounding box regression, and object presence prediction.

Classification Loss. To mitigate the severe class imbalance typical in detection tasks, we adopt a focal-modulated binary cross-entropy formulation:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \cdot w_{\text{pos}} \cdot \log(p_i) + (1-y_i)\cdot p_i^{\gamma}\cdot \log(1-p_i)\Big], \tag{6}$$

where $p_i = \sigma(z_i)$ denotes the sigmoid-transformed logits, $y_i \in \{0,1\}$ is the binary ground-truth label, $w_{\text{pos}} = 10.0$ amplifies the contribution of positive instances, and $\gamma = 2$ emphasizes difficult negative examples during training.

Presence Loss. A dedicated scalar head predicts whether any object of interest appears in the image. Its supervision signal is derived from the sigmoid focal loss:

$$\mathcal{L}_{\text{presence}} = -\alpha(1-p_t)^{\gamma}\log(p_t), \tag{7}$$

with $p_t = p$ if the image contains at least one object ($y = 1$), and $p_t = 1 - p$ otherwise. The hyperparameters are set to $\alpha = 0.5$ and $\gamma = 2$ to balance positive/negative contributions and focus on challenging cases.

Table I: Loss hyperparameters for SAM3 training.

Component	Parameter	Value
Classification Loss	Loss weight λ_{ce}	20.0
	Presence loss weight $\lambda_{\text{presence}}$	20.0
	Positive weight w_{pos}	10.0
	Focal α	0.25
	Focal γ	2
	Number of queries	200
Bounding Box Loss	L1 weight λ_{bbox}	5.0
	GIoU weight λ_{giou}	2.0
Segmentation Loss	Mask weight λ_{mask}	200.0
	Dice weight λ_{dice}	10.0
	Focal α	0.25
	Focal γ	2.0
Pseudo-Label Loss	Weight λ_{pseudo}	0.1
Hungarian Matcher	Classification cost λ_{cls}	2.0
	BBox cost λ_{bbox}	5.0
	GIoU cost λ_{giou}	2.0

Bounding Box Regression. Localization accuracy is enforced through a hybrid regression objective combining L1 distance and Generalized Intersection over Union (GIoU):

$$\mathcal{L}_{\text{bbox}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{b}_i - \mathbf{b}_i^*\|_1, \qquad \mathcal{L}_{\text{giou}} = \frac{1}{N}\sum_{i=1}^{N} \big(1 - \text{GIoU}(\mathbf{b}_i, \mathbf{b}_i^*)\big). \tag{8}$$

Here, $\text{GIoU}(\mathbf{b}, \mathbf{b}^*)$ is defined as

$$\text{GIoU}(\mathbf{b}, \mathbf{b}^*) = \text{IoU}(\mathbf{b}, \mathbf{b}^*) - \frac{|C \setminus (\mathbf{b} \cup \mathbf{b}^*)|}{|C|}, \tag{9}$$

where C represents the smallest axis-aligned bounding box that fully encloses both $\mathbf{b}$ and $\mathbf{b}^*$.

II.2 Segmentation Loss Details

For mask prediction, we combine focal loss and Dice loss to effectively address pixel-level foreground–background imbalance while promoting spatial coherence.

Mask Focal Loss. A per-pixel sigmoid focal loss is applied to each location (i, j) in the output mask:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{HW}\sum_{i,j}\Big[\alpha_{\text{mask}} M_{ij}(1-p_{ij})^{\gamma_{\text{mask}}}\log(p_{ij}) + (1-\alpha_{\text{mask}})(1-M_{ij})p_{ij}^{\gamma_{\text{mask}}}\log(1-p_{ij})\Big], \tag{10}$$

where $p_{ij} \in [0, 1]$ is the predicted probability, $M_{ij} \in \{0, 1\}$ is the ground-truth mask value, and we use $\alpha_{\text{mask}} = 0.25$, $\gamma_{\text{mask}} = 2$ to down-weight well-classified easy examples.

Dice Loss. To directly optimize region overlap, we employ the soft Dice loss:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i,j}\hat{M}_{ij}M_{ij} + 1}{\sum_{i,j}\hat{M}_{ij} + \sum_{i,j}M_{ij} + 1}, \tag{11}$$

where $\hat{M}_{ij} = \sigma(\cdot)$ denotes the sigmoid-normalized mask logits. The constant term $+1$ in numerator and denominator ensures numerical stability when both masks are empty.

II.3 Knowledge-Guided Pseudo-Label Objective

The pseudo-label objective is applied only to targets in the fixed set $\mathcal{S}_{\text{unseen}}$ that satisfy the dependency condition defined in the main text. The objective below is the per-target contribution, summed over eligible targets; a target without a valid dependency configuration contributes zero. A full-image fallback prompt is not a pseudo label and produces no pseudo loss. Let $\tilde{Y}_n = (\tilde{M}_n, \tilde{\mathbf{b}}_n)$ denote the anatomical pseudo target selected from the posterior-guided support approximation described in Appendix V.6. The decoder produces a set of query predictions $\{(\hat{M}_q, \hat{\mathbf{b}}_q, \ell_q)\}_{q=1}^{Q}$ for the same target prompt. We select the query with the largest compatibility score,

$$q_n^* = \arg\max_q \; C(q \mid \mathcal{I}, \tilde{Y}_n, \mathbf{p}_n^{\text{enhanced}}) \approx \arg\max_q \; \sigma(\ell_q)\, p_{\text{presence}}, \tag{12}$$

with a box-consistency fallback when confidence scores are unavailable. This query selection can be interpreted as a highest-scoring assignment between the latent anatomical target and the finite decoder hypotheses.

