FREEECHO: LEARNING PROMPT-FREE ECHOCARDIOGRAPHY VIDEO SEGMENTATION VIA DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Echocardiography video segmentation delineates cardiac structures across frames to support quantitative assessment. Prompt-free models enable automated segmentation but remain limited in accuracy and generalization across data sources, whereas prompt-based models achieve stronger performance by relying on external localization inputs that hinder scalable inference. We present FreeEcho, a knowledge distillation framework that converts existing prompt-based echocardiography video segmentation models into prompt-free counterparts. FreeEcho adapts the model prediction pipeline, combines ground-truth supervision with teacher-derived distillation signals, and employs a prompt removal curriculum during training. Experiments across multiple models and public datasets show that the distilled students consistently approach or achieve their teachers' performance while substantially outperforming direct prompt removal. They also outperform existing prompt-free models, with clear advantages in generalization across different data sources. Ablation studies clarify the effects of distillation design choices.

1 INTRODUCTION

Echocardiography videos provide a noninvasive view of cardiac anatomy and motion for assessing cardiac function (Lang et al., 2015; Ouyang et al., 2020). Their segmentation delineates cardiac structures across frames, providing essential spatial information for quantifying cardiac anatomy and its dynamic changes. However, manual delineation is time-consuming and difficult to scale, motivating automated segmentation for efficient and consistent analysis. In practical applications, segmentation models must achieve both high accuracy and reliable generalization to diverse data sources. Variations in factors such as scanners and patient populations can substantially alter video appearance and degrade segmentation performance (Chen et al., 2023; Reddy et al., 2023).

Existing echocardiography video segmentation models can be broadly distinguished by whether inference requires external localization prompts. Prompt-free models predict cardiac masks directly from videos, enabling automated analysis (Ouyang et al., 2020; Wang et al., 2025), but achieving both high segmentation accuracy and reliable generalization across data sources remains challenging (Chen et al., 2023). In contrast, prompt-based models use localization inputs, such as points or bounding boxes, to identify the target structure and typically achieve stronger segmentation performance (Hu et al., 2025; Deng et al., 2024; Ma et al., 2025). However, their reliance on additional prompts increases user interaction and limits large-scale automated analysis, while directly removing these prompts can substantially degrade performance. Knowledge distillation transfers teacher knowledge to a student through predictions or intermediate representations (Hinton et al., 2015; Lopez-Paz et al., 2015). This paradigm is well suited to our setting, where a high-performance prompt-based teacher provides supervision to a student that learns to localize and segment targets without prompts. Therefore, distillation transfers the teacher's strong segmentation capability while removing the need for external prompts at inference.

In this paper, we identify a clear gap between prompt-free and prompt-based echocardiography video segmentation through a systematic evaluation of representative models: prompt-free models exhibit limited cross-source generalization, while prompt-based models suffer substantial performance degradation once prompts are removed. To address this gap, we propose FreeEcho, a distillation framework that effectively converts high-performance prompt-based echocardiography video segmentation models into prompt-free counterparts. FreeEcho introduces architecture-specific

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

segmentation adaptation to enable prompt-free inference, integrates ground-truth masks with complementary teacher-derived supervision for effective knowledge transfer, and employs a prompt-removal curriculum for stable distillation. After distillation, the resulting student performs echocardiography video segmentation without external prompts. Extensive experiments across multiple models and public datasets show that FreeEcho consistently converts different prompt-based models into prompt-free counterparts while maintaining strong segmentation performance and cross-source generalization, demonstrating its effectiveness and generality. The resulting models also outperform existing prompt-free models. Further ablation studies validate the key design choices of the framework. Our main contributions can be summarized as follows:

- We systematically evaluate multiple echocardiography video segmentation models and identify a clear gap between strong prompt-based segmentation capability and prompt-free deployment: existing prompt-free models exhibit limited segmentation performance, particularly in generalization across different data sources, while directly removing prompts from prompt-based models leads to substantial performance degradation.
- We introduce FreeEcho, a distillation framework that transforms prompt-based echocardiography video segmentation models into prompt-free counterparts through prompt-free segmentation adaptation, complementary teacher-derived supervision, and a prompt-removal curriculum.
- We conduct extensive experiments across multiple public echocardiography video datasets and segmentation models. The distilled students consistently approach or achieve their prompt-based teachers' performance and outperform direct prompt removal. Furthermore, they can achieve superior segmentation accuracy and generalization capabilities compared to existing prompt-free methods. Ablation studies further clarify the effects of key distillation choices.

2 Related Work

2.1 Echocardiography Video Segmentation

2.1.1 Prompt-free and Prompt-based Segmentation

Prompt-free echocardiography video segmentation models predict cardiac masks without external localization cues. Some methods focus on image segmentation and can be applied frame-by-frame to videos (Ronneberger et al., 2015; Leclerc et al., 2019; Zhu et al., 2024). More recent approaches extend segmentation to videos by exploiting temporal information across frames (Wu et al., 2022; 2023; Lin et al., 2024a; Wang et al., 2025; 2026). Although their fully automated inference supports large-scale analysis, their segmentation performance remains limited, particularly in generalization across different data sources (Chen et al., 2023). In contrast, prompt-based models rely on manually provided localization prompts, such as points or bounding boxes, to guide segmentation (Deng et al., 2024; Hu et al., 2025; Ma et al., 2025). Building on SAM (Kirillov et al., 2023) and its ultrasound adaptations (Lin et al., 2023; 2024b), recent methods extend promptable segmentation to echocardiography videos (Deng et al., 2024; Hu et al., 2025).

2.1.2 Per-frame and Memory-based Segmentation

From the perspective of temporal modeling, echocardiography video segmentation methods follow two paradigms: per-frame prediction and memory-based propagation. Per-frame models process each frame independently without explicitly modeling temporal dependencies (Leclerc et al., 2019; Ouyang et al., 2020; Hu et al., 2025), whereas memory-based models retain information from previous frames to guide subsequent predictions through temporal states or memory retrieval (Wang et al., 2025; 2026; Deng et al., 2024; Ravi et al., 2025; Ma et al., 2025; Zheng et al., 2026). This architectural distinction is critical for knowledge distillation, as effective knowledge transfer must account for how each paradigm represents and propagates task-relevant information, motivating their separate consideration in FreeEcho. Despite different temporal mechanisms, prompt-based models in both settings still rely on external prompts, limiting fully automated segmentation.

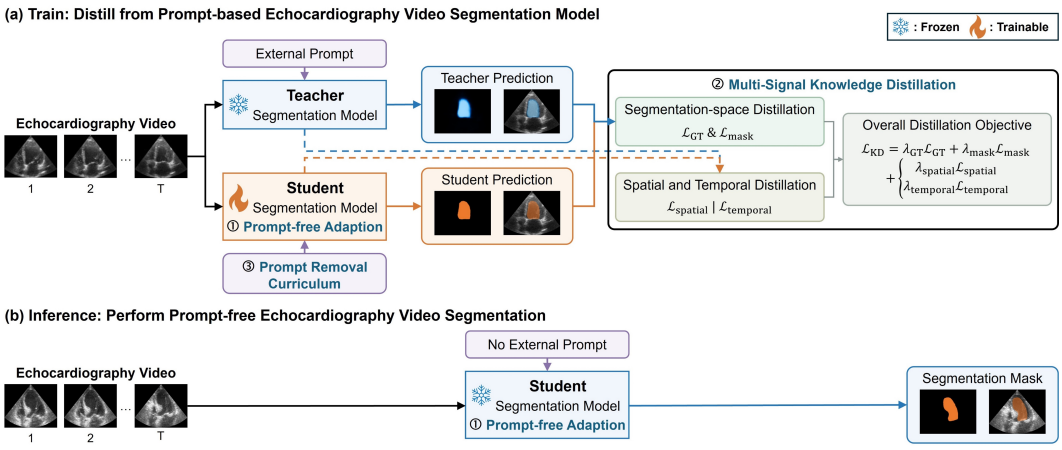


Figure 1: Overview of FreeEcho. (a) The student is trained by distilling a frozen prompt-based teacher, incorporating prompt-free adaptation, multi-signal supervision, and the prompt-removal curriculum. (b) The distilled student segments echocardiography videos without external prompts.

