What, Where, and Boundary: Hierarchical Cognitive Decomposition for Echocardiography Video Segmentation

Long Zheng, Zhi Li *, **Weidong Wang, Chenxi Zhu, Zhenyu Dai, Minyi Guo**
State Key Laboratory of Public Big Data, College of Computer Science and Technology
Guizhou University, China
{gs.lzheng23,zhili,gs.wdwang24,cxzhu24,zydai,myguo}@gzu.edu.cn

Abstract

Segmenting the left ventricle in echocardiographic videos remains difficult because the target undergoes continuous deformation and rapid displacement across the cardiac cycle, while low tissue contrast further obscures its endocardial boundaries. Inspired by the roles of anatomical knowledge and focused inspection in clinical interpretation, we organize segmentation around three functional questions: *What is the target? Where is it now? Where does the boundary lie?* We propose Hierarchical Cognitive Decomposition (HCD), a cognition-inspired framework that translates these questions into three complementary stages of identity anchoring, spatial localization, and boundary delineation. To capture invariant target identity amid changing appearances, Static Identity Anchoring (SIA) anchors the first annotated frame as a semantic reference, enabling stable recognition across the cardiac cycle. To maintain spatial focus despite inter-frame motion, Adaptive Gaze Prior (AGP) converts the previous frame's prediction into a spatial attention prior that localizes the current target position. To resolve boundary ambiguity under low tissue contrast, Structural Boundary Perception (SBP) extracts multi-scale boundary cues within the localized region. On the CAMUS and EchoNet-Dynamic benchmarks, HCD achieves state-of-the-art performance with only 1.35M trainable parameters (out of 35.2M total) at 68 FPS, demonstrating the effectiveness of this clinically inspired decomposition for accurate and efficient segmentation. The code is available at `https://github.com/ZhengLong777/HCD`.

1 Introduction

Cardiovascular diseases are a leading cause of mortality worldwide, driving the clinical demand for precise and efficient cardiac assessment. Echocardiography is the primary non-invasive modality for evaluating cardiac function, and accurate left ventricular segmentation is essential for deriving clinical metrics such as ejection fraction and stroke volume [1]. However, as shown in Fig. 1, segmenting the left ventricle poses severe challenges: continuous deformation, rapid displacement across the cardiac cycle, and low tissue contrast that renders endocardial boundaries ambiguous.

To address these challenges, existing echocardiographic segmentation methods have explored temporal consistency modeling [2], memory-based propagation [3, 4], and adaptation of pretrained visual representations [5]. These approaches strengthen cross-frame association and semantic or structural representations, providing complementary information for segmentation under changing appearances. Accurate anatomical delineation also involves interpreting uncertain local evidence in relation to the

*Corresponding author.

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

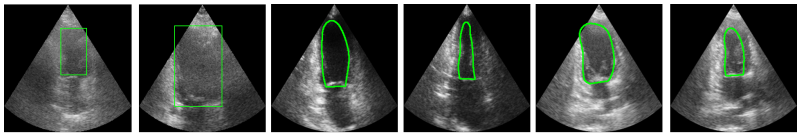


Figure 1: Illustrative challenges for echocardiography video segmentation: (a–b) indistinct or blurred contours, and (c–f) the substantial changes in the target's shape and scale throughout the cardiac cycle.

intended target. Building on these advances, we examine how target identity, spatial location, and local boundary evidence can be coordinated to support contour estimation.

Clinical image interpretation offers a useful perspective on this problem. Research on medical image perception highlights the relationship between global image information and focused local inspection [6, 7]. For anatomical delineation, knowledge of the intended structure provides context for selecting relevant regions and interpreting uncertain contour cues, while regional attention establishes a focus for examining finer details. Drawing on this perspective, we organize segmentation around three complementary questions: *What is the target? Where is it now? Where does the boundary lie?* These questions distinguish the roles of target identity, spatial localization, and boundary delineation while connecting them within the segmentation process.

We propose Hierarchical Cognitive Decomposition (HCD), a cognition-inspired framework that translates this organization into progressive feature refinement. Static Identity Anchoring (SIA) retrieves patient- and sequence-specific target information from the annotated first frame, providing a semantic reference across changing appearances. Adaptive Gaze Prior (AGP) derives a soft spatial prior from the preceding prediction and uses it to focus the reference-conditioned features on the current region of interest. Structural Boundary Perception (SBP) then refines these guided features alongside finer-resolution image features, supporting contour delineation under weak tissue contrast. Through these interactions, the target reference informs the representations used for localization, and the resulting spatial guidance carries into structural refinement.

Our main contributions are:

- We propose HCD, a cognition-inspired framework that organizes echocardiographic video segmentation around three complementary functions: identity anchoring, spatial localization, and boundary delineation.
- We design three dedicated modules, SIA, AGP, and SBP, that connect target reference retrieval, prediction-derived spatial guidance, and multi-scale structural refinement within a unified framework.
- Experiments on CAMUS and EchoNet-Dynamic demonstrate that HCD achieves state-of-the-art segmentation accuracy and clinically consistent LV_{EF} estimation, while requiring only 1.35M trainable parameters at 68 FPS.

2 Related Work

2.1 Echocardiography Video Segmentation

Video object segmentation provides a foundation for using target information and temporal correspondence across frames. OSMN [8] and RGMP [9] combine target appearance with spatial cues from preceding predictions to support mask propagation. Memory-based approaches, including STM [10], XMem [11], and Cutie [12], further develop mechanisms for retrieving relevant information from past observations as appearances evolve. Foundation models have expanded these capabilities through pretrained visual representations, with SAM2 extending promptable segmentation to videos [13, 14]. These developments provide reusable mechanisms for target conditioning, cross-frame association, and visual representation learning.