The selected query is then supervised by a mask, box, and confidence objective:

$$\begin{aligned}\mathcal{L}_{\text{pseudo}} = \lambda_m \Big[\mathcal{L}_{\text{BCE}}(\hat{M}_{q_n^*}, \tilde{M}_n) + \mathcal{L}_{\text{Dice}}(\hat{M}_{q_n^*}, \tilde{M}_n)\Big] \\ + \lambda_b \|\hat{\mathbf{b}}_{q_n^*} - \tilde{\mathbf{b}}_n\|_1 + \lambda_g \mathcal{L}_{\text{GIoU}}(\hat{\mathbf{b}}_{q_n^*}, \tilde{\mathbf{b}}_n) + \lambda_c \mathcal{L}_{\text{conf}}.\end{aligned} \tag{13}$$

For the confidence term, the positive query receives a soft target proportional to its spatial agreement with the proposed anatomical support,

$$y_{q_n^*}^{\text{soft}} = \text{clip}\left(\sigma(\ell_{q_n^*})^{1/4}\, \text{IoU}(\hat{\mathbf{b}}_{q_n^*}, \tilde{\mathbf{b}}_n)^{3/4}, \epsilon, 1\right), \tag{14}$$

while unmatched queries are treated as negatives with focal reweighting. The pseudo target provides a graph-derived support constraint for the image-conditioned decoder query.

II.4 Hungarian Matching

Consistent with the DETR methodology, we establish a one-to-one correspondence between predictions and ground-truth objects via bipartite matching. The total assignment cost between prediction i and target j is given by:

$$\mathcal{C}_{ij} = \lambda_{\text{cls}} \mathcal{C}_{\text{cls}}(i,j) + \lambda_{\text{bbox}} \|\mathbf{b}_i - \mathbf{b}_j^*\|_1 + \lambda_{\text{giou}} \big(1 - \text{GIoU}(\mathbf{b}_i, \mathbf{b}_j^*)\big), \tag{15}$$

where $\lambda_{\text{cls}} = 2.0$, $\lambda_{\text{bbox}} = 5.0$, and $\lambda_{\text{giou}} = 2.0$. The classification cost $\mathcal{C}_{\text{cls}}(i,j)$ is computed using the same focal-weighted cross-entropy as in the classification loss. The optimal permutation $\hat{\sigma}$ that minimizes the total matching cost, $\hat{\sigma} = \arg\min_\sigma \sum_i \mathcal{C}_{i,\sigma(i)}$, is efficiently obtained using the Hungarian algorithm.

III Segmentation Evaluation Metrics

To assess segmentation performance, we use four complementary metrics that quantify region overlap and boundary accuracy. Predictions are matched to ground-truth instances by selecting the candidate with the highest Intersection over Union (IoU) for each image-category pair. If no prediction can be matched to a ground-truth instance, the instance is assigned zero for overlap-based metrics (DSC and mIoU) and infinite error for boundary-based metrics (HD95 and ASSD). Boundary distances are computed on the image grid and reported in pixel units under a unit-spacing convention. Final scores are computed at the instance level and aggregated across the dataset, with detailed per-category results provided separately.

III.1 Dice Similarity Coefficient (DSC)

The DSC quantifies the spatial agreement between the predicted mask $\hat{M}$ and the reference mask M by measuring their region overlap:

$$\text{DSC}(M, \hat{M}) = \frac{2|M \cap \hat{M}|}{|M| + |\hat{M}|}. \tag{16}$$

Scaled between 0 (no overlap) and 1 (perfect alignment), this metric is computationally derived from the IoU using the transformation $\text{DSC} = 2\,\text{IoU}/(1 + \text{IoU})$, ensuring mathematical equivalence for binary segmentations while optimizing calculation efficiency.

III.2 Mean Intersection over Union (mIoU)

IoU serves as the fundamental measure of region similarity, defined as the ratio of the intersecting area to the union of the predicted and ground-truth regions:

$$\text{IoU}(M, \hat{M}) = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|}. \tag{17}$$

The mIoU represents the arithmetic mean of IoU scores calculated across all successfully matched ground-truth instances within every category, providing a robust indicator of overall segmentation accuracy.

III.3 95th Percentile Hausdorff Distance (HD95)

While standard Hausdorff Distance captures the maximum surface deviation, it is often overly sensitive to outliers. To address this, we utilize HD95, which evaluates the 95th percentile of the minimum Euclidean distances between contour points. Let ∂M and $\partial \hat{M}$ denote the boundary point sets of the ground truth and prediction, respectively. The symmetric HD95 is formulated as:

$$\text{HD95}(M, \hat{M}) = \max\Big(\text{P}_{95}\Big(\{d(p, \partial \hat{M}) \mid p \in \partial M\}\Big),\ \text{P}_{95}\Big(\{d(q, \partial M) \mid q \in \partial \hat{M}\}\Big)\Big), \tag{18}$$

where $d(x, S) = \min_{y \in S} \|x - y\|_2$ and P_{95} denotes the 95-th percentile operator. Under the unit-spacing convention on the image grid, lower HD95 values indicate smaller high-percentile boundary deviation while reducing sensitivity to isolated outlier pixels.

III.4 Average Symmetric Surface Distance (ASSD)

Complementing HD95, the ASSD provides a global assessment of boundary accuracy by averaging the bidirectional surface distances. It is defined as:

$$\text{ASSD}(M, \hat{M}) = \frac{1}{2}\left(\frac{1}{|\partial M|}\sum_{p \in \partial M} d(p, \partial \hat{M}) + \frac{1}{|\partial \hat{M}|}\sum_{q \in \partial \hat{M}} d(q, \partial M)\right). \tag{19}$$

Unlike the percentile-based HD95, ASSD incorporates contributions from all boundary points, measuring the average deviation between the predicted and reference contours. As with HD95, distances follow the unit-spacing convention on the image grid, and lower values indicate closer average boundary alignment.

IV Abbreviation

Below is the correspondence between abbreviations, full names, and descriptions commonly used in cardiac ultrasound (echocardiography).

V Knowledge-Grounded Anatomical Support Priors

This section formalizes the view-level clinical constraints that define the structured anatomical support prior used by the cardiac knowledge graph and geometric support inference module.

V.1 Knowledge Graph Construction Details

Anatomical Sources. We construct the medical knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from standardized cardiac anatomy references and echocardiographic view guidelines [1]. The graph is then reviewed by domain experts for anatomical validity. Its ontology spans cardiac views; each view activates only the structures and relations observable in that acquisition plane, as formalized in Appendix V.4.

Node Definition. Each node $v_i \in \mathcal{V}$ represents a distinct cardiac concept relevant to echocardiographic analysis. We include chambers (LV, RV, LA, RA), myocardial structures (myocardium, interventricular septum, interatrial septum, posterior wall), valves (mitral, tricuspid, aortic, and pulmonary valves), and large vessels (ascending aorta, descending aorta, and pulmonary artery). Each node is associated with two forms of knowledge: a morphology description, which specifies appearance or tissue properties, and a relation description, which specifies anatomical connectivity or spatial adjacency.