2.2 Knowledge Distillation

Knowledge distillation transfers knowledge from a teacher to a student through supervisory signals such as predictive distributions or intermediate representations (Hinton et al., 2015; Zagoruyko & Komodakis, 2017; Romero et al., 2015; Liu et al., 2019). Beyond model compression, it can also operate across different input conditions, where the teacher has access to information unavailable to the student (Lopez-Paz et al., 2015; Gupta et al., 2016). In this setting, the teacher and student may share the same architecture but differ in their available inputs, allowing distillation to transfer capabilities rather than reduce model size. Related works have used learned or self-generated prompts to reduce user interaction in medical image segmentation (Tang et al., 2026; Xu et al., 2024), while other studies distill foundation-model knowledge for semi-supervised segmentation (Huang et al., 2025; Lu et al., 2026; Zhao et al., 2025). However, prompt-generation methods introduce additional modules and still rely on the model's original prompt-dependent inference pipelines, while semi-supervised distillation primarily aims to address data scarcity rather than eliminate the prompt dependency. FreeEcho instead investigates eliminating the dependence of prompt-based echocardiography video segmentation models on external prompt inputs through knowledge distillation.

3 Method

3.1 Problem Formulation and Framework Overview

Given an L-frame echocardiography video $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_{L-1})$, the prompt-based teacher $F^{\mathrm{T}}_{\theta^{\mathrm{T}}}$ predicts segmentation masks using external prompts $\mathbf{P}$. FreeEcho distills the teacher into a student $F^{\mathrm{S}}_{\theta^{\mathrm{S}}}$ that performs prompt-free segmentation at inference:

$$F^{\mathrm{T}}_{\theta^{\mathrm{T}}}(\mathbf{X}, \mathbf{P}) \xrightarrow{\text{Distill}} F^{\mathrm{S}}_{\theta^{\mathrm{S}}}(\mathbf{X}, \varnothing)\,. \tag{1}$$

During distillation at epoch e, the teacher receives external prompts $\mathbf{P}$, whereas the student receives the curriculum-controlled prompt condition $\mathbf{P}^{\mathrm{S}}_e$. Let $\mathcal{D}_{\mathrm{train}}$ denote the distribution of training tuples $(\mathbf{X}, \mathbf{P}, \mathbf{Y}_{\mathcal{K}})$, where $\mathbf{Y}_{\mathcal{K}} = \{\mathbf{y}_t \mid t \in \mathcal{K}\}$ contains the ground-truth masks for annotated frames indexed by $\mathcal{K}$. The distillation problem is formulated as:

$$\theta^{\mathrm{S}*} = \arg\min_{\theta^{\mathrm{S}}} \mathbb{E}_{(\mathbf{X},\mathbf{P},\mathbf{Y}_{\mathcal{K}})\sim\mathcal{D}_{\mathrm{train}}} \left[\mathcal{L}_{\mathrm{KD}}\left(F^{\mathrm{S}}_{\theta^{\mathrm{S}}}(\mathbf{X}, \mathbf{P}^{\mathrm{S}}_e), \mathrm{sg}\left[F^{\mathrm{T}}_{\theta^{\mathrm{T}}}(\mathbf{X}, \mathbf{P})\right], \mathbf{Y}_{\mathcal{K}}\right)\right], \tag{2}$$

where $\mathcal{L}_{\mathrm{KD}}$ denotes the overall distillation objective for transferring the teacher's segmentation knowledge to the student. The operator sg stops gradients, such that only θ^{S} is optimized.

As illustrated in Figure 1, we design FreeEcho, a distillation framework that transforms high-performance prompt-based echocardiography video segmentation models into prompt-free counterparts. FreeEcho introduces an architecture-specific adaptation that enables different models to operate in a prompt-free manner and establishes a foundation for the distillation process (Section 3.2).

It combines ground-truth masks with multiple complementary teacher-derived supervision signals to facilitate effective knowledge transfer during distillation (Section 3.3). It also incorporates a prompt-removal curriculum that gradually withdraws student prompts during distillation for stable knowledge transfer (Section 3.4).

3.2 Prompt-Free Segmentation Adaptation

Prompt-based echocardiography video segmentation models are not directly compatible with prompt-free inference because localization prompts participate in their prediction pipeline. FreeEcho introduces adaptations to these models while preserving their original segmentation pipeline. Since per-frame and memory-based models incorporate prompt information differently during segmentation, separate adaptations are required.

Per-frame models. A per-frame model independently predicts the mask of each frame from the current image and an external prompt. FreeEcho preserves the original model architecture while invoking its prompt processing pipeline without external inputs. The resulting prompt-free representation is incorporated into the original prediction process, allowing the student to infer the target location and boundaries directly from each frame. The prompt-free prediction is formulated as:

$$\hat{\mathbf{y}}_t^{\mathrm{S}} = F_{\theta^{\mathrm{S}}}^{\mathrm{S}}(\mathbf{x}_t, \varnothing), \quad t \in \{0, 1, \ldots, L-1\}. \tag{3}$$

Memory-based models. Unlike per-frame models, memory-based models explicitly maintain temporal memory from previously processed frames to guide segmentation of subsequent frames. Therefore, prompt-free adaptation must preserve both the prediction of the current frame and the initialization and update of temporal memory. Naively replacing prompts with a fully empty input can disrupt the original process and cause errors. FreeEcho introduces a memory-preserving prompt-free interface for this setting. It replaces external inputs with a model-defined prompt-free representation $\phi_{\varnothing}^{\mathrm{S}}$, while preserving the model's original pipeline. On the initial frame, the student predicts a mask without memory and uses this prediction to construct the memory state. For each subsequent frame, the student combines the current frame, the propagated memory, and the same prompt-free representation to predict the mask and update the memory. Let $\mathbf{m}_t^{\mathrm{S}}$ denote the student memory after processing frame t, and $U_{\theta^{\mathrm{S}}}$ denote the memory-update operation. The memory update and prompt-free prediction for memory-based students are formulated as:

$$\mathbf{m}_t^{\mathrm{S}} = U_{\theta^{\mathrm{S}}}\left(\mathbf{m}_{t-1}^{\mathrm{S}}, \mathbf{x}_t, \hat{\mathbf{y}}_t^{\mathrm{S}}\right), \quad \hat{\mathbf{y}}_t^{\mathrm{S}} = F_{\theta^{\mathrm{S}}}^{\mathrm{S}}\left(\mathbf{x}_t, \mathbf{m}_{t-1}^{\mathrm{S}}, \phi_{\varnothing}^{\mathrm{S}}\right), \quad t = 1, \ldots, L-1. \tag{4}$$

3.3 Multi-Signal Knowledge Distillation

The teacher and student perform segmentation under different input conditions: the teacher receives external prompts, whereas the student must infer the target directly from the video. To facilitate effective distillation under this input asymmetry, FreeEcho employs a weighted combination of segmentation-space distillation with auxiliary spatial and temporal distillation.