In echocardiography, segmentation must accommodate continuous cardiac deformation while delineating anatomical structures under weak tissue contrast. Beyond frame-wise CNN prediction [15, 16],

temporal consistency and semi-supervised learning exploit information across frames and reduce reliance on dense annotations [2, 17, 18]. Neighborhood correlation mining and hierarchical spatiotemporal aggregation further strengthen contextual modeling [19, 20]. Reference propagation and foundation-model adaptation offer complementary approaches: MemSAM integrates SAM with a space-time memory bank [3], while GDKVM develops linear key-value association with gated memory updates for echocardiographic videos [4]. Recent methods also connect temporal processing with semantic and structural refinement. EchoVim introduces dynamic interaction and semantic token-attentive refinement [21], EchoForge combines structural prototype propagation with pseudo-label refinement [22], and E^3SAM2 incorporates entropy-aware adaptation and explicit edge supervision [5].

Together, these methods advance temporal association, target conditioning, and structural refinement for echocardiographic segmentation. In the presence of weak or ambiguous boundary cues, accurate delineation benefits from coordinating target identification and localization with the interpretation of local evidence in relation to the intended anatomy. This motivates examining how target information can guide both regional localization and detailed contour estimation.

2.2 Hierarchical Perception in Medical Image Interpretation

Research in medical image perception examines how global image information and focused local inspection contribute to expert interpretation [6]. In mammogram interpretation, rapid initial fixation of abnormalities provides evidence that global perception can guide attention toward relevant regions [7]. Related observations have been reported in simulation-based echocardiography, where experts fixated case-specific regions of interest earlier than non-experts [23]. These findings motivate linking broader image information with the selection of regions for closer examination.

For anatomical delineation, regional attention must be complemented by an understanding of the structure being outlined. Clinical quantification guidelines define ventricular contours with reference to anatomical interfaces and landmarks [1]. These definitions provide context for interpreting local image evidence, particularly where tissue contrast is weak or boundaries are incomplete. Knowledge of the intended anatomy therefore informs both where to inspect and how to interpret uncertain contour cues, while focused inspection supports the examination of finer structural details.

Drawing on these perceptual and clinical perspectives, HCD formulates three complementary questions: *What is the target? Where is it now? Where does the boundary lie?* These questions connect identity anchoring, spatial localization, and boundary delineation. Within this framework, the annotated first frame provides a patient- and sequence-specific example of the intended anatomy, serving as a concrete reference for subsequent segmentation. This perspective motivates using target information to guide regional attention and the interpretation of uncertain contour cues.

3 Method

3.1 Overview

The overall architecture of HCD is shown in Fig. 2. Given an echocardiography video $V = \{I_t\}_{t=0}^{T-1}$, the goal is to predict segmentation masks $\{p_t\}_{t=0}^{T-1}$ for all frames. Following SAM2-UNet [24], we adopt a frozen SAM2 Hiera-Small encoder [14] with lightweight Adapters to extract multi-scale features per frame. HCD then follows a functional hierarchy for identity anchoring, spatial localization, and boundary delineation. Refined multi-scale features are decoded under deep supervision, and per-frame predictions are smoothed via a sliding-window average. Only the Adapters, stage-specific modules, and decoder are optimized during training.

3.2 Cognitive Decomposition

HCD implements the clinically inspired decomposition in Section 2.2 through three complementary operators. The identity operator ϕ_{id} retrieves target-specific information from a fixed first-frame reference, providing patient- and sequence-specific guidance. The localization operator ϕ_{loc} constrains spatial attention using the preceding prediction and the temporal continuity of cardiac motion. The boundary operator ϕ_{bd} enhances contour-sensitive features for fine endocardial delineation. Together, these operators translate the *what*, *where*, and *boundary* questions into progressive feature refinement.

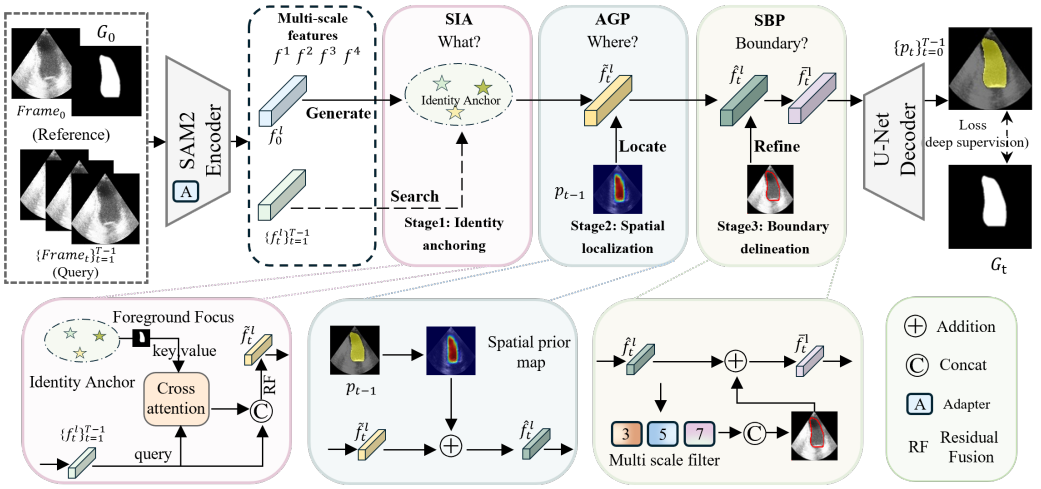


Figure 2: Overview of the proposed HCD framework. Given the first-frame annotation G_0 and encoder features $\{f_t^l\}$, HCD progressively refines per-frame representations through three complementary processing stages. SIA (*what the target is*) anchors the first-frame features as a static semantic reference for identity retrieval. AGP (*where it is now*) converts the previous prediction p_{t-1} into a soft spatial prior for current-frame localization. SBP (*where the boundary lies*) extracts multi-scale boundary cues through parallel depthwise filters to sharpen endocardial contour delineation.