Table II: Correspondence between Abbreviations and Full Terms

Abbreviation	Full Term	Context / Notes
A4C	Apical Four-Chamber view	Standard echocardiographic view
A2C	Apical Two-Chamber view	Additional apical view
PLAX	Parasternal Long-Axis view	Long-axis acquisition plane
PSAX	Parasternal Short-Axis view	Short-axis acquisition plane
Myo	Myocardium	Myocardial wall
ASD	Atrial Septal Defect	Congenital septal abnormality
ED	End-Diastole / End-Diastolic	Phase of the cardiac cycle (e.g., LVEDV)
ES	End-Systole / End-Systolic	Phase of the cardiac cycle (e.g., LVESV)
LV	Left Ventricle	Cardiac chamber
LA	Left Atrium	Cardiac chamber
RV	Right Ventricle	Cardiac chamber
RA	Right Atrium	Cardiac chamber
MR	Mitral Valve	Anatomical structure
TR	Tricuspid Valve	Anatomical structure
IAS	Interatrial Septum	Anatomical structure separating atria
IVS	Interventricular Septum	Anatomical structure separating ventricles

ED/ES: Often used as prefixes or suffixes (e.g., LV_{ED} for Left Ventricle at End-Diastole).
A4C: Essential view for assessing all four chambers and the mitral/tricuspid valves.
IAS/IVS: Critical for detecting septal defects (ASD/VSD).

Edge Definition. Edges encode typed anatomical relations drawn from a finite relation set $\mathcal{R}$, including tissue continuity, annular enclosure, separation by a membrane, and adjacency through a valve. These relations are used in two ways. First, they define the neighborhood over which concept embeddings are propagated. Second, they determine which geometric support proposal family is valid for an unseen target and contribute to the relation relevance weights used in geometric support inference. For prompt fusion, the relation channel corresponds to the selected edge embedding in the main text; for geometric inference, it corresponds to the active dependency relation used to instantiate and weight the target support proposal. The graph is symmetrized for relation embedding updates, so an edge such as LV–LA contributes the same relational evidence regardless of the prompt direction.

Anatomical Prior Specification. The graph encodes clinical anatomical relations: LV is connected to LA through the mitral valve, RV and LV are separated by the interventricular septum, and myocardium forms an annular wall around the ventricular cavity. The graph and stochastic support ranges are specified before model training from anatomical references, independently of target annotations. The supervision and evaluation protocol is detailed in Appendix VI.

V.2 Anchor-Conditioned Anatomical Support Approximation

Let $\mathcal{A}_0 = \{(\hat{M}_i, \hat{\mathbf{b}}_i)\}_{i \in \mathcal{S}_{\text{seen}}}$ denote the visible anchor structures localized by the model in an image. For an unseen target n, let $\mathcal{S}_{\text{avail}}(n)$ contain the categories of localized seen anchors and any model-inferred intermediate supports available before estimating n. An inferred intermediate category remains in $\mathcal{S}_{\text{unseen}}$; support availability does not alter the seen/unseen partition. We write $\mathcal{A}$ for the corresponding available support set, initialized by $\mathcal{A}_0$ and augmented only with model-inferred intermediate supports. The knowledge graph selects a dependency set $\mathcal{D}(n) \subseteq \mathcal{S}_{\text{avail}}(n)$ and valid relation types $\rho \in \mathcal{R}$. This subsection instantiates the main-text notation ϕ_ρ and τ_ρ at the implementation level: ϕ_ρ is realized as a relation-specific support proposal score, and τ_ρ is represented by latent geometric variables

$$\theta_n = (\rho_n, \gamma_n, \eta_n), \tag{20}$$

where ρ_n is the active anatomical relation, γ_n parameterizes relation-specific scale or extent, and η_n captures residual image-specific translation or shape variation. We use the following factorization to organize the relation choice and geometric perturbations:

$$P(\theta_n \mid \mathcal{G}) = P(\rho_n \mid \mathcal{G})P(\gamma_n \mid \rho_n)P(\eta_n \mid \rho_n). \tag{21}$$

In implementation, $P(\rho_n \mid \mathcal{G})$ is parameterized by the normalized relation relevance weights in Eq. 2. Thus, learnable edge embeddings determine the mixture over valid dependency relations in addition to contributing to prompt grounding. Conditioned on θ_n and the anchor structures, the support proposal is obtained by

$$(\bar{M}_n, \bar{\mathbf{b}}_n) = f_{\rho_n}(\mathcal{A}_{\rho_n}, \gamma_n, \eta_n), \tag{22}$$

where $\mathcal{A}_{\rho_n} \subseteq \mathcal{A}$ is the relation-specific anchor subset and f_{ρ_n} denotes a relation-conditioned support operator. The resulting $\bar{M}_n$ and $\bar{\mathbf{b}}_n$ are candidate supports used to instantiate pseudo targets. Uncertainty enters through stochastic perturbations of γ_n and η_n, and learning gradients are applied through the image-conditioned decoder losses.

V.3 Relation-Specific Geometric Support Proposals

The cardiac graph activates shared relation-conditioned support proposal families. Each family maps available anchor supports, a relation type, and latent geometric variables to a set of candidate target supports. The operators are defined in normalized image coordinates and are weighted by the learned relation relevance scores in Eq. 2. The final pseudo target is selected using the image-specific anchor configuration and decoder evidence.

Completion and Continuity Proposals. For chamber-like or tissue-continuity relations, the proposal operator estimates a candidate support from the relative configuration of available neighboring structures. This family captures that an unseen structure may be spatially implied by multiple visible anchors. Missing or unreliable anchors reduce the proposal confidence and trigger weaker, less localized supports.

Enclosure and Surrounding Proposals. For relations such as surrounding tissue or wall-like structures, the proposal operator encourages candidate supports to remain spatially coupled to an inferred central structure while allowing the latent thickness and extent to vary.