3.3.1 Segmentation-Space Distillation

Segmentation-space distillation provides a reliable foundation for transferring the teacher's predictive behavior to the student, while preserving consistency with the underlying segmentation task. Therefore, FreeEcho employs two complementary distillation objectives in the segmentation space, including ground-truth segmentation supervision and teacher-derived soft-mask supervision.

Ground-truth supervision anchors student training to the target segmentation objective and reduces the influence of potentially erroneous teacher-derived signals. Rather than averaging supervision uniformly across annotated frames, FreeEcho employs frame-dependent weighting that imposes stronger supervision on earlier frames, improving the distilled student's segmentation performance, particularly for memory-based models whose later predictions depend strongly on early-frame behavior. With w_t^{GT} denoting the supervision weight for annotated frame t, the ground-truth segmentation loss is defined as:

$$\mathcal{L}_{\mathrm{GT}} = \frac{1}{\sum_{t \in \mathcal{K}} w_t^{\mathrm{GT}}} \sum_{t \in \mathcal{K}} w_t^{\mathrm{GT}} \left[\mathcal{L}_{\mathrm{BCE}}\left(\mathbf{z}_t^{\mathrm{S}}, \mathbf{y}_t\right) + \mathcal{L}_{\mathrm{Dice}}\left(\sigma(\mathbf{z}_t^{\mathrm{S}}), \mathbf{y}_t\right)\right], \tag{5}$$

where $\mathbf{z}_t^{\mathrm{S}}$ denotes the student logit map and $\hat{\mathbf{y}}_t^{\mathrm{S}} = \sigma(\mathbf{z}_t^{\mathrm{S}})$ denotes the probability map. $\mathcal{L}_{\mathrm{BCE}}$ denotes binary cross-entropy with logits, which provides pixel-wise supervision. $\mathcal{L}_{\mathrm{Dice}}$ denotes the soft Dice loss, which promotes region-level overlap between the prediction and the ground truth.

Teacher-derived soft-mask supervision complements ground-truth supervision, which is limited to sparsely annotated frames and discrete binary targets. It extends supervision to all frames in the input video, including those without ground-truth masks. The teacher's soft foreground probabilities further encode graded information about target localization and boundary uncertainty, helping the student reproduce the teacher's strong segmentation behavior without external prompts. Let $\mathbf{z}_t^{\mathrm{T}}$ and $\mathbf{z}_t^{\mathrm{S}}$ denote the teacher and student foreground logits on a common output grid Ω. The teacher-derived soft-mask distillation loss is averaged over all video frames:

$$\mathcal{L}_{\mathrm{mask}} = \frac{\tau_{\mathrm{S}}^2}{L} \sum_{t=0}^{L-1} \mathcal{L}_{\mathrm{BCE}} \left(\frac{\mathbf{z}_t^{\mathrm{S}}}{\tau_{\mathrm{S}}}, \mathrm{sg} \left(\sigma \left(\frac{\mathbf{z}_t^{\mathrm{T}}}{\tau_{\mathrm{T}}} \right) \right) \right), \tag{6}$$

where τ_{T} and τ_{S} are the teacher and student temperatures, respectively.

3.3.2 Auxiliary Spatial and Temporal Distillation

Beyond segmentation-space distillation, FreeEcho incorporates auxiliary spatial and temporal supervision tailored to different echocardiography video segmentation architectures, providing complementary guidance and helping the student to learn prompt-free segmentation. For per-frame segmentation models, intermediate spatial representations can provide additional supervision beyond the final segmentation output. FreeEcho introduces coarse-mask supervision as an auxiliary spatial distillation signal specifically for per-frame models. These representations encode the approximate location and spatial extent of the target cardiac structure, providing the student with direct guidance on the teacher's intermediate localization behavior. With $\mathbf{z}_{t,c}^{\mathrm{T}}$ and $\mathbf{z}_{t,c}^{\mathrm{S}}$ denoting the teacher and student coarse logit maps on the coarse output grid Ω_c, the spatial distillation loss is defined as:

$$\mathcal{L}_{\mathrm{spatial}} = \frac{\tau_{\mathrm{S}}^2}{L} \sum_{t=0}^{L-1} \mathcal{L}_{\mathrm{BCE}} \left(\frac{\mathbf{z}_{t,c}^{\mathrm{S}}}{\tau_{\mathrm{S}}}, \mathrm{sg} \left(\sigma \left(\frac{\mathbf{z}_{t,c}^{\mathrm{T}}}{\tau_{\mathrm{T}}} \right) \right) \right). \tag{7}$$

Unlike per-frame models, memory-based models primarily rely on temporal representations stored in memory to condition predictions on subsequent frames, which cannot be effectively transferred to the prompt-free student via the previous supervision. Therefore, FreeEcho transfers the teacher's temporal segmentation behavior by aligning the object-level memory representations produced after the initial frame. Let $\mathbf{o}_t^{\mathrm{T}}$ and $\mathbf{o}_t^{\mathrm{S}}$ denote the teacher and student object-level memory representations for frame t. The temporal distillation loss is:

$$\mathcal{L}_{\mathrm{temporal}} = \frac{1}{L-1} \sum_{t=1}^{L-1} \left[1 - \cos \left(\mathbf{o}_t^{\mathrm{S}}, \mathrm{sg}(\mathbf{o}_t^{\mathrm{T}}) \right) \right], \tag{8}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. The initial frame is excluded because the teacher and student initialize target propagation under different prompt conditions.

3.3.3 Weighted Distillation Objective

The overlall distillation objective $\mathcal{L}_{\mathrm{KD}}$ is instantiated as:

$$\begin{aligned} \mathcal{L}_{\mathrm{KD}} &= \lambda_{\mathrm{GT}} \mathcal{L}_{\mathrm{GT}} + \lambda_{\mathrm{mask}} \mathcal{L}_{\mathrm{mask}} + \mathcal{L}_{\mathrm{aux}}, \\ \mathcal{L}_{\mathrm{aux}} &= \begin{cases} \lambda_{\mathrm{spatial}} \mathcal{L}_{\mathrm{spatial}}, & \text{per-frame models}, \\ \lambda_{\mathrm{temporal}} \mathcal{L}_{\mathrm{temporal}}, & \text{memory-based models}, \end{cases} \end{aligned} \tag{9}$$

where the nonnegative coefficients λ_{GT}, λ_{mask}, $\lambda_{\mathrm{spatial}}$, and $\lambda_{\mathrm{temporal}}$ denote the effective weights of their respective losses. The combined objective jointly anchors student predictions to ground-truth masks and distills the teacher's segmentation knowledge through a set of specialized supervision signals tailored to the underlying mechanisms of echocardiography video segmentation models.

Table 1: Comparison of echocardiography video segmentation performance under zero-shot transfer. Dice (mean ± SD) is reported, with the best and second-best results within each group shown in **bold** and underlined.

