EchoPilot: Training-Free Ultrasound Video Segmentation via Scale-Space Semantic Prompting and Reliability-Gated Memory

Ruiqiang Xiao[1], Zhaohu Xing[1], Yijun Yang[1], Zhenyan Han[2], Weiming Wang[3], Kaishun Wu[1], and Lei Zhu[1✉]

[1] The Hong Kong University of Science and Technology (Guangzhou)
leizhu@hkust-gz.edu.cn
[2] Third Affiliated Hospital of Sun Yat-Sen University
[3] Hong Kong Metropolitan University

Abstract. Ultrasound video segmentation is clinically valuable yet difficult due to speckle noise, weak boundaries, and rapid anatomical deformation. Recent promptable foundation models enable point-guided segmentation, but their direct deployment in ultrasound remains unreliable: a single point provides insufficient spatial context to resolve scale ambiguity, and greedy memory updates amplify early errors into severe temporal drift. We present **EchoPilot**, a *training-free* framework for ultrasound video segmentation under *sparse first-frame interaction*, requiring only a single point click and an anatomical category name. EchoPilot orchestrates a frozen medical vision–language model (VLM) for semantic localization, a vision foundation model (VFM) for dense geometric feature extraction, and a promptable video segmentor for mask prediction and propagation. To resolve initialization ambiguity, we propose **Scale-Space Semantic Prompting**, which first selects an optimal contextual view via a parameter-free **S.E.E.D.** (Semantic Energy–Entropy Density) criterion, and then synthesizes geometrically precise auxiliary point prompts from dense foundation features without additional user interaction. To reduce propagation drift, a **Reliability-Gated Memory** update is further introduced to selectively freeze the segmentor's memory bank under uncertain predictions, preventing error accumulation. We also contribute the first **dynamic fetal placenta** ultrasound video segmentation dataset with 671 annotated frames. Across three ultrasound video datasets, EchoPilot achieves state-of-the-art performance under the sparse-interactive setting, consistently outperforming training-free baselines and finetuned specialists.
Project page: https://keeplearning-again.github.io/EchoPilot/.

Keywords: Video Object Segmentation · Sparse Interaction · Training-Free · Ultrasound.

1 Introduction

Video Object Segmentation (VOS) typically assumes a semi-supervised setting where a first-frame mask is propagated through the video [24,6,4,3,18,17]. Recent

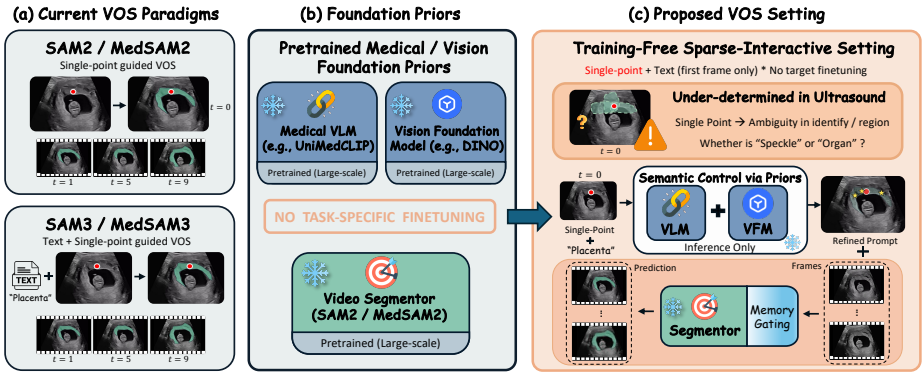

Fig. 1. Setting and concept of EchoPilot. (a) Existing point-guided and text-guided VOS paradigms suffer from initialization ambiguity or require task-specific fine-tuning. (b) EchoPilot orchestrates three frozen foundation priors—a medical VLM, a VFM, and a video segmentor—without any fine-tuning. (c) Our category-anchored sparse interaction requires only a single point click and an anatomical category name on the first frame.

foundation models for interactive segmentation—notably SAM 2 [20] and medical adaptations [26,27,15,2,28,10]—have reshaped this paradigm by enabling *point-guided* video segmentation, where sparse clicks or boxes can replace dense masks. Text-conditioned variants like SAM 3 [1] and MedSAM 3 [13] further allow users to specify semantic intent through natural language, suggesting a practical path toward clinician-friendly VOS with minimal interaction (Fig. 1(a)).