Separation and Junction Proposals. For membranes, septa, and valve-like structures, the proposal operator favors supports near compatible interfaces or junctions between related structures. The exact support width, orientation, and placement are treated as latent relation geometry and are resolved through posterior-guided scoring with decoder evidence. If one required boundary structure is itself unseen, its model-inferred support proposal is used.

V.4 View-Conditioned Generalization Beyond A4C

The A4C setting is one view-conditioned instantiation of the proposed topology-to-geometry framework. For an echocardiographic view v, we define a view-conditioned graph

$$\mathcal{G}^{(v)} = (\mathcal{V}^{(v)}, \mathcal{E}^{(v)}), \qquad \mathcal{E}^{(v)} \subseteq \mathcal{V}^{(v)} \times \mathcal{R} \times \mathcal{V}^{(v)}, \tag{23}$$

where $\mathcal{V}^{(v)}$ contains the structures visible or clinically relevant in view v, and $\mathcal{E}^{(v)}$ selects valid anatomical relations from the shared relation vocabulary $\mathcal{R}$. The relation families and support proposal operators $\{f_r\}_{r\in\mathcal{R}}$ are shared across views; only the visible anchors, valid dependency edges, and clinically meaningful target set change with the acquisition plane. Therefore, extending from A4C to another view requires re-instantiating the graph with view-specific anchors.

The geometric support proposals are anchor-conditioned. Let T denote a 2D similarity transform applied consistently to all anchor masks and boxes in $\mathcal{A}$, and let η_T denote the corresponding transformed residual offset. For an operator defined entirely by anchor-relative distances and directions, with its geometric parameters transformed consistently, the ideal continuous support satisfies

$$f_r(T(\mathcal{A}), \gamma, \eta_T) = T(f_r(\mathcal{A}, \gamma, \eta)). \tag{24}$$

Relative displacement vectors transform with the rotation and scale of T, relative lengths scale consistently, and dimensionless thickness ratios are unchanged; constructing the support from these

Table III: Examples of view-conditioned graph instantiations for echocardiographic segmentation. The table describes theoretical instantiations of the shared relation families. For each target, its own annotation is excluded from the available supports.

View	Anchors or intermediate supports	Example targets	Active relation families
A4C	visible chamber anchors and intermediate supports	LV, Myo, IAS, IVS, mitral valve	chamber completion; annular surround; membrane separation; valve junction
A2C	LA, LV	LV, Myo, mitral valve	chamber continuity; annular surround; valve junction
Parasternal long-axis	LA, LV, LVOT and aorta	Myo, IVS, posterior wall, mitral and aortic valves	wall adjacency; septal boundary; valve junction
Parasternal short-axis	LV cavity, RV crescent	Myo, IVS	annular surround; LV–RV separation

quantities therefore commutes with T. This argument applies to the continuous support geometry, with discretization and axis-aligned box extraction introducing approximation. It does not imply invariance to a different 3D acquisition plane, anatomical truncation, or a topology change. Those cases require view-appropriate visible anchors and valid graph relations; appearance shifts may additionally require adaptation.

V.5 Anatomical Scope and Conditions for Generalization

The framework operates on shared relation types and image-specific supports. The current ontology covers most major cardiac structure groups through chambers, myocardial walls, septa, valves, and great vessels. In principle, any cardiac structure with observable anatomical dependencies can be incorporated by specifying its concept, valid relations, and compatible geometric operators. The anatomical relations determine the representable structure set, while available images and reference masks determine the set of structures that can be quantitatively evaluated.

Latent geometry accommodates variation in scale, thickness, and position relative to the observed anchors. Absent or discontinuous tissue can violate an expected anatomical relation and lead to support proposals for missing structures. Detection confidence reflects both target presence and image quality, motivating target-presence calibration.

The same reasoning applies to other anatomies by defining clinically valid nodes and relations, choosing visible anchors with feasible annotations, and instantiating compatible geometric operators. Geometric priors are less informative for diffuse or variably located lesions, severely truncated images, or spatial relations that cannot be resolved from a single 2D plane. Temporal reasoning, pathology-conditioned relations, and confidence calibration are natural extensions.

V.6 Posterior-Guided Scoring for Geometric Inference

The posterior-guided score combines the anchor-conditioned geometric support proposal with image-conditioned decoder evidence. Let q index a decoder query and let $\hat{Y}_q = (\hat{M}_q, \hat{\mathbf{b}}_q, \ell_q)$ be its predicted mask, box, and confidence logit. For an unseen target n, we use the following scoring form:

$$S(Y_n, q \mid \mathcal{I}, \mathcal{A}, \mathcal{G}) = \Psi_{\text{dec}}(\hat{Y}_q, Y_n, \mathcal{I}, \mathbf{p}_n^{\text{enhanced}}) + \Psi_{\text{geo}}(Y_n \mid \mathcal{A}, \mathcal{G}), \tag{25}$$

where $Y_n = (M_n, \mathbf{b}_n)$ is the latent target support, Ψ_{geo} scores compatibility with the anchor-conditioned support proposals, and Ψ_{dec} scores compatibility with decoder evidence. The geometric support distribution used by Ψ_{geo} is approximated by averaging or sampling anatomical perturbations:

$$P_{\text{geo}}(Y_n \mid \mathcal{A}, \mathcal{G}) = \int P_{\text{geo}}(Y_n \mid \mathcal{A}, \theta_n, \mathcal{G}) P(\theta_n \mid \mathcal{G})\, d\theta_n, \tag{26}$$

using stochastic perturbations of scale, thickness, and offset. In implementation, Ψ_{geo} is a monotonic score derived from this support distribution. The decoder-evidence term favors queries that both respond to the enhanced target prompt and agree spatially with the proposed support:

$$\Psi_{\text{dec}}(\hat{Y}_q, Y_n, \mathcal{I}, \mathbf{p}_n^{\text{enhanced}}) \propto \log \sigma(\ell_q) + \log p_{\text{presence}} - \lambda_m \mathcal{L}_{\text{mask}}(\hat{M}_q, M_n) - \lambda_b \mathcal{L}_{\text{box}}(\hat{\mathbf{b}}_q, \mathbf{b}_n). \tag{27}$$

The pseudo target used for training is the highest-scoring support proposal,

$$\tilde{Y}_n = \arg\max_{Y_n} \max_{q} S(Y_n, q \mid \mathcal{I}, \mathcal{A}, \mathcal{G}), \tag{28}$$

and the matched decoder query is the highest-scoring query under the decoder-evidence term above. In this posterior-guided formulation, target support is proposed from image-specific anchor anatomy, latent relation geometry is explored through stochastic perturbation, and visual evidence determines which decoder hypothesis is finally reinforced.