Method	Prompt	Source	Zero-shot Transfer		
		CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Prompt-free Models					
U-Net	No	0.920 ± 0.041	0.732 ± 0.107	0.655 ± 0.137	0.485 ± 0.286
DeepLabV3+	No	0.889 ± 0.048	0.525 ± 0.196	0.312 ± 0.202	0.340 ± 0.269
EchoNet CNN	No	0.764 ± 0.131	0.561 ± 0.124	0.605 ± 0.145	0.219 ± 0.102
nnU-Net	No	0.924 ± 0.046	0.615 ± 0.239	0.600 ± 0.242	0.030 ± 0.096
GDKVM	No	0.924 ± 0.033	0.855 ± 0.090	0.818 ± 0.131	0.515 ± 0.346
OSA	No	$\mathbf{0.926 \pm 0.039}$	$\mathbf{0.869 \pm 0.059}$	$\mathbf{0.839 \pm 0.105}$	$\mathbf{0.527 \pm 0.320}$
Per-frame Model: EchoONE					
Teacher	Yes	$\mathbf{0.939 \pm 0.032}$	$\mathbf{0.898 \pm 0.055}$	$\mathbf{0.877 \pm 0.080}$	$\mathbf{0.673 \pm 0.244}$
Direct prompt removal	No	0.894 ± 0.057	0.853 ± 0.100	0.808 ± 0.148	0.284 ± 0.288
DistillSAM	No	0.928 ± 0.033	0.885 ± 0.071	0.833 ± 0.184	0.596 ± 0.300
KnowSAM	No	0.925 ± 0.041	0.873 ± 0.077	0.853 ± 0.112	0.546 ± 0.254
FreeEcho (ours)	No	0.937 ± 0.036	$\mathbf{0.898 \pm 0.059}$	0.867 ± 0.112	0.614 ± 0.284
Memory-based Model: MedSAM2					
Teacher	Yes	$\mathbf{0.937 \pm 0.032}$	$\mathbf{0.900 \pm 0.050}$	$\mathbf{0.884 \pm 0.067}$	$\mathbf{0.739 \pm 0.178}$
Direct prompt removal	No	0.000 ± 0.000	0.000 ± 0.006	0.000 ± 0.004	0.009 ± 0.071
DistillSAM	No	0.918 ± 0.036	0.867 ± 0.098	0.848 ± 0.107	0.458 ± 0.339
KnowSAM	No	0.915 ± 0.042	0.873 ± 0.079	0.854 ± 0.101	0.505 ± 0.309
FreeEcho (ours)	No	0.929 ± 0.039	0.878 ± 0.073	0.863 ± 0.102	0.535 ± 0.257

3.4 Prompt Removal Curriculum

Prompts provide strong target localization guidance during segmentation, making their direct removal during distillation challenging. FreeEcho addresses this discrepancy through a prompt-removal curriculum that gradually reduces the probability of providing prompts to the student during distillation, allowing the student to increasingly infer target-relevant information directly from the video while maintaining stable knowledge transfer. At epoch e, the student prompt is sampled as:

$$\mathbf{P}_e^{\mathrm{S}} = \begin{cases} \widetilde{\mathbf{P}}, & \text{if } b_e = 1, \\ \varnothing, & \text{if } b_e = 0, \end{cases} \qquad b_e \sim \mathrm{Bernoulli}\left(\max\left(0, 1 - \frac{e}{E_a}\right)\right), \tag{10}$$

where $\widetilde{\mathbf{P}}$ is the teacher-derived candidate prompt, and E_a is the number of annealing epochs. Its sampling probability gradually decreases to zero, after which training proceeds entirely without prompts. As prompts are withdrawn, the student increasingly relies on prompt-free prediction, while the prompt-based teacher continues to provide distillation supervision.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate FreeEcho on four established echocardiography video segmentation datasets. CAMUS (Leclerc et al., 2019) contains 1,000 videos from 500 patients. EchoNet-Dynamic (EchoNet-Dyn) (Ouyang et al., 2020) comprises 10,030 adult videos. EchoNet-Pediatric (EchoNet-Ped) (Reddy et al., 2023) contains 7,643 pediatric videos, while CardiacUDA (Yang et al., 2023) comprises 992 videos from two hospitals using different scanners. Their variation in patient populations, institutions, and acquisition devices provides a diverse and representative testing platform.

Models. We select the prompt-based models EchoONE (Hu et al., 2025) and MedSAM2 (Ma et al., 2025) as distillation teachers, which are representative, state-of-the-art per-frame and memory-based

Table 2: Comparison of echocardiography video segmentation performance under in-domain training. Dice (mean ± SD) is reported, with the best and second-best results within each group shown in **bold** and underlined.

Method	Prompt	In-domain Training			
		CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Prompt-free Models					
U-Net	No	0.920 ± 0.041	0.891 ± 0.047	0.896 ± 0.055	0.828 ± 0.137
DeepLabV3+	No	0.889 ± 0.048	0.902 ± 0.057	0.898 ± 0.067	0.832 ± 0.133
EchoNet CNN	No	0.764 ± 0.131	0.907 ± 0.047	0.671 ± 0.145	0.619 ± 0.283
nnU-Net	No	0.924 ± 0.046	0.909 ± 0.045	0.894 ± 0.081	0.838 ± 0.138
GDKVM	No	0.924 ± 0.033	$\mathbf{0.911 \pm 0.043}$	0.900 ± 0.056	0.838 ± 0.153
OSA	No	$\mathbf{0.926 \pm 0.039}$	0.909 ± 0.051	$\mathbf{0.903 \pm 0.060}$	$\mathbf{0.842 \pm 0.141}$
Per-frame Model: EchoONE					
Teacher	Yes	$\mathbf{0.939 \pm 0.032}$	$\mathbf{0.926 \pm 0.038}$	$\mathbf{0.908 \pm 0.049}$	$\mathbf{0.873 \pm 0.081}$
Direct prompt removal	No	0.894 ± 0.057	0.684 ± 0.268	0.838 ± 0.130	0.000 ± 0.000
DistillSAM	No	0.928 ± 0.033	0.901 ± 0.062	0.872 ± 0.089	0.781 ± 0.285
KnowSAM	No	0.925 ± 0.041	0.919 ± 0.043	0.900 ± 0.070	0.842 ± 0.130
FreeEcho (ours)	No	0.937 ± 0.036	$\mathbf{0.926 \pm 0.039}$	$\mathbf{0.908 \pm 0.060}$	0.867 ± 0.116
Memory-based Model: MedSAM2					
Teacher	Yes	$\mathbf{0.937 \pm 0.032}$	$\mathbf{0.930 \pm 0.035}$	$\mathbf{0.918 \pm 0.043}$	$\mathbf{0.881 \pm 0.077}$
Direct prompt removal	No	0.000 ± 0.000	0.001 ± 0.026	0.000 ± 0.000	0.004 ± 0.033
DistillSAM	No	0.918 ± 0.036	0.912 ± 0.031	0.892 ± 0.074	0.839 ± 0.154
KnowSAM	No	0.915 ± 0.042	0.911 ± 0.033	0.888 ± 0.075	0.392 ± 0.426
FreeEcho (ours)	No	0.929 ± 0.039	0.919 ± 0.028	0.903 ± 0.059	0.857 ± 0.140

models. The existing prompt-free models included in our comparison span representative general segmentation models U-Net (Ronneberger et al., 2015) and DeepLabV3+ (Chen et al., 2018), the medical segmentation model nnU-Net (Isensee et al., 2021), and echocardiography video segmentation models EchoNet CNN (Ouyang et al., 2020), GDKVM (Wang et al., 2025), and OSA (Wang et al., 2026).