Let $\{f_t^l\}_{l=1}^4 = \varepsilon(I_t)$ denote the multi-scale features extracted by the encoder, where l indexes the feature hierarchy from shallow to deep layers. In abstract form, HCD applies these operators progressively over the multi-scale feature representation:

$$\tilde{f}_t = \phi_{\text{id}}(f_t;\, f_0, G_0), \tag{1}$$

$$\hat{f}_t = \phi_{\text{loc}}(\tilde{f}_t;\, p_{t-1}), \tag{2}$$

$$\bar{f}_t = \phi_{\text{bd}}(\hat{f}_t,\, f_t), \tag{3}$$

$$p_t = D(\bar{f}_t). \tag{4}$$

Here, f_t, $\tilde{f}_t$, $\hat{f}_t$, and $\bar{f}_t$ denote multi-scale feature sets rather than a single feature map. The progressive notation $f_t \to \tilde{f}_t \to \hat{f}_t \to \bar{f}_t$ reflects sequential refinement through identity anchoring, spatial localization, and boundary delineation. These operators are instantiated as SIA (Section 3.3), AGP (Section 3.4), and SBP (Section 3.5), where the specific feature levels, input configurations, and architectural details of each operator are described in the respective sections.

SIA and AGP operate at the deeper feature levels $l \in \{4, 3\}$, where target reference information and prediction-derived spatial guidance are combined. SBP refines the guided features $\hat{f}_t^3$ alongside the original, finer-resolution features f_t^2, bringing regional guidance and local structural detail into the representations used for decoding.

3.3 Static Identity Anchoring (SIA)

As the first stage of HCD, SIA addresses *what the target is* by providing a patient- and sequence-specific reference. This reference complements the anatomical representations learned during training and conditions current-frame features on the structure to be segmented.

In the present implementation, SIA retains the features and annotation of the first frame as a fixed semantic reference. Current-frame queries retrieve relevant target information through cross-attention, so the retrieved representation varies with the current observation even though the stored reference remains unchanged. AGP subsequently introduces spatial guidance from the preceding prediction. The reference-update ablation compares this fixed-reference implementation with EMA and FDGU within the evaluated HCD configuration (Appendix A.3).

Formally, SIA instantiates the identity anchoring operator ϕ_{id} defined in Section 3.2. For feature levels $l \in \{4, 3\}$, it maps the encoder feature f_t^l to an identity-enhanced representation:

$$\tilde{f}_t^l = \phi_{SIA}(f_t^l;\, f_0^l, G_0), \quad l \in \{4, 3\} \tag{5}$$

where f_0^l denotes the first-frame feature at level l, and G_0 is the first-frame annotation. The current feature f_t^l queries the anchored reference f_0^l through multi-head cross-attention with projections:

$$Q_t^l = W_Q^l f_t^l, \quad K_0^l = W_K^l f_0^l, \quad V_0^l = W_V^l f_0^l \tag{6}$$

To emphasize target-region information, mask-derived spatial gating is applied to the value features:

$$\hat{V}_0^l = V_0^l \odot M^l = V_0^l \odot \left[\tilde{G}_0^l + (1-\lambda)(1-\tilde{G}_0^l)\right] \tag{7}$$

where $\tilde{G}_0^l$ is the first-frame annotation G_0 downsampled to the spatial resolution of level l. $\lambda \in [0, 1]$ controls background suppression strength, with λ=0 leaving the reference unchanged and λ=1 applying hard foreground gating by zeroing the background values. The effect of varying the gating strength is analyzed in Tab. 4 (supplement). The reference feature is then computed via scaled dot-product attention:

$$R_t^l = \mathrm{softmax}\left(\frac{Q_t^l (K_0^l)^\top}{\sqrt{d}}\right) \hat{V}_0^l \tag{8}$$

where d denotes the per-head dimension. The retrieved reference R_t^l is concatenated with the original feature and fused through a residual block, yielding the identity-enhanced feature:

$$\tilde{f}_t^l = \mathrm{Fuse}^l([f_t^l,\, R_t^l]) \tag{9}$$

3.4 Adaptive Gaze Prior (AGP)

With target guidance provided by SIA, AGP addresses *where the target is now*. It uses the preceding prediction as a recent estimate of ventricular location and extent, converting it into a soft spatial prior that guides the reference-conditioned features. Dilation and smoothing broaden the estimated region to accommodate inter-frame displacement, while learnable refinement adapts the prior before feature injection.

Formally, AGP instantiates the spatial localization operator ϕ_{loc} defined in Section 3.2. For feature levels $l \in \{4, 3\}$, it maps the identity-enhanced feature $\tilde{f}_t^l$ to a spatially focused representation:

$$\hat{f}_t^l = \phi_{loc}(\tilde{f}_t^l;\, p_{t-1}), \quad l \in \{4, 3\} \tag{10}$$

where p_{t-1} denotes the segmentation prediction from the previous frame. AGP first generates a soft spatial prior A_t by converting p_{t-1} through area downsampling to the spatial resolution of f^1, morphological dilation via max-pooling, Gaussian smoothing, and learnable convolutional refinement:

$$A_t = \psi(\mathcal{G}(\mathrm{MaxPool}_k(\mathrm{Down}_1(p_{t-1})))) \tag{11}$$

where $\mathrm{Down}_1(\cdot)$ denotes area downsampling to the resolution of f^1, $\mathrm{MaxPool}_k$ is dilation with kernel size k, $\mathcal{G}$ is a fixed Gaussian kernel, and ψ is a learnable refinement network. The resulting prior A_t is bilinearly resized to the target spatial resolution of level l, and injected as an additive spatial bias:

$$\hat{f}_t^l = \tilde{f}_t^l + \alpha_l A_t, \quad l \in \{4, 3\} \tag{12}$$

where α_l is a scale-specific weighting factor. For the first frame, where no previous prediction is available, p_{t-1} is replaced by the annotated mask G_0.

3.5 Structural Boundary Perception (SBP)

With target and spatial guidance available, the final stage of HCD addresses boundary delineation, namely *where the boundary lies*. Low tissue contrast can still leave local contour details ambiguous. Motivated by the role of focused local inspection discussed in Section 2.2, SBP introduces dedicated processing for these details before decoding. It uses learnable multi-scale depthwise filters on the guided features and finer-resolution encoder features to support endocardial contour estimation.