Ultrasound is uniquely suited to real-time bedside assessment due to its portability, safety, and low cost [16], yet it remains one of the most challenging modalities for automated segmentation, as images are inherently degraded by speckle noise, acoustic artifacts, weak and fluctuating boundaries [23,11]. Consequently, sparse point prompts on a single frame often *under-determine* the segmentation task (Fig. 1(c), top): a click specifies a location but may still match multiple plausible structures. Although an anatomical category provides semantic intent, it does not specify the target size, surrounding context, or boundary shape in noisy ultrasound videos. The appropriate contextual scale around the click—neither too large, which includes confusing nearby anatomy, nor too small, which removes useful cues for identification—therefore cannot be inferred from the point alone. In the video setting, such initialization ambiguity is particularly detrimental: once a segmentor commits to an incorrect interpretation, greedy memory-based propagation reinforces and compounds the error over time, resulting in persistent object drift [7,21,5].

A common solution in ultrasound segmentation research is to collect task-specific video annotations and fine-tune models for each anatomy, protocol, or device domain. However, annotation is expensive, susceptible to inter-observer variability [25], and fragile to domain shift across hardware and imaging set-

tings [14,8]. While training-free VOS has been explored under dense first-frame masks in surgical instrument segmentation [28], the combination of *training-free inference* with *sparse first-frame interaction* remains largely unaddressed for ultrasound video segmentation. We therefore study **training-free, sparse-interactive ultrasound VOS**, where all models are kept frozen at inference time and tracking is driven only by a single first-frame point click and an anatomical category name.

We propose EchoPilot, a training-free pipeline that targets the two dominant failure modes in this setting: *initialization ambiguity* and *error amplification during propagation*. The key observation is that different pretrained priors capture complementary cues for sparse-interactive VOS (Fig. 1(b)): a medical vision–language model (VLM) provides text-grounded semantic evidence about *what* to segment, a vision foundation model (VFM) supplies dense appearance features to refine *where* to prompt, and a promptable video segmentor handles temporal correspondence for mask propagation. A Reliability-Gated Memory update policy further regulates the memory bank, preventing uncertain predictions from contaminating subsequent frames and reducing drift. EchoPilot is directly compatible with existing point-guided video segmentors (*e.g.*, SAM 2 [20], MedSAM 2 [15]). We also curate the first dynamic fetal placenta ultrasound VOS dataset with 671 annotated frames. We evaluate on three ultrasound datasets spanning diverse anatomies, boundary conditions, and motion dynamics.

Our contributions are as follows: (1) we introduce **EchoPilot**, a plug-and-play framework for training-free ultrasound VOS from a single first-frame click and an anatomical category name; (2) we propose **Scale-Space Semantic Prompting** with **S.E.E.D.**, a parameter-free energy–entropy criterion for VLM-guided scale selection, together with **Reliability-Gated Memory** to reduce propagation drift; (3) we curate the first dynamic fetal placenta ultrasound VOS dataset and demonstrate consistent gains over training-free baselines and fine-tuned specialists across three ultrasound video datasets.

2 Method

2.1 Problem Setup and Overview

Given an ultrasound video $\{I_t\}_{t=0}^{T-1}$, a user provides a single positive point p_0 and an anatomical category name $\mathcal{C}$ on the first frame I_0. Our goal is to predict a segmentation mask M_t for each frame t in a **training-free** manner. The category name is a task-level descriptor (*e.g.*, "placenta") fixed per clinical protocol. After initialization, the segmentor operates solely on **point prompts**.

As illustrated in Fig. 2, EchoPilot consists of two stages. **Stage I** (t=0) resolves initialization ambiguity by determining *what* to segment and *where* to prompt: it identifies a contextual scale where the target anatomy is semantically recognizable and spatially dominant, then synthesizes geometrically precise auxiliary point prompts within the selected region. **Stage II** (t>0) stabilizes temporal propagation by controlling *when* to update the segmentor's memory bank, preventing unreliable predictions from contaminating subsequent frames.