Evaluation. We systematically evaluate whether FreeEcho transfers the segmentation capability of prompt-based teachers to prompt-free students across architectures and data sources. For zero-shot transfer, all compared models are trained or adapted on CAMUS, aligning their task-specific training source for a fair cross-source comparison without target-data training or checkpoint selection. For in-domain evaluation, models are trained separately on each dataset and evaluated on its held-out test set. We also compare FreeEcho with existing distillation-based methods, DistillSAM (Tang et al., 2026) and KnowSAM (Huang et al., 2025), which also aim to reduce reliance on external prompts. Segmentation accuracy is quantified by the Dice similarity coefficient (higher is better), reported as mean ± standard deviation across annotated frames.

4.2 Echocardiography Video Segmentation Performance

4.2.1 Segmentation Performance under Zero-Shot Transfer

Our main experiment evaluates the effectiveness and generality of FreeEcho while providing a systematic and controlled comparison of existing prompt-free and prompt-based echocardiography video segmentation models. Table 1 reports segmentation performance across four different datasets. It includes representative prompt-free models and two SOTA prompt-based teacher models that instantiate per-frame and memory-based segmentation model architectures. For each teacher architecture, we compare the original prompt-based teacher with prompt-free configurations obtained through direct prompt removal, two existing distillation-based methods, and FreeEcho.

As shown in Table 1, FreeEcho consistently outperforms direct prompt removal and current distillation-based methods for both segmentation architectures across all evaluation data sources.

Table 3: Ablation study on multi-signal distillation supervision in FreeEcho. Dice (mean ± SD) is reported, with the best result within each group shown in **bold**.

Configuration	Source	Zero-shot Transfer		
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Per-frame Model: EchoONE				
$\mathcal{L}_{\mathbf{KD}}$ **(default)**	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.898 \pm 0.059}$	$\mathbf{0.867 \pm 0.112}$	$\mathbf{0.614 \pm 0.284}$
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{GT}}\mathcal{L}_{\mathrm{GT}}$	0.930 ± 0.035	0.888 ± 0.065	0.856 ± 0.105	0.582 ± 0.299
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}}$	0.932 ± 0.031	0.891 ± 0.069	0.832 ± 0.174	0.606 ± 0.289
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{spatial}}\mathcal{L}_{\mathrm{spatial}}$	0.926 ± 0.041	0.869 ± 0.080	0.851 ± 0.115	0.552 ± 0.253
Memory-based Model: MedSAM2				
$\mathcal{L}_{\mathbf{KD}}$ **(default)**	$\mathbf{0.929 \pm 0.039}$	$\mathbf{0.878 \pm 0.073}$	$\mathbf{0.863 \pm 0.102}$	$\mathbf{0.535 \pm 0.257}$
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{GT}}\mathcal{L}_{\mathrm{GT}}$	0.920 ± 0.045	0.870 ± 0.095	0.849 ± 0.110	0.496 ± 0.332
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}}$	0.924 ± 0.043	0.875 ± 0.086	0.857 ± 0.099	0.477 ± 0.344
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{temporal}}\mathcal{L}_{\mathrm{temporal}}$	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.039 ± 0.159

Furthermore, FreeEcho can produce prompt-free students that largely preserve the strong segmentation performance of their prompt-based teachers. Directly removing the prompt deprives the prompt-based model of the guidance information it relies on for segmentation, leading to a significant performance degradation; in contrast, FreeEcho enables the student to gradually learn to perform segmentation without relying on the prompt during distillation, thereby achieving superior results. Additionally, FreeEcho specifically incorporates multi-signal distillation tailored for current echocardiography video segmentation models, enabling it to outperform image-oriented DistillSAM and KnowSAM. These results demonstrate that FreeEcho effectively transfers the segmentation knowledge of prompt-based teachers to prompt-free students, outperforming related methods while avoiding the severe degradation induced by directly removing prompts.

Table 1 further illustrates that existing prompt-free models remain less accurate than prompt-based models and exhibit limited generalization across different data sources, with segmentation performance degrading more significantly compared to prompt-based models. In contrast, the two prompt-free models obtained via FreeEcho outperform all listed prompt-free methods, particularly under zero-shot evaluation on different data sources. This comparison demonstrates that FreeEcho provides an effective and feasible approach for obtaining high-performance prompt-free echocardiography video segmentation models.

4.2.2 Segmentation Performance under In-Domain Training

We further evaluate FreeEcho in the in-domain training setting. Table 2 shows that FreeEcho remains the most effective approach for prompt-free conversion among the compared methods, consistently outperforming direct prompt removal and existing distillation-based methods across both segmentation architectures and all four datasets. The resulting prompt-free students also achieve segmentation performance closer to that of their original prompt-based teachers. Moreover, the students distilled via FreeEcho consistently exceed current prompt-free models across different datasets. These findings further demonstrate the effectiveness of FreeEcho for obtaining high-performing prompt-free echocardiography video segmentation models under dataset-specific training.

4.3 Ablation Studies on Distillation Design Choices

4.3.1 Effectiveness of Distillation Supervision

To evaluate the effectiveness of the proposed multi-signal distillation and assess the contribution of each supervision signal, we compare the full objective in Equation 9 with variants that remove one signal at a time. Table 3 reports the resulting prompt-free segmentation performance under zero-shot transfer. As shown in Table 3, the full distillation objective achieves the highest accuracy for both architectures on every dataset. Removing either ground-truth or teacher-derived soft-mask supervision degrades performance because the two signals address different gaps: ground-truth masks

Table 4: Ablation study on the prompt removal curriculum in FreeEcho. Dice (mean ± SD) is reported, with the best result within each group shown in **bold**.

Configuration	Source	Zero-shot Transfer		
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Per-frame Model: EchoONE				
Linear annealing (default)	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.898 \pm 0.059}$	$\mathbf{0.867 \pm 0.112}$	$\mathbf{0.614 \pm 0.284}$
Adaptive prompt gating	0.928 ± 0.041	0.887 ± 0.069	0.863 ± 0.105	0.575 ± 0.255
Prompt-free throughout	0.924 ± 0.043	0.880 ± 0.074	0.862 ± 0.102	0.564 ± 0.256
Memory-based Model: MedSAM2				
Linear annealing (default)	$\mathbf{0.929 \pm 0.039}$	$\mathbf{0.878 \pm 0.073}$	$\mathbf{0.863 \pm 0.102}$	$\mathbf{0.535 \pm 0.257}$
Adaptive prompt gating	0.914 ± 0.065	0.872 ± 0.098	0.851 ± 0.116	0.515 ± 0.339
Prompt-free throughout	0.917 ± 0.039	0.867 ± 0.095	0.850 ± 0.112	0.499 ± 0.327

anchor predictions to annotated cardiac structures, while soft masks transfer the teacher's localization and boundary information across all frames. The per-frame model suffers its largest decline under the spatial distillation signal ablation because removing it prevents the student from learning the teacher's intermediate behavior, making it difficult to reproduce the teacher's segmentation performance. In the memory-based model, removing temporal distillation prevents the student from learning the important memory mechanism used to guide segmentation, causing errors within its memory to propagate to subsequent frames and thus sharply degrading performance. These results validate the effectiveness of multi-signal distillation and confirm the need to retain complementary teacher-derived supervision signals during distillation, particularly the auxiliary signals.