V.7 Model Architecture Details

Visual Encoder. We adopt the ViT-L/14 backbone from SAM3 as the visual encoder, which processes input images at resolution 1024×1024. The encoder produces multi-scale features at strides of 4, 8, 16, and 32 pixels, with embedding dimension $d = 256$.

Text Encoder. We use the CLIP ViT-L/14 text encoder to extract textual embeddings from anatomical descriptions. The encoder outputs 768-dimensional features, which are projected to $d = 256$ dimensions via a learned linear layer.

Graph Neural Network. The graph fusion module implements F_{GNN} in the main text using three concept-level channels: the current node embedding, the mean embedding of its anatomical neighbors, and the relation embedding of the connected edge. For each anatomical concept v_i, we define

$$\mathbf{h}_i^{\text{self}} = \mathbf{e}_i^{\text{vis}}, \qquad \mathbf{h}_i^{\text{neigh}} = \frac{1}{|\mathcal{N}(v_i)|} \sum_{v_j \in \mathcal{N}(v_i)} \mathbf{e}_j^{\text{vis}}, \tag{29}$$

and select the relation channel as

$$\mathbf{h}_i^{\text{edge}} = \text{SelectEdge}\left(\{\mathbf{r}_{ij} \mid v_j \in \mathcal{N}(v_i)\}\right), \tag{30}$$

where $\text{SelectEdge}(\cdot)$ selects the edge embedding with the highest relation relevance. These channels are concatenated and passed through two residual fusion layers:

$$\mathbf{z}_i = \text{MLP}_{\text{fuse}}\left([\mathbf{h}_i^{\text{self}}; \mathbf{h}_i^{\text{neigh}}; \mathbf{h}_i^{\text{edge}}]\right) + \mathbf{h}_i^{\text{self}}. \tag{31}$$

The fusion MLP uses GELU activations and LayerNorm, producing a knowledge-grounded prompt embedding with dimension $d = 256$. For relation weighting, the scalar score in Eq. 2 is implemented as

$$a_{i,n} = \text{MLP}_{\text{rel}}([\mathbf{z}_i; \mathbf{z}_n; \mathbf{r}_{i,n}]). \tag{32}$$

The same grounded node and edge embeddings are therefore used for semantic prompt grounding and relation weighting in geometric support inference.

Momentum-Based Knowledge Refinement. The visual node embeddings and relation embeddings are refined through a momentum update. For each observed structure v_i with image-derived visual embedding $\mathbf{v}_i$ and text embedding $\mathbf{t}_i$, the node embedding is updated as

$$\mathbf{e}_i^{\text{vis}} \leftarrow m\mathbf{e}_i^{\text{vis}} + (1-m)\left(\beta_{\text{vis}}\mathbf{v}_i + \beta_{\text{text}}\mathbf{t}_i\right), \tag{33}$$

where $m = 0.9$, $\beta_{\text{vis}} = 0.7$, and $\beta_{\text{text}} = 0.3$. For each edge (v_i, v_j), the relation embedding is updated by

$$\mathbf{r}_{ij} \leftarrow m\mathbf{r}_{ij} + (1-m)|\mathbf{t}_i - \mathbf{t}_j|. \tag{34}$$

The symmetric formulation keeps undirected anatomical relations consistent while allowing visual evidence from seen structures to gradually ground the graph representation.

Decoder. We adopt the mask decoder from SAM3, which processes multi-scale visual features and text prompts through a transformer-based architecture. The decoder outputs segmentation masks at 256×256 resolution, which are bilinearly upsampled to the original image size.

Optimization. We use AdamW for 10 epochs with a fixed learning rate of 1×10^{-5}, weight decay of 0.01, and batch size of 8, matching the implementation details in Section 4.

VI Zero-shot Evaluation Protocol

In the zero-shot setting, the model is trained on CardiacUDA using annotations of the seen structures LA, RA, and RV. Masks, boxes, points, scribbles, and validation signals for the unseen targets LV and Myo are excluded from network optimization, knowledge-graph updates, pseudo-label construction, checkpoint selection, and hyperparameter tuning. Model selection uses validation performance on seen structures. Ground-truth masks of unseen structures are used only after prediction for offline metric computation. The graph-level anatomical descriptions and support ranges follow the prior specification in Appendix V.1.

Medical Zero-Shot Setting. In clinical imaging, the target ontology is commonly defined by the acquisition protocol and diagnostic question: relevant organs, chambers, vessels, or tissue boundaries are known before dense annotations are collected. Acquiring reliable pixel-level labels requires expert domain knowledge, standardized definitions, and careful quality control. The task is to localize these clinically specified targets through anatomical relations and image-specific evidence from seen structures.

Anchor-Supported Target-Level Zero-Shot Setting. Our protocol is a target-level zero-shot setting supported by annotated or reliably localized anatomical anchors. In clinical ultrasound, a limited set of reference structures is often routinely visible and comparatively reliable to annotate within a given view. These seen structures provide the image-specific coordinate frame, scale, and anatomical context required for geometric support inference. When no valid dependency configuration is available, the decoder uses the knowledge-enhanced prompt with a non-localizing full-image interface prompt, and no geometric pseudo target or pseudo loss is generated. Partial anchor availability can still support an applicable relation, although localization becomes less constrained.

Prompting Conditions. In the main zero-shot comparisons, bbox/point-dependent baselines receive a full-image box specifying image extent, and text-driven baselines receive the target name. Echo-SAM uses the target name and a pre-defined clinical anatomy graph, with image-specific localization inferred from visible anchors and the geometric support score. A full-image box can disadvantage methods designed for localized prompts. Appendix X.2 therefore compares methods under shared fixed anatomy-prior boxes and shared oracle target boxes. The fixed boxes use general anatomical knowledge without target labels; oracle boxes are derived from target ground truth and evaluate mask refinement with supplied target localization.