Formally, SBP instantiates the boundary delineation operator ϕ_{bd} defined in Section 3.2. It operates on two feature levels with distinct inputs:

$$\bar{f}_t^3 = \phi_{SBP}^{(3)}(\hat{f}_t^3), \quad \bar{f}_t^2 = \phi_{SBP}^{(2)}(f_t^2) \tag{13}$$

where $\hat{f}_t^3$ is the identity- and location-enhanced feature produced by SIA and AGP, while f_t^2 is the original encoder feature at the shallower level. This design allows SBP to detect boundaries within semantically localized representations at f^3 while exploiting higher spatial resolution for fine-grained geometric detail at f^2.

At each level $l \in \{3, 2\}$, the filters first reduce the input channel dimensionality, then apply parallel depthwise convolutions with kernel sizes $k \in \mathcal{K}$ to capture boundary responses at multiple scales. The multi-scale responses are concatenated and projected back to the original dimensionality, yielding boundary enhancements B_t^3 and B_t^2. The enhanced features are obtained through residual injections:

$$\bar{f}_t^3 = \hat{f}_t^3 + \beta_3 \cdot B_t^3, \quad \bar{f}_t^2 = f_t^2 + \beta_2 \cdot B_t^2 \tag{14}$$

where β_l is a scale-specific weighting factor. SBP is optimized end to end through the segmentation loss without explicit boundary supervision.

4 Experiments

4.1 Datasets and Evaluation Metrics

HCD is evaluated on two public benchmarks: **CAMUS** [25] (500 patients, apical 2- and 4-chamber videos with frame-by-frame ED-to-ES annotations) and **EchoNet-Dynamic** [26] (10,030 apical 4-chamber videos with sparse ED/ES annotations). We report mean Dice coefficient (mDice), 95th-percentile Hausdorff Distance (HD95), and for clinical evaluation, Left Ventricular Ejection Fraction (LV_{EF}) estimated via Simpson's rule [27] with Pearson correlation, mean bias, and standard deviation.

4.2 Implementation Details

Following SAM2-UNet [24], HCD uses a frozen SAM2 Hiera-Small encoder [14]. The lightweight adapters, HCD modules, and decoder are jointly optimized. Training is performed on a single NVIDIA A100 GPU for 100 epochs on CAMUS and 50 epochs on EchoNet-Dynamic. We use AdamW [28] with a learning rate of 1×10^{-3} and a batch size of 4.

Images are resized to 256×256 for CAMUS and 128×128 for EchoNet-Dynamic. Each video is uniformly sampled to 10 frames, with ED and ES as the first and last frames. Gamma correction, scaling, rotation, and contrast adjustment are each applied with a probability of 0.5. Frames are propagated autoregressively during both training and inference. On both datasets, supervision is applied only to the annotated ED and ES frames. Following [29, 30], we combine weighted IoU and weighted binary cross-entropy losses, with equal-weight deep supervision at all three decoder outputs.

For AGP, a 7×7 max-pooling dilation is followed by smoothing with a 5×5 Gaussian kernel ($\sigma = 1.0$) and refinement through the lightweight convolutional network ψ. SBP uses depthwise convolution kernels of sizes $\mathcal{K} = \{3, 5, 7\}$. At inference, predictions are smoothed using a sliding-window average of size 3. We follow the GDKVM [4] experimental protocol for fair comparison.

4.3 Comparison with State-of-the-art Methods

We compare HCD against eight task-specific methods (PKEchoNet [18], DSA [31], MemSAM [3], SimLVSeg [32], E^{3}SAM2 [5], EchoForge [22], EchoVim [21], and GDKVM [4]) and four related methods (XMem++ [33], Cutie [12], VideoMamba [34], and Vision LSTM [35]). Baseline results are taken from GDKVM [4], except for EchoForge [22] and EchoVim [21], whose results are quoted from their respective publications, and E^{3}SAM2 [5], which we evaluate in this study.

Table 1: Comparison with state-of-the-art methods on CAMUS and EchoNet-Dynamic. Segmentation is evaluated using mDice and HD95 (mm); LV_{EF} estimation is evaluated using Pearson correlation (corr) and bias±std (percentage points). HCD results are from a single training run. A dash denotes unavailable results.

Method	**CAMUS**				**EchoNet-Dynamic**			
	mDice ↑	**HD95 ↓**	**corr ↑**	**bias±std (%)**	**mDice ↑**	**HD95 ↓**	**corr ↑**	**bias±std (%)**
XMem++ ICCV'23	89.38	4.03	0.746	1.70 ± 21.90	87.51	3.14	0.692	5.77 ± 10.40
Cutie CVPR'24	91.09	3.89	0.787	1.67 ± 21.70	88.96	2.89	0.695	4.30 ± 11.70
VideoMamba ECCV'24	91.96	3.48	0.780	-4.49 ± 19.40	90.22	2.79	0.764	-2.92 ± 9.38
Vision LSTM ICLR'25	92.14	3.79	0.806	-0.31 ± 18.80	90.24	2.65	0.768	-2.42 ± 9.10
PKEchoNet AAAI'23	93.49	3.42	0.862	-1.53 ± 16.40	92.60	2.53	0.852	1.29 ± 9.62
DSA TMI'24	94.25	3.27	0.891	0.86 ± 13.40	92.91	2.46	0.868	1.42 ± 9.18
MemSAM CVPR'24	93.63	3.47	0.878	-0.89 ± 12.30	92.71	2.56	0.859	1.09 ± 9.44
SimLVSeg UMB'24	92.54	3.65	0.895	1.83 ± 13.80	91.91	2.65	0.794	0.91 ± 9.56
E^3SAM2 AAAI'26	93.42	3.67	0.891	-2.41 ± 5.55	92.41	3.17	0.776	0.24 ± 8.69
EchoForge CVPR'26	94.89	3.12	0.913	$0.23 \pm -$	93.63	3.05	**0.887**	$-0.51 \pm -$
EchoVim ACM MM'25	94.81	3.02	0.908	0.42 ± 4.87	93.68	**2.35**	0.884	1.63 ± 8.71
GDKVM ICCV'25	95.11	3.05	0.904	-0.19 ± 11.30	93.46	2.38	0.872	-0.70 ± 9.15
HCD (Ours)	**95.21**	**2.59**	**0.914**	4.18 ± 4.93	**94.38**	2.36	0.842	-2.20 ± 6.79

Quantitative results. As shown in Table 1, HCD achieves the highest mDice among the compared methods on both CAMUS and EchoNet-Dynamic. It also obtains the lowest HD95 on CAMUS and competitive boundary accuracy on EchoNet-Dynamic. These results support the effectiveness of combining identity anchoring, spatial localization, and boundary refinement for echocardiographic video segmentation.