4.3.2 Effects of the Prompt Removal Curriculum

We further conduct an ablation study to assess the effectiveness of our prompt-removal curriculum. Linear annealing in Equation 10 is compared with training without prompts throughout and adaptive prompt gating, in which a threshold on model feedback determines whether the student receives a prompt. All students are evaluated without external prompts under the zero-shot transfer setting. Table 4 shows that the prompt-free models obtained via distillation using the linear annealing strategy achieve the best segmentation performance across different datasets. This advantage can be attributed to the progressive transition: as distillation proceeds, the gradual removal of prompts encourages students to progressively learn to localize targets and perform segmentation without relying on prompts. In contrast, distilling without prompts throughout provides no initial localization guidance, while adaptive gating does not ensure a progressive transition.

5 Conclusion

In this paper, we present FreeEcho, a knowledge distillation framework that converts existing high-performance prompt-based echocardiography video segmentation models into prompt-free counterparts. FreeEcho combines prompt-free segmentation adaptation, multi-signal knowledge distillation, and a prompt-removal curriculum to support models with different architectures. Experiments across multiple datasets show that the distilled students retain much of their teachers' segmentation capability while consistently outperforming direct prompt removal and existing distillation-based methods. FreeEcho also produces stronger prompt-free models, with particularly clear advantages over existing prompt-free methods when applied to echocardiography videos from different data sources. Ablation studies demonstrate the benefits of complementary supervision signals and progressive prompt removal, providing practical guidance for adapting prompt-based models to automated echocardiography video segmentation.

AI USE STATEMENT

We used generative AI tools to polish the manuscript's English by suggesting revisions to grammar, phrasing, and readability. We used these tools to identify potentially relevant literature; the authors checked the existence, bibliographic details, and relevance of every AI-suggested reference included in the manuscript against the original publication, and no unverified or fabricated references were retained. We used generative AI tools to prepare initial drafts of some sections from author-provided outlines and research materials. AI tools were not used to generate the reported experimental results, and the authors carefully reviewed all AI-assisted material to ensure that no fabricated references or unsupported scientific claims were included. The authors take full responsibility for the final manuscript, experimental results, and associated code, including all AI-assisted content.

ETHICS STATEMENT

This work studies prompt-free echocardiography video segmentation using publicly available benchmark datasets and pretrained segmentation models. All datasets were obtained through official distribution channels and used in strict accordance with their respective terms of use and access requirements. Our focus is methodological evaluation rather than clinical deployment. Any downstream clinical use should account for patient privacy, performance differences across patient populations and acquisition settings, and the need for independent clinical validation and expert oversight.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the FreeEcho framework, distillation objectives, and prompt-removal curriculum in Section 3. Experimental settings, datasets, and evaluation procedures are presented in Section 4, with further details on evaluation protocols, metrics, model configurations, and training durations in Appendix A.

REFERENCES

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018. URL https://www.ecva.net/papers/eccv_2018/papers_ECCV/html/Liang-Chieh_Chen_Encoder-Decoder_with_Atrous_ECCV_2018_paper.php.

Tongwaner Chen, Menghua Xia, Yi Huang, Jing Jiao, and Yuanyuan Wang. Cross-domain echocardiography segmentation with multi-space joint adaptation. *Sensors*, 23(3):1479, 2023. doi: 10.3390/s23031479.

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*, pp. 9622–9631, 2024.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2827–2836, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jiongtong Hu, Wufeng Xue, Jun Cheng, Yingying Liu, Wei Zhuo, and Dong Ni. EchoONE: Segmenting multiple echocardiography planes in one model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5207–5216, 2025.

Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, Chen Gong, and Dong Liang. Learnable prompting SAM-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(5):2295–2306, 2025.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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.

Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Løvstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, 2019.

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, 2024a.

Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. *arXiv preprint arXiv:2309.06824*, 4(11), 2023.

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

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00271. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Structured_Knowledge_Distillation_for_Semantic_Segmentation_CVPR_2019_paper.html.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Wenjing Lu, Yi Hong, and Yang Yang. Harmonizing generalization and specialization: uncertainty-informed collaborative learning for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2026.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

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, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

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. doi: 10.1016/j.echo.2023.01.015.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Springer, 2015.

Jiyang Tang, Hu Han, Shiguang Shan, and Xilin Chen. Distillation-SAM: Knowledge distillation based auto-prompt embedding learning for surgical image segmentation. *IEEE Transactions on Medical Imaging*, 2026.

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 (ICCV)*, pp. 12191–12200, 2025.

Rui Wang, Huisi Wu, and Jing Qin. OSA: Echocardiography video segmentation via orthogonalized state update and anatomical prior-aware feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1428–1438, 2026.

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.

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, pp. 2803–2811, 2023.

Qing Xu, Jiaxuan Li, Xiangjian He, Ziyu Liu, Zhen Chen, Wenting Duan, Chenxin Li, Maggie M. He, Fiseha B. Tesema, Wooi P. Cheah, Yi Wang, Rong Qu, and Jonathan M. Garibaldi. ESP-MedSAM: Efficient self-prompting SAM for universal domain-generalized medical image segmentation. *arXiv preprint arXiv:2407.14153*, 2(3), 2024.

Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, and Xiaomeng Li. GraphEcho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11844–11853. IEEE, 2023.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Haifeng Zhao, Haiyang Li, Lei-Lei Ma, and Dengdi Sun. Segment anything model meets semi-supervised medical image segmentation: A novel perspective. *Advances in Neural Information Processing Systems*, 38:170128–170148, 2025.

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, pp. 13423–13431, 2026. doi: 10.1609/aaai.v40i16.38346.

Yaqi Zhu, Changchun Xiong, Heng Zhao, and Yudong Yao. SAM-Att: a prompt-free SAM-related model with an attention module for automatic segmentation of the left ventricle in echocardiography. *IEEE Access*, 12:50335–50346, 2024.

A Experimental Details

A.1 Datasets and Evaluation Protocols

We evaluate left-ventricular (LV) segmentation using the four datasets introduced in Section 4. Table 5 summarizes the annotated test frames used in each evaluation. CAMUS includes two- and four-chamber views, while the EchoNet-Ped and CardiacUDA evaluations use four-chamber views. The CardiacUDA in-domain experiment uses a 15% site holdout, whereas zero-shot transfer uses the full four-chamber evaluation split. The two CardiacUDA protocols therefore contain different frame sets and are reported separately.

Table 5: Evaluation sets for LV segmentation. Counts denote annotated frames, not videos or independent patients. CAMUS is the training source in the zero-shot transfer setting.

Dataset	Views	Zero-shot Transfer	In-domain Training
CAMUS	Two- and four-chamber	200 (source)	200
EchoNet-Dyn	Four-chamber	2,552	2,552
EchoNet-Ped	Four-chamber	706	706
CardiacUDA	Four-chamber	566	86

Zero-shot transfer. All models use CAMUS as their task-specific training source. Their parameters remain fixed when evaluated on EchoNet-Dyn, EchoNet-Ped, and CardiacUDA, without target-data training or target-based checkpoint selection. The CAMUS column reports source performance and is not itself a zero-shot result. A public checkpoint follows the same source-based definition: public availability alone does not make its evaluation zero-shot. This protocol controls the downstream training source, rather than assuming identical foundation-model pretraining across architectures.