Scope of A4C Evaluation. Our main experiments focus on the apical four-chamber (A4C) view because most publicly available adult echocardiographic benchmarks with dense multi-structure annotations are collected or reported in A4C, which is also one of the most widely used clinical views. This view supports clinically important measurements, including chamber size, ventricular morphology, and left ventricular ejection fraction. The shared relation families can be instantiated for other acquisition planes through the view-conditioned formulation in Appendix V.4. Quantitative evaluation in additional views requires corresponding images and dense reference masks.

VII The Pseudo-Algorithm of Echo-SAM

Algorithm 1 presents the complete procedure for zero-shot segmentation of unseen structures.

Computational Complexity. The algorithm complexity is dominated by the visual encoder ($O(HW \cdot d^2)$ for ViT) and the decoder ($O(HW \cdot d^2)$). The GNN computation is negligible with $O(|\mathcal{E}| \cdot d^2)$, where $\mathcal{E}$ is the edges in our knowledge graph.

VIII Additional Conventions

Prompt Construction. For each anatomical structure, we construct text prompts following the template: "[structure name] in an echocardiographic image." For structures with multiple synonyms, we generate multiple prompts and average their embeddings.

Algorithm 1 Zero-Shot Segmentation Pipeline

Require: Image $\mathcal{I}$, knowledge graph $\mathcal{G}$, target structure $s_n \in \mathcal{S}_{\text{unseen}}$
Ensure: Segmentation mask $\mathbf{M}_n$ for s_n

Step 1: Visual Feature Extraction
$\mathbf{F}_{\text{vis}} \leftarrow \text{VisualEncoder}(\mathcal{I})$
$\mathcal{A} \leftarrow \text{LocalizeSeenAnchors}(\mathbf{F}_{\text{vis}}, \mathcal{S}_{\text{seen}})$
Step 2: Knowledge-Grounded Embedding
$\mathbf{z}_n \leftarrow \text{GNN}(\mathbf{e}_n^{\text{vis}}, \{\mathbf{e}_j^{\text{vis}}\}_{v_j \in \mathcal{N}(v_n)}, \{\mathbf{r}_{jn}\}_{v_j \in \mathcal{N}(v_n)})$
$\mathbf{p}_n^{\text{enhanced}} \leftarrow (1-\alpha) \cdot \mathbf{p}_n^{\text{text}} + \alpha \cdot \mathbf{z}_n$
Step 3: Geometric Support Inference
Add any required intermediate supports inferred without unseen labels to $\mathcal{A}$
Identify available dependency supports $\mathcal{D}_{\text{avail}}(s_n)$ from $\mathcal{G}$ and $\mathcal{A}$
if an applicable graph relation has sufficient available dependency supports **then**
    Sample latent geometry $\theta_n = (\rho_n, \gamma_n, \eta_n)$ from $P(\theta_n|\mathcal{G})$
    Estimate anchor-conditioned support score $S_{\text{geo}}(Y_n|\mathcal{A}, \mathcal{G})$
    Select highest-scoring pseudo target $\tilde{Y}_n = (\tilde{M}_n, \tilde{\mathbf{b}}_n)$
else
    No pseudo target is generated; the training pseudo loss for s_n is zero
    Use a full-image interface prompt with $\mathbf{p}_n^{\text{enhanced}}$ at inference
end if
Step 4: Prompt-Based Segmentation
Let $\mathbf{b}_{\text{prompt}} = \tilde{\mathbf{b}}_n$ when inferred support exists; otherwise use the full-image box
Decode queries using $(\mathbf{F}_{\text{vis}}, \mathbf{p}_n^{\text{enhanced}}, \mathbf{b}_{\text{prompt}})$
$\mathbf{M}_n \leftarrow$ mask of the highest-scoring target query
return $\mathbf{M}_n$

Bounding Box Parameterization. Bounding boxes are parameterized as $\mathbf{b} = (x, y, w, h)$, where (x, y) denotes the center coordinates, and (w, h) denotes the width and height. All coordinates are normalized to $[0, 1]$ relative to image dimensions.

Coordinate Convention. Bounding boxes are normalized to the image domain, and masks are generated on the decoder output grid before being compared with predicted masks. These conventions provide a common coordinate system for support-score query assignment and the pseudo-label objective.

IX Supplementary Visualization

Figure I provides qualitative examples for additional cardiac targets. The interatrial septum predictions are located between LA and RA, the interventricular septum follows the boundary separating LV and RV, the mitral valve response is concentrated near the LA–LV junction, and the myocardium predictions form an annular support around the LV cavity. These examples illustrate support proposals for separating membranes, surrounding myocardium, and junctional valve structures. Dense reference masks for these additional targets are unavailable in the evaluated public benchmarks, so these examples support qualitative assessment.