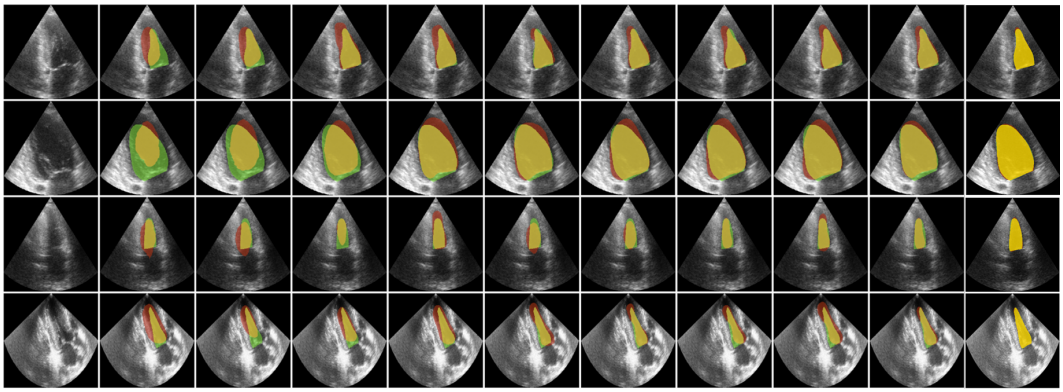


Figure 3: Visual comparison with state-of-the-art methods on the CAMUS test set. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

Qualitative results. Figure 3 presents qualitative comparisons on challenging CAMUS cases with weak endocardial boundaries. HCD better preserves ventricular cavity completeness and apical contours, while reducing over-segmentation and local boundary artifacts. The larger overlap regions with the ground truth are consistent with the quantitative segmentation results.

Clinical metrics. Table 1 also reports LV_{EF} estimation, with HCD achieving correlations of 0.914 on CAMUS and 0.842 on EchoNet-Dynamic. On EchoNet-Dynamic, its error spread is smaller than those of GDKVM and EchoVim, although its correlation is lower. Regression and Bland–Altman plots are provided in Fig. 4 for CAMUS and Fig. 9 in the appendix for EchoNet-Dynamic.

Efficiency. As shown in Fig. 5, HCD achieves 95.21 mDice at 68 FPS on a single RTX 2080Ti with 35.2M total parameters, of which only 1.35M (3.8%) are trainable. Other methods in the figure are reported on an RTX 3090 [4], including GDKVM at 37 FPS; these comparisons are indicative because the hardware differs.

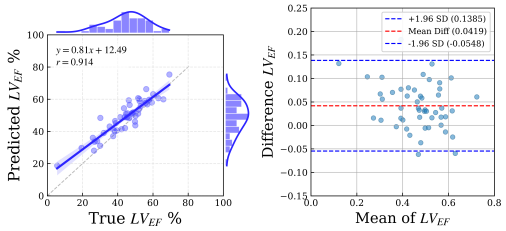


Figure 4: Linear regression and Bland–Altman analysis of LV_{EF} on CAMUS.

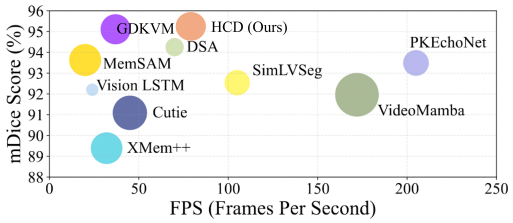


Figure 5: Accuracy–efficiency comparison on CAMUS. Bubble size denotes total parameter count.

4.4 Ablation Studies

We evaluate individual and combined configurations on CAMUS to examine the contributions of SIA, AGP, and SBP (Table 2).

Table 2: Component ablation on CAMUS. The empty row denotes the baseline. mDice (%); HD95 (mm).

SIA	AGP	SBP	mDice ↑	HD95 ↓
			93.77	3.45
✓			94.91	2.80
	✓		94.66	2.82
		✓	94.11	3.21
✓	✓		95.10	2.64
✓		✓	94.72	2.88
	✓	✓	94.86	2.72
✓	✓	✓	**95.21**	**2.59**

Module contributions. Each module improves over the baseline when used individually, with SIA providing the largest mDice gain. Combining SIA and AGP achieves 95.10 mDice and an HD95 of 2.64 mm. Adding SBP further improves mDice to 95.21 and reduces HD95 to 2.59 mm, yielding the best results on both metrics. Its effect depends on the module combination, as SIA with SBP alone does not improve over SIA. Overall, reference conditioning and spatial localization account for the main improvements, while SBP provides modest additional improvements in overlap and boundary accuracy in the complete pipeline.

Role diagnostic visualization. Figure 6 shows spatially selective SIA attention on the fixed reference and an AGP prior covering the ventricle in each query frame, illustrating their respective retrieval and localization roles. SIA attention is averaged over heads and query tokens within the predicted ventricle. Additional analysis of SIA reference gating is provided in Appendix A.2.