In-domain training. Models are trained separately on each dataset and evaluated on the corresponding held-out test set. The same annotated evaluation frames are used for the model comparisons within each dataset. Differences between zero-shot and in-domain results reflect both the training setting and, for CardiacUDA, the evaluation split; they should not be interpreted as paired improvements on identical test frames.

Prompt conditions. Prompt-based references use either a point on every frame or a point on the end-diastolic (ED) frame, depending on their prediction pipeline. Per-frame models process images independently, whereas memory-based models propagate target information from the conditioning frame. FreeEcho students receive no external localization prompts at inference. The prompt-based teacher is used only during training and is not required by the deployed student. Direct prompt removal evaluates the original model without its localization input; it does not include the additional student training performed by FreeEcho.

A.2 Evaluation Metric

For annotated frame i, let B_i denote the predicted binary foreground mask and G_i the reference LV mask. We compute the Dice similarity coefficient as

$$d_i = \frac{2|B_i \cap G_i|}{|B_i| + |G_i|}. \tag{11}$$

For N annotated frames, the reported mean and standard deviation are

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i, \qquad s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \bar{d})^2}. \tag{12}$$

All quantitative tables report $\bar{d} \pm s$ to three decimal places, following the main text. This is the population standard deviation across annotated frames, not variation across training seeds or a confidence

interval. Frames from the same video are correlated. Accordingly, the reported differences describe segmentation accuracy and do not by themselves establish statistical significance or full-cycle temporal consistency.

A.3 Model Configurations and Training

Table 6 summarizes model size, initialization, and encoder training for the evaluated implementations. Parameter counts include frozen weights. These configurations define the models used in our comparisons; they do not imply identical capacity or pretraining across methods.

Table 6: Configurations before dataset-specific supervised training. Frozen refers to the image encoder. Counts include all model parameters.

Model	Parameters	Initialization	Encoder
Prompt-free Models			
U-Net	24.5M	ImageNet ResNet-34	Frozen
DeepLabV3+	39.6M	ImageNet ResNet-50	Frozen
EchoNet CNN	39.6M	ImageNet ResNet-50	Frozen
nnU-Net	24.5M	ImageNet ResNet-34	Frozen
GDKVM	35.1M	ImageNet ResNet-50/18	Frozen
OSA	35.1M	ImageNet ResNet-50/18	Frozen
Prompt-based Models			
Medical-SAM2 (per-frame)	38.9M	MedSAM2-T	Frozen
SAMUS	133M	SAM ViT-B	Frozen
EchoONE	139M	SAM ViT-B	Frozen
MemSAM	152M	SAM ViT-B	Frozen
MedSAM2	38.9M	MedSAM2-T	Frozen
E^3SAM2	39.2M	MedSAM2-T	Frozen

Teacher and student instantiations. FreeEcho is instantiated with EchoONE and MedSAM2 as representative per-frame and memory-based architectures. Each student is initialized from its corresponding prompt-based teacher, which remains frozen during distillation. EchoONE retains its prediction pipeline and uses coarse-mask supervision for $\mathcal{L}_{\text{spatial}}$. MedSAM2 retains temporal memory and uses object-level representations for $\mathcal{L}_{\text{temporal}}$. As defined in Section 3.3, ground-truth supervision is restricted to annotated frames, while teacher-derived soft masks supervise the input sequence. These architecture-specific signals are combined with the prompt-removal curriculum in Section 3.4.

B Additional Segmentation Results

B.1 Segmentation Performance under Zero-Shot Transfer

Table 7 extends Table 1 with additional prompt-based models. SAMUS uses per-frame point prompts, while MemSAM and E^3SAM2 use an ED-frame point and propagate information through memory. The per-frame Medical-SAM2 configuration is listed separately from the memory-based MedSAM2 teacher in the main text. These models broaden the segmentation comparison; they are not additional FreeEcho teacher instantiations.

The additional references show that prompt-based accuracy varies across both architectures and data sources. SAMUS gives the highest source Dice among these references, whereas E^3SAM2 performs best on EchoNet-Ped and CardiacUDA. The prompt-free EchoONE student exceeds MemSAM and the per-frame Medical-SAM2 configuration on every dataset, and approaches the stronger prompt-based references on CAMUS and the two EchoNet datasets. A larger gap remains on CardiacUDA. These results complement the teacher–student comparisons in the main text: FreeEcho achieves strong prompt-free performance, while some prompt-based models remain more accurate.

Table 7: Additional segmentation comparisons under zero-shot transfer. All models use CAMUS as the task-specific training source. Dice (mean ± SD) is reported, with the best result within each group shown in **bold**. Prompt-based references receive per-frame or ED-frame points; FreeEcho students receive no external prompts.

Model	Source	Zero-shot Transfer		
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Additional Prompt-based Models				
Medical-SAM2 (per-frame)	0.920 ± 0.043	0.864 ± 0.073	0.838 ± 0.110	0.533 ± 0.301
SAMUS	$\mathbf{0.939 \pm 0.032}$	$\mathbf{0.900 \pm 0.053}$	0.866 ± 0.085	0.661 ± 0.262
MemSAM	0.923 ± 0.045	0.802 ± 0.130	0.733 ± 0.176	0.397 ± 0.230
E^3SAM2	0.936 ± 0.027	0.897 ± 0.045	$\mathbf{0.879 \pm 0.059}$	$\mathbf{0.734 \pm 0.171}$
FreeEcho Prompt-free Students				
EchoONE student	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.898 \pm 0.059}$	$\mathbf{0.867 \pm 0.112}$	$\mathbf{0.614 \pm 0.284}$
MedSAM2 student	0.929 ± 0.039	0.878 ± 0.073	0.863 ± 0.102	0.535 ± 0.257

B.2 Segmentation Performance under In-Domain Training

Table 8 reports the same additional references after dataset-specific training. Both FreeEcho students match or exceed MemSAM and the per-frame Medical-SAM2 configuration across the four datasets. The EchoONE student also matches SAMUS on EchoNet-Dyn and CardiacUDA at the reported precision and slightly exceeds it on EchoNet-Ped. E^3SAM2 remains a stronger prompt-based reference on the three non-CAMUS datasets. Together with Table 2, these comparisons distinguish the segmentation accuracy achievable with localization prompts from the performance obtained without them.

Table 8: Additional segmentation comparisons under in-domain training. Dice (mean ± SD) is reported, with the best result within each group shown in **bold**. CardiacUDA uses the 86-frame holdout, rather than the 566-frame zero-shot transfer set.

Model	In-domain Training			
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Additional Prompt-based Models				
Medical-SAM2 (per-frame)	0.920 ± 0.043	0.919 ± 0.042	0.896 ± 0.062	0.848 ± 0.121
SAMUS	$\mathbf{0.939 \pm 0.032}$	0.926 ± 0.040	0.906 ± 0.061	0.867 ± 0.122
MemSAM	0.923 ± 0.045	0.906 ± 0.051	0.887 ± 0.069	0.843 ± 0.109
E^3SAM2	0.936 ± 0.027	$\mathbf{0.929 \pm 0.035}$	$\mathbf{0.916 \pm 0.040}$	$\mathbf{0.881 \pm 0.080}$
FreeEcho Prompt-free Students				
EchoONE student	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.926 \pm 0.039}$	$\mathbf{0.908 \pm 0.060}$	$\mathbf{0.867 \pm 0.116}$
MedSAM2 student	0.929 ± 0.039	0.919 ± 0.028	0.903 ± 0.059	0.857 ± 0.140

C Additional Ablation Studies

C.1 Effectiveness of Distillation Supervision under In-Domain Training

Table 9 complements the zero-shot transfer results in Table 3 with dataset-specific training. It compares the full FreeEcho objective with configurations that remove ground-truth supervision, soft-mask distillation, or the architecture-specific auxiliary signal. All resulting students are evaluated without external prompts.