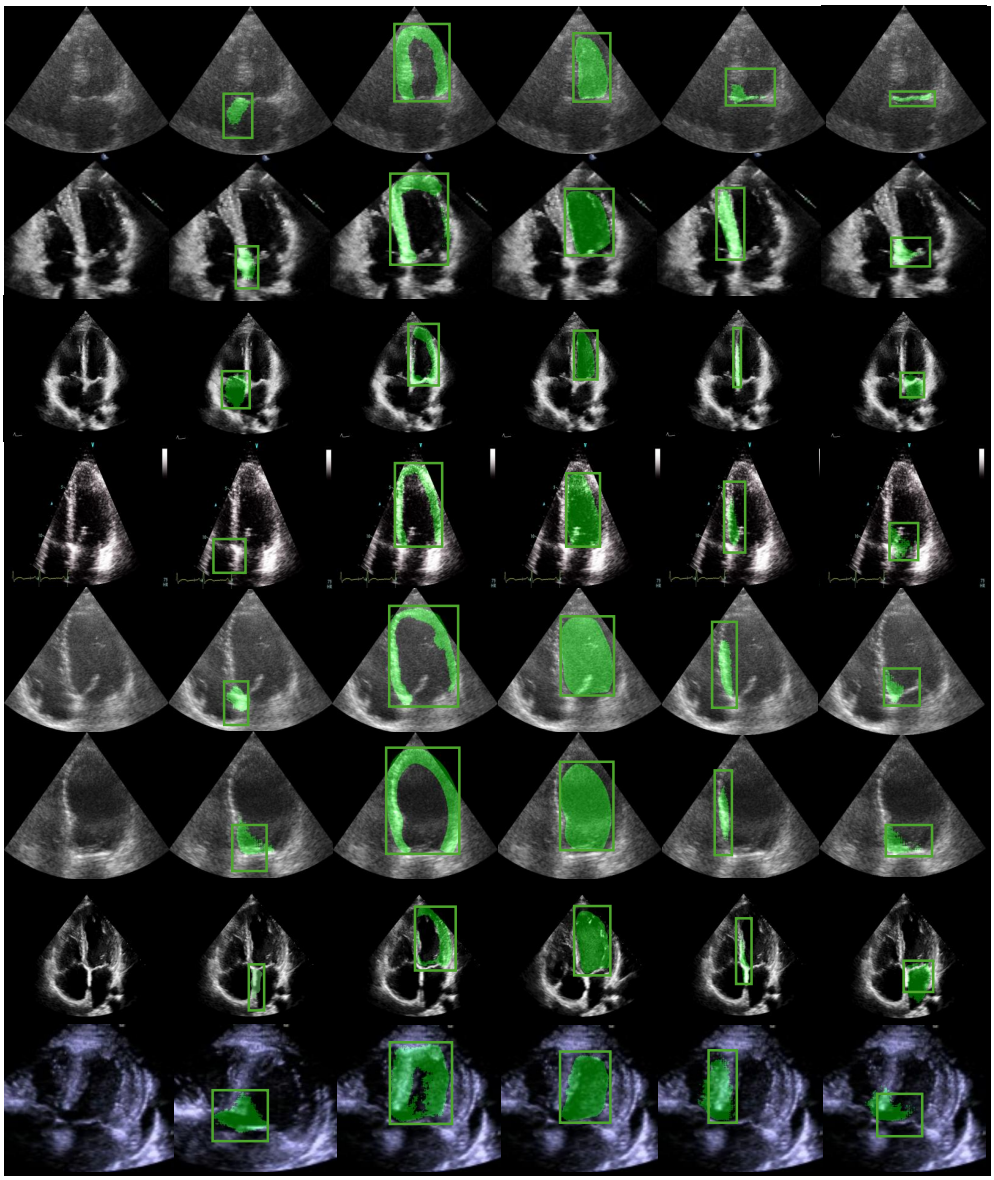


Figure I: Supplementary zero-shot visualizations for additional unseen cardiac structures. Each row shows an input echocardiographic frame followed by Echo-SAM predictions for the interatrial septum, myocardium, left ventricle, interventricular septum, and mitral valve. Green overlays and boxes indicate the inferred target support. These examples show anatomically coherent localization across chamber cavities, septal structures, the myocardial wall, and the mitral valve region under weak ultrasound boundaries.

X Additional Results

This section reports complementary evaluations of anchor reliability, prompting fairness, and structural detection under abnormal anatomy.

Statistical reporting. Unless otherwise stated, DSC is reported in percent as the sample mean $\pm$ sample standard deviation (SD) across three runs. For run-level scores d_r, we use $\bar{d} = \frac{1}{3}\sum_{r=1}^{3} d_r$ and $\mathrm{SD} = \sqrt{\frac{1}{2}\sum_{r=1}^{3}(d_r - \bar{d})^2}$. Anchor robustness evaluations use three independent runs; box-prompt diagnostics use three repeated inference runs without retraining or fine-tuning. Thus, their SDs describe different sources of variability and are not interchangeable. The ASD result is reported as ROC-AUC with a 95% confidence interval (CI) for the population mean ROC-AUC, rather than run-level DSC variability. Numerical differences below are descriptive; SDs alone do not establish statistical significance.

X.1 Sensitivity to Anchor Localization and Availability

Joint mask–box perturbation. We retain every anchor and apply the same sampled center/scale transform to each anchor's mask and bounding box. At perturbation level l, the center displacement is at most l times the original box diagonal, while width and height are independently scaled within $[1-l, 1+l]$. We consider mild ($< 5\%$), moderate (5–30%), and severe (30–50%) perturbations. Five perturbation seeds are averaged within each run before computing the mean and SD across three independent runs.

Table IV: Joint perturbation of anchor masks and boxes. DSC (%), mean $\pm$ run-level SD over three independent runs; five perturbation seeds are averaged within each run.

Perturbation	CAMUS LV	CAMUS Myo	HMC-QU Myo
None	49.39 ± 1.10	60.24 ± 2.87	57.09 ± 2.44
Mild ($< 5\%$)	51.43 ± 1.64	61.91 ± 2.90	60.22 ± 3.05
Moderate (5–30%)	50.90 ± 2.39	63.36 ± 2.61	61.57 ± 2.49
Severe (30–50%)	32.65 ± 3.55	37.16 ± 3.53	36.24 ± 3.26

Table IV shows no degradation under mild or moderate perturbations in these evaluations. One possible explanation is that perturbing pseudo-target construction broadens spatial support and reduces sensitivity to ambiguous ultrasound boundaries. Severe perturbation lowers DSC by 16.74, 23.08, and 20.85 points on CAMUS-LV, CAMUS-Myo, and HMC-QU-Myo, respectively, demonstrating a regime in which anchor errors propagate to target localization.

Missing anchors. We separately remove complete anchor references, including the mask, box, and relation input, while leaving the retained anchors unperturbed. Within each independent run, all removal combinations with the same number of retained anchors are averaged. The three initial seen anchors are LA, RA, and RV.

Table V: Removal of complete anchor references with retained anchors unperturbed. DSC (%), mean $\pm$ SD over three independent runs; all removal combinations of each size are averaged within a run.

Available anchors	CAMUS LV	CAMUS Myo	HMC-QU Myo
All three anchors	49.39 ± 1.10	60.24 ± 2.87	57.09 ± 2.44
One anchor missing	48.48 ± 1.37	56.52 ± 2.55	55.76 ± 2.81
Two anchors missing	47.69 ± 1.49	57.77 ± 2.04	54.19 ± 2.22
No anchor	32.93 ± 5.66	39.74 ± 2.27	35.98 ± 2.70

As shown in Table V, retaining only one anchor reduces DSC by 1.70, 2.47, and 2.90 points relative to the all-anchor setting. Partial loss therefore preserves useful spatial evidence in these tasks. With no anchor, the scores match the graph-only ablation: knowledge-grounded prompt fusion remains available, but geometric support cannot be generated. Together, the two analyses distinguish tolerance to modest localization errors or partial anchor loss from the substantial degradation caused by severely corrupted or entirely absent spatial references.