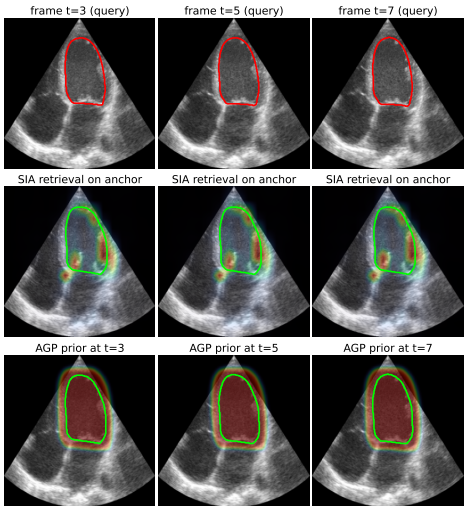


Figure 6: SIA and AGP visualization on CAMUS. Top: query frames with predictions (red). Middle: SIA attention on the fixed reference. Bottom: smoothed AGP priors before learnable refinement. Green contours denote ground truth.

First-frame anchor sensitivity. Without retraining, we perturb only the first-frame mask at inference using dilation, erosion, translation, mixed morphology with translation, and contour noise. Figure 7 shows gradual performance degradation as anchor Dice decreases. The effect also depends on the perturbation type: at an anchor Dice of approximately 0.84, dilation reduces mDice by about 6.2 points, compared with 1.8 points for contour noise. This highlights the greater impact of systematic geometric errors than local contour irregularities in the evaluated setting.

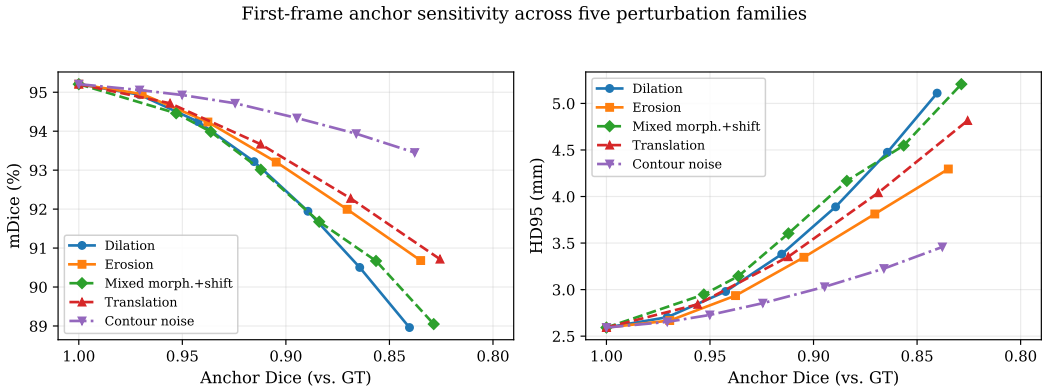


Figure 7: First-frame anchor sensitivity on CAMUS. Left: segmentation mDice. Right: HD95. Anchor Dice measures agreement between the perturbed mask and its ground truth and decreases from left to right.

Reference-update ablation. Within HCD, the sequence-fixed reference achieves higher mDice and lower HD95 than the evaluated exponential moving average (EMA) and feature-dissimilarity-gated update (FDGU) strategies on 10-frame CAMUS clips (Table 5 in Appendix A.3), supporting its use together with frame-dependent spatial refinement in this setting.

4.5 Extension to Multi-Structure Segmentation

We extend HCD to jointly segment the LV cavity, myocardium, and left atrium (four classes including background), using the same CAMUS split, test patients, data pipeline, and semi-supervised protocol as the main experiments. Supervision remains limited to the first and last frames of each clip, allowing

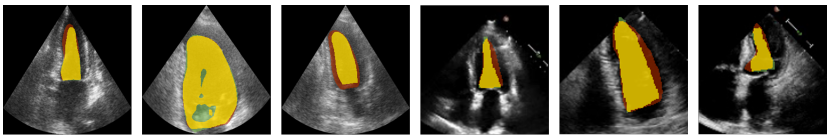
direct comparison of LV-cavity results with Table 1. Only the three 1×1 decoder heads are widened from one to four output channels. SIA gates the reference using the union of the annotated structures, while AGP uses the total foreground probability from the softmax output. SIA and AGP remain shared across structures, while SBP and the frozen backbone are unchanged.

Table 3: Joint three-structure segmentation on CAMUS. Values are averaged over three training runs. The foreground average excludes background.

Structure	Dice (%) ↑	HD95 (mm) ↓
LV cavity	94.18	3.22
Myocardium	89.44	3.28
Left atrium	92.54	3.63
Foreground average	92.05	3.38

With shared guidance modules, HCD achieves an average Dice of 92.05% and an HD95 of 3.38 mm across the three foreground structures, demonstrating its applicability to joint multi-structure segmentation. Table 3 shows that broader target coverage trades some LV-specific accuracy: compared with the dedicated model, LV Dice decreases from 95.21 to 94.18, HD95 increases from 2.59 to 3.22 mm, and the multi-structure model achieves a mean LV_{EF} correlation of 0.866, compared with 0.914 for the dedicated model. The dedicated model therefore remains preferable when only LV analysis is required.

4.6 Discussion and Limitations



(a) (b) (c) (d) (e) (f)

Figure 8: Failure cases on CAMUS (a–c) and EchoNet-Dynamic (d–f).

Representative failure cases are shown in Fig. 8. On CAMUS, HCD under-segments when the endocardial border is almost entirely obscured by acoustic shadows. On EchoNet-Dynamic, failures occur under extreme foreshortening, where the first-frame reference becomes less representative. Further improving multi-structure accuracy and extending HCD to non-periodic medical video scenarios represent promising directions for future investigation.

5 Conclusion

We propose HCD, a cognition-inspired framework that organizes echocardiographic video segmentation around three functional questions: *what the target is*, *where it is now*, and *where its boundary lies*. SIA, AGP, and SBP implement identity anchoring, spatial localization, and boundary delineation at different feature levels. HCD achieves state-of-the-art segmentation accuracy and supports LV_{EF} estimation on CAMUS and EchoNet-Dynamic with only 1.35M trainable parameters at 68 FPS, supporting the effectiveness of this clinically inspired decomposition for medical video segmentation.