The full configuration achieves the highest mean Dice for both architectures across all four datasets, including a tie for EchoONE on EchoNet-Dyn. Removing ground-truth supervision produces the

Table 9: Additional supervision comparisons under in-domain training. Dice (mean ± SD) is reported, with the best result within each architecture shown in **bold**. Spatial and temporal supervision follow the architecture-specific definitions in Section 3.3.

Configuration	In-domain Training			
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Per-frame Model: EchoONE				
$\mathcal{L}_{\mathbf{KD}}$ **(default)**	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.926 \pm 0.039}$	$\mathbf{0.908 \pm 0.060}$	$\mathbf{0.867 \pm 0.116}$
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{GT}}\mathcal{L}_{\mathrm{GT}}$	0.930 ± 0.035	0.906 ± 0.050	0.898 ± 0.069	0.626 ± 0.158
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}}$	0.932 ± 0.031	$\mathbf{0.926 \pm 0.039}$	0.907 ± 0.051	0.859 ± 0.104
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{spatial}}\mathcal{L}_{\mathrm{spatial}}$	0.926 ± 0.041	0.921 ± 0.042	0.902 ± 0.065	0.855 ± 0.114
Memory-based Model: MedSAM2				
$\mathcal{L}_{\mathbf{KD}}$ **(default)**	$\mathbf{0.929 \pm 0.039}$	$\mathbf{0.919 \pm 0.028}$	$\mathbf{0.903 \pm 0.059}$	$\mathbf{0.857 \pm 0.140}$
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{GT}}\mathcal{L}_{\mathrm{GT}}$	0.920 ± 0.045	0.915 ± 0.021	0.900 ± 0.066	0.804 ± 0.200
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}}$	0.924 ± 0.043	0.918 ± 0.031	0.901 ± 0.070	0.840 ± 0.143
$\mathcal{L}_{\mathrm{KD}} - \lambda_{\mathrm{temporal}}\mathcal{L}_{\mathrm{temporal}}$	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.039 ± 0.159

largest EchoONE reduction on CardiacUDA, from 0.867 to 0.626, while removing spatial supervision reduces accuracy on every dataset. The effect of soft-mask removal is smaller under in-domain training: EchoONE retains the same rounded mean on EchoNet-Dyn, but performs worse on the other three datasets. Thus, the contribution of an individual supervision signal depends on the evaluation setting.

For MedSAM2, configurations without ground-truth or soft-mask supervision retain substantial segmentation capability but remain below the full objective. Removing temporal supervision produces zero mean Dice on CAMUS and both EchoNet datasets, and 0.039 on the CardiacUDA holdout. These results support the use of architecture-specific auxiliary supervision for memory-based students. They report the outcomes of the evaluated training configurations; they do not establish that the same loss is necessary for every possible prompt-free adaptation.

C.2 Effects of the Prompt Removal Curriculum under In-Domain Training

Table 10 extends the curriculum comparison in the main text to dataset-specific training. Linear annealing follows Equation 10. Adaptive prompt gating uses model feedback to decide whether a student receives a prompt, while training without prompts throughout omits student prompts from the start. The additional short-annealing configuration reduces the EchoONE annealing interval to four epochs and continues without prompts for sixteen epochs.

The default linear curriculum achieves the highest mean Dice for both architectures on all four datasets. Shortening the EchoONE annealing interval lowers its accuracy across the four in-domain evaluations. MedSAM2 is particularly sensitive to adaptive gating on CardiacUDA, where Dice decreases from 0.857 to 0.413. Prompt-free training from the start remains below the default, although it exceeds adaptive gating for MedSAM2 on three datasets. These comparisons support progressive prompt removal, while also showing that the alternatives have architecture- and dataset-dependent behavior.

C.3 Effects of the Annealing Interval under Zero-Shot Transfer

Table 11 evaluates whether the shorter EchoONE annealing interval preserves cross-source performance. Both configurations use twenty training epochs; the default withdraws prompts over twelve epochs, while the shorter variant does so over four. The resulting CAMUS-trained checkpoints are evaluated without target-data adaptation.

The longer annealing interval improves source Dice and all three transfer scores. The largest mean difference occurs on CardiacUDA, where Dice increases from 0.573 to 0.614. Together with the

Table 10: Prompt-removal curricula under in-domain training. Dice (mean ± SD) is reported, with the best result within each architecture shown in **bold**. All students are evaluated without external prompts.

Configuration	In-domain Training			
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Per-frame Model: EchoONE				
Linear annealing (default)	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.926 \pm 0.039}$	$\mathbf{0.908 \pm 0.060}$	$\mathbf{0.867 \pm 0.116}$
Four-epoch annealing	0.927 ± 0.040	0.920 ± 0.043	0.904 ± 0.067	0.855 ± 0.113
Adaptive prompt gating	0.928 ± 0.041	0.918 ± 0.040	0.903 ± 0.068	0.853 ± 0.121
Prompt-free throughout	0.924 ± 0.043	0.914 ± 0.045	0.904 ± 0.067	0.850 ± 0.124
Memory-based Model: MedSAM2				
Linear annealing (default)	$\mathbf{0.929 \pm 0.039}$	$\mathbf{0.919 \pm 0.028}$	$\mathbf{0.903 \pm 0.059}$	$\mathbf{0.857 \pm 0.140}$
Adaptive prompt gating	0.914 ± 0.065	0.882 ± 0.083	0.890 ± 0.103	0.413 ± 0.428
Prompt-free throughout	0.917 ± 0.039	0.902 ± 0.063	0.883 ± 0.059	0.835 ± 0.155

Table 11: Effects of the EchoONE annealing interval under zero-shot transfer. Dice (mean ± SD) is reported, with the best result shown in **bold**. Both configurations use twenty training epochs.

Configuration	Source	Zero-shot Transfer		
	CAMUS	EchoNet-Dyn	EchoNet-Ped	CardiacUDA
Twelve-epoch annealing	$\mathbf{0.937 \pm 0.036}$	$\mathbf{0.898 \pm 0.059}$	$\mathbf{0.867 \pm 0.112}$	$\mathbf{0.614 \pm 0.284}$
Four-epoch annealing	0.927 ± 0.040	0.879 ± 0.075	0.859 ± 0.111	0.573 ± 0.260

in-domain results, this comparison supports the default gradual transition to prompt-free prediction. It compares two available intervals and does not establish a universally optimal annealing duration.

C.4 Scope of the Supplementary Results

The experiments evaluate LV segmentation with two teacher architectures and four datasets. They demonstrate prompt-free conversion under these conditions, rather than a guarantee for all prompt-based models or cardiac structures. Evaluation on annotated frames also leaves full-cycle temporal consistency and downstream cardiac-function estimation unmeasured. The reported frame-level standard deviations characterize variation across evaluated images; repeated training runs would be needed to quantify optimization variability.