X.2 Shared-Box Diagnostics for Prompting Fairness

The original full-image prompt supplies no target location and can be weak for box-dependent models. We therefore compare MedSAM [15], Med-SA [14], MemSAM [11], and Echo-SAM under two shared-box conditions. For the *fixed anatomy-prior box*, one coarse category-level box is specified for each of RV, LV, RA, LA, and Myo before evaluation, using general anatomical knowledge without target labels. For the *oracle GT box*, the box is derived from the target ground truth. Each method receives the same box for the corresponding target under each condition. For Echo-SAM, the shared box replaces its inferred localization support. All values use three repeated inference runs without retraining or fine-tuning. The original column retains each method's main evaluation protocol, including inferred support for Echo-SAM.

Table VI: Shared-box diagnostics on CAMUS-LV. DSC (%), mean ± SD across three repeated inference runs without retraining. Oracle boxes use target ground truth and do not evaluate zero-shot localization.

Method	Original protocol	Fixed anatomy-prior box	Oracle GT box
MedSAM	33.60 ± 1.08	56.27 ± 2.48	79.45 ± 2.30
Med-SA	33.44 ± 1.42	50.50 ± 2.35	68.94 ± 2.71
MemSAM	0.00 ± 0.00	9.64 ± 1.86	24.68 ± 1.59
Echo-SAM	49.39 ± 1.10	65.53 ± 2.70	90.92 ± 2.81

Table VII: Shared-box diagnostics on CAMUS-Myo under the same protocol as Table VI. DSC (%), mean ± SD across three repeated inference runs.

Method	Original protocol	Fixed anatomy-prior box	Oracle GT box
MedSAM	16.95 ± 1.76	6.51 ± 1.11	19.23 ± 1.39
Med-SA	6.14 ± 1.56	7.52 ± 1.88	16.67 ± 1.20
MemSAM	0.00 ± 0.00	3.78 ± 1.28	17.78 ± 1.64
Echo-SAM	60.24 ± 2.87	48.64 ± 3.06	81.71 ± 2.45

On CAMUS-LV (Table VI), Echo-SAM exceeds the strongest baseline by 9.26 points with the fixed anatomy-prior box and by 11.47 points with the oracle box. On CAMUS-Myo (Table VII), all three baselines remain below 20 DSC even with oracle localization. A box specifies location but does not encode the surrounding relation between myocardium and the ventricular cavity; baseline predictions can include the cavity. The coarse fixed box also reduces Echo-SAM from 60.24 to 48.64 DSC because it replaces more specific inferred support and can cover both structures. Under that same fixed box, Echo-SAM still exceeds the strongest baseline by 41.12 points. The advantage persists under shared localization inputs. Oracle-box results assess segmentation with supplied target localization.

X.3 Structural Detection under Congenital Abnormality

An atrial septal defect (ASD) can make the interatrial septum (IAS) discontinuous or partially absent, violating a normal-anatomy relation between the atria. We evaluate whether IAS detection confidence remains informative on CardiacNet-ASD [42]. Lower target-detection confidence is interpreted as a higher ASD score; equivalently, the score can be written as $1 - c_{\mathrm{IAS}}$ for confidence $c_{\mathrm{IAS}} \in [0, 1]$. No ASD classification labels or IAS segmentation labels are used to train or fine-tune Echo-SAM; ASD labels are used only for evaluation.

Echo-SAM achieves 73.71 ROC-AUC (95% CI: 65.47–82.95). Its point estimate is 1.46 points above ResNet3D, while remaining below the dedicated disease-classification methods reported by CardiacNet. The supervision regimes are stated in Table VIII. The available disease-level labels support evaluation of the IAS confidence signal for ASD detection; IAS segmentation accuracy remains to be measured using dense tissue masks. Target-presence calibration and pathology-conditioned relations are directions for extending the analysis of abnormal anatomy.

Table VIII: ASD assessment on CardiacNet-ASD. Comparator scores are reproduced from CardiacNet [42], Table 2; they use disease-specific training. Echo-SAM uses IAS detection confidence without ASD or IAS supervision. The 95% CI for the population mean ROC-AUC is reported only for Echo-SAM.

Method	ROC-AUC (%)	95% CI
Disease-specific methods reported by CardiacNet		
ResNet3D	72.25	—
AGXNet	76.52	—
EchoNet	83.62	—
DeepGuide	85.02	—
DiffMIC	82.81	—
HiFuse	81.08	—
CardiacNet	91.24	—
Structural detection without ASD-specific training		
Echo-SAM	73.71	[65.47, 82.95]

X.4 Failure Analysis

The anchor experiments identify two measurable failure regimes: severe displacement of spatial references and complete anchor loss. More generally, four conditions can undermine prediction: (i) absent or severely mislocalized anchors produce weak or misplaced support; (ii) acoustic shadowing or weak boundaries prevent precise mask delineation even when support is plausible; (iii) image truncation removes structures needed by an active graph relation; and (iv) abnormal anatomy invalidates expected topology and can produce a false-positive proposal for absent or discontinuous tissue. The ASD evaluation examines a detection signal associated with the fourth condition.

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: Please see our Abstract and Introduction Sections.

   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: Please see our Conclusion section and the Broader Impact in appendix section I.

   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: [Yes]

Justification: We have provided the full set of assumptions and a complete (and correct) proof for all necessary parts.

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: We have provided all information needed in our experimental section and appendix. (Please see section 4 and our appendix.

   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: We will release all our code publicly upon the paper being accepted.

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: We have specified all the training and test details in our Result section. (Please see section 4)

   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: the significance of the difference in segmentation results is validated by the T-test, our results are with $P \leq 0.05$, which denotes statistically significant.

   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: Please see section 4

   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:

   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: Please see the section I in our Appendix.

    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: [N/A]

    Justification:

    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: We cite these works and comply with their relevant agreements.

    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:

    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: [N/A]

    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: [N/A]

    Justification: [N/A]

    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: [Yes]

Justification: [Yes]

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.