6 Acknowledgments

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

References

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

[2] 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.

[3] 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.

[4] 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.

[5] Long Zheng, Zhi Li, Weidong Wang, Zhenyu Dai, and Shuyun Li. E^3SAM2: Entropy-aware and edge-guided adaptation of SAM2 for echocardiography video segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13423–13431, 2026.

[6] Calvin F Nodine and Harold L Kundel. The cognitive side of visual search in radiology. In *Eye Movements from Physiology to Cognition*, pages 573–582. Elsevier, 1987.

[7] Harold L Kundel, Calvin F Nodine, Emily F Conant, and Susan P Weinstein. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*, 242(2):396–402, 2007.

[8] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6499–6507. IEEE, 2018.

[9] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385. IEEE, 2018.

[10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019.

[11] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an Atkinson–Shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[12] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024.

[13] 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.

[14] 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.

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

[16] Ghada Zamzmi, Sivaramakrishnan Rajaraman, Li-Yueh Hsu, Vandana Sachdev, and Sameer Antani. Real-time echocardiography image analysis and quantification of cardiac indices. *Medical Image Analysis*, 80: 102438, 2022.

[17] Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022.

[18] 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.

[19] Xiaolong Deng and Huisi Wu. Echocardiography video segmentation via neighborhood correlation mining. *IEEE Transactions on Medical Imaging*, 44(12):5172–5182, 2025.

[20] Jingxing Guo, Guilian Chen, Yimu Sun, Huisi Wu, and Jing Qin. Hierarchical spatiotemporal context aggregation and speckle-aware deformable convolution for echocardiography video segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3865–3874, 2025.

[21] Jingxing Guo, Guilian Chen, Yimu Sun, Huisi Wu, and Jing Qin. EchoVim: Making vision mamba docile for echocardiography video segmentation via dynamic interaction and semantic token-attentive refinement. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3261–3269, 2025.

[22] Yunpeng Fang, Yimu Sun, Jingxing Guo, Huisi Wu, and Jing Qin. Semi-supervised echocardiography video segmentation via anchor semantic awareness and continuous pseudo-label reforging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8535–8544, 2026.

[23] Christina Hafner, Vincenz Scharner, Martina Hermann, Philipp Metelka, Benedikt Hurch, Daniel Alexander Klaus, Wolfgang Schaubmayr, Michael Wagner, Andreas Gleiss, Harald Willschke, et al. Eye-tracking during simulation-based echocardiography: a feasibility study. *BMC Medical Education*, 23(1):490, 2023.

[24] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. SAM2-UNet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence*, 4(1):2, 2026.

[25] 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.

[26] 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.

[27] Edward D Folland, Alfred F Parisi, Paul F Moynihan, D Ray Jones, Charles L Feldman, and Donald E Tow. Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. a comparison of cineangiographic and radionuclide techniques. *Circulation*, 60(4):760–766, 1979.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[29] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–273. Springer, 2020.

[30] Jun Wei, Shuhui Wang, and Qingming Huang. F^3Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12321–12328, 2020.

[31] 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.

[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] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. XMem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023.

[34] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. VideoMamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024.

[35] Benedikt Alkin, Maximilian Beck, Korbinian Pöppel, Sepp Hochreiter, and Johannes Brandstetter. Vision-lstm: xlstm as generic vision backbone. In *International Conference on Learning Representations*, volume 2025, pages 23804–23826, 2025.

A Technical appendices and supplementary material

A.1 Experiments Continued

Clinical Metrics on EchoNet-Dynamic. Figure 9 complements the EchoNet-Dynamic results in Table 1. The regression plot illustrates the association between predicted and reference LV_{EF}, while the Bland–Altman plot shows the negative mean bias. The correlation is 0.842 and the bias±std is -2.20 ± 6.79 percentage points, indicating that accurate segmentation does not necessarily yield unbiased functional estimates.

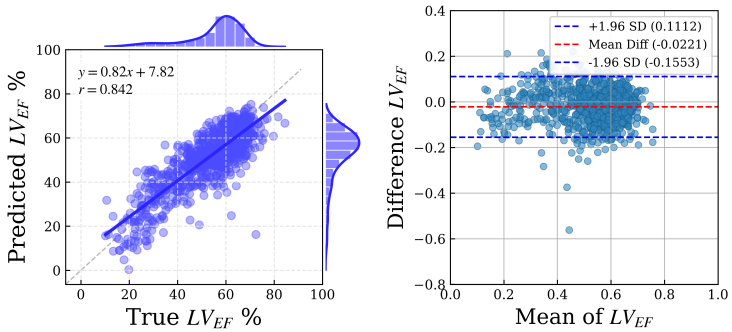


Figure 9: Linear regression and Bland–Altman plots for clinical metric LV_{EF} on EchoNet-Dynamic.

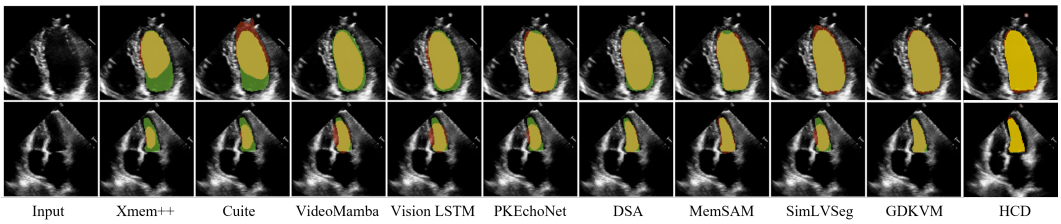


Figure 10: Visual comparison on the EchoNet-Dynamic test set. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

A.2 Additional Module Analysis

Ablation on Reference Gating in SIA. We vary the background suppression strength λ, where λ=0 keeps the reference unchanged and λ=1 denotes hard foreground gating. Hard foreground gating achieves the best performance in Table 4, supporting the use of foreground-focused reference retrieval in SIA.

Table 4: Ablation on SIA reference gating.

λ	Foreground weight	Background weight	Dice	HD95 (mm)
0.0	1.0	1.0	93.04	4.50
0.3	1.0	0.7	93.54	3.70
0.5	1.0	0.5	94.48	3.16
0.7	1.0	0.3	94.60	3.44
0.9	1.0	0.1	94.70	3.05
1.0	1.0	0.0	**95.21**	**2.59**

A.3 Reference-Update Ablation

We compare a sequence-fixed reference, an exponential moving average (EMA), and a feature-dissimilarity-gated update (FDGU) within the same HCD framework on CAMUS using 10-frame clips. All variants share the same reference retrieval architecture, AGP, SBP, decoder, temporal smoothing, and training settings; only the reference-update strategy differs.

Table 5: Reference-update strategies within HCD on CAMUS under the 10-frame clip setting. All other components and training settings are held fixed.

Reference strategy	mDice (%) ↑	HD95 (mm) ↓
EMA	93.53	3.96
FDGU	94.28	3.28
Fixed (HCD)	**95.21**	**2.59**

Let r_t^l and f_t^l denote the reference feature and current-frame feature at level $l \in \{3, 4\}$. All strategies are initialized from the annotated first frame, with $r_0^l = f_0^l$. The sequence-fixed strategy keeps $r_t^l = r_0^l$ throughout the clip. EMA updates the reference using a fixed interpolation coefficient:

$$r_t^l = \alpha r_{t-1}^l + (1 - \alpha) f_t^l, \tag{15}$$

where $\alpha \in [0, 1)$ controls the contribution of the previous reference feature.

FDGU instead derives the interpolation weight from the cosine dissimilarity between the previous reference and the current-frame feature:

$$d_t^l = \left[1 - \cos(r_{t-1}^l, f_t^l)\right]_+, \qquad g_t^l = 1 - \exp\left(-\frac{d_t^l}{\tau}\right), \tag{16}$$

followed by

$$r_t^l = r_{t-1}^l + g_t^l (f_t^l - r_{t-1}^l). \tag{17}$$

Here, $[x]_+ = \max(x, 0)$ and $\tau > 0$ controls gate sensitivity, with larger values producing more conservative updates. The corresponding reference mask follows the same interpolation strategy.

As shown in Table 5, the sequence-fixed reference achieves the highest mDice and lowest HD95 among the evaluated configurations. These results support maintaining a fixed target reference together with frame-dependent spatial refinement within HCD under the evaluated 10-frame CAMUS setting.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the three-stage cognitive decomposition framework (SIA, AGP, SBP), quantitative performance claims on CAMUS and EchoNet-Dynamic, and efficiency metrics (1.35M trainable parameters, 68 FPS). Limitations are discussed in Section 4.6.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 4.6 discusses two failure modes: under-segmentation under acoustic shadows on CAMUS, and degraded performance under extreme foreshortening on EchoNet-Dynamic where the first-frame reference becomes less representative. Section 4.5 evaluates joint multi-structure segmentation and reports its LV-accuracy trade-off; further accuracy improvements and non-periodic video applications are identified as future directions.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: This paper does not include theoretical results, proofs, or formal theorems. The contributions are architectural and empirical.

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: Section 4.2 provides full implementation details including backbone (frozen SAM2 Hiera), optimizer (AdamW, lr=1e-3), batch size (4), training epochs (100/50), input resolution (256×256/128×128), frame sampling strategy, and data augmentation. The experimental protocol strictly follows GDKVM [4]. Module designs are formally described with equations in Sections 3.3–3.5.

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: [No]

   Justification: Code will be released upon acceptance. Both datasets used (CAMUS and EchoNet-Dynamic) are publicly available. The paper provides sufficient architectural and implementation details in Section 4.2 and Sections 3.3–3.5 to enable reproduction of the main results.

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: Section 4.2 specifies all training details including GPU (NVIDIA A100), optimizer (AdamW), learning rate (1e-3), batch size (4), training epochs, input resolution, frame sampling, data augmentation protocol, inference smoothing (sliding-window average, window size 3), and the GDKVM experimental protocol followed for fair comparison.

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: [No]

   Justification: Segmentation results are reported as point estimates without error bars or significance tests. The multi-structure results in Section 4.5 are averaged over three training runs. The bias$\pm$std clinical metrics describe the spread of LVEF errors across test cases.

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: Section 4.2 specifies training on a single NVIDIA A100 GPU. The main inference speed (68 FPS) is measured on a single RTX 2080Ti and reported in Figure 5. Total parameters (35.2M) and trainable parameters (1.35M) are reported in Section 4.3.

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: This research uses only publicly available anonymized medical imaging datasets (CAMUS and EchoNet-Dynamic) that were collected with appropriate ethical approvals by their respective creators. No new data collection or human subjects research was conducted in this work.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The proposed method targets automated left ventricular segmentation for cardiac function assessment, with potential to improve diagnostic efficiency and consistency in clinical echocardiography. No foreseeable negative societal impacts are identified, as the method is designed for assistive rather than autonomous clinical decision-making.

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: This paper does not release new datasets or pre-trained models with high risk for misuse. The method is designed for a specific medical imaging task using existing public benchmarks.

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: CAMUS dataset is publicly available under its original research license (Leclerc et al., IEEE TMI 2019). EchoNet-Dynamic is publicly available under the Stanford University dataset license (Ouyang et al., Nature 2020). The SAM2 backbone is used under its Apache 2.0 license. All assets are properly cited in the references.

13. **New assets**

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

    Answer: [No]

    Justification: This paper does not introduce new datasets. The proposed HCD model will be released upon acceptance with training code and model checkpoints.

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: This paper does not involve crowdsourcing or direct research with human subjects. All experiments use existing publicly available anonymized echocardiography datasets.

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: This paper uses only existing publicly available anonymized datasets (CAMUS and EchoNet-Dynamic) that were collected and approved by their respective institutions. No new human subjects research was conducted.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were not used as any important, original, or non-standard component of the core methods in this research. Any LLM usage was limited to writing assistance and does not impact the core methodology or scientific contributions.