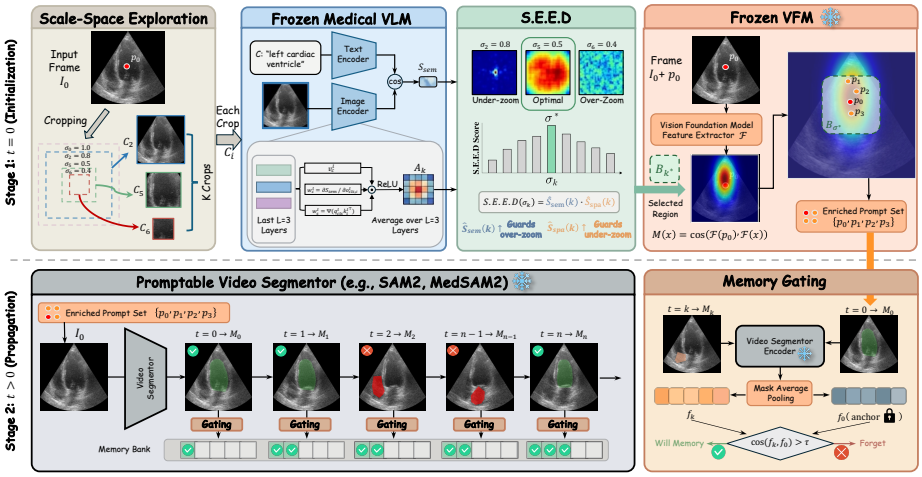

Fig. 2. Overview of EchoPilot. At t=0 (Stage I), the S.E.E.D. criterion selects an optimal contextual scale via a frozen medical VLM, and a VFM refines point prompts within the selected region. For t>0 (Stage II), a reliability-gated write policy controls which predicted frames enter the SAM 2/MedSAM 2 memory bank, suppressing error accumulation.

2.2 Stage I: Scale-Space Semantic Prompting

A single point click is ambiguous in ultrasound since the visible context depends on the observation scale: too-narrow views are dominated by speckle, while too-wide views introduce competing structures. We search for a scale where the target anatomy is both semantically identifiable and spatially prominent.

Multi-scale Crops and VLM-based Attribution. We generate K center crops $\{C_k\}_{k=1}^{K}$ around p_0 at decreasing scale factors σ_k. For each crop, a frozen medical VLM encodes the image and category name via its image encoder $\mathcal{E}_{img}$ and text encoder $\mathcal{E}_{txt}$, yielding a clamped cosine similarity:

$$\mathcal{S}_{sem}(k) = \max\big(0,\ \cos\big(\mathcal{E}_{img}(C_k),\ \mathcal{E}_{txt}(\mathcal{C})\big)\big). \tag{1}$$

This score indicates *whether* the crop matches the target but not *where*. Following [30], we extract a text-conditioned spatial attribution map $A_k \in \mathbb{R}^{h\times w}$ (where $h\times w$ is the VLM spatial token grid) by backpropagating $\mathcal{S}_{sem}(k)$ through the VLM's attention layers. For spatial token i at layer l:

$$A_k^{(l)}(i) = \mathrm{ReLU}\left(\sum_c \underbrace{\left(\frac{\partial \mathcal{S}_{sem}}{\partial o_{\mathrm{cls},c}^{(l)}}\right)}_{w_c^{(l)}\ \text{(channel)}} \cdot \underbrace{\Psi\left(q_{\mathrm{cls}}^{(l)}\, k_i^{(l)\top}\right)}_{w_s^{(l)}\ \text{(spatial)}} \cdot v_{i,c}^{(l)}\right), \tag{2}$$

where $w_c^{(l)}$ is the channel importance weight (gradient of $\mathcal{S}_{sem}$ with respect to the class-token attention output $o_{\text{cls},c}^{(l)}$), $w_s^{(l)}(i)$ is the spatial importance weight (normalized attention affinity between the class-token query $q_{\text{cls}}^{(l)}$ and the key $k_i^{(l)}$ of token i), and $v_{i,c}^{(l)}$ is the corresponding value at channel c. $\Psi(\cdot)$ denotes affinity normalization that suppresses sparse responses [30]. We aggregate over the last L layers ($A_k = \frac{1}{L}\sum_l A_k^{(l)}$) to reduce layer-specific localization noise.

S.E.E.D. Contextual Scale Selection. $\mathcal{S}_{sem}$ alone cannot distinguish under-zoomed crops where the anatomy is recognizable but occupies only a small fraction of the view. We introduce **S.E.E.D.** (Semantic Energy–Entropy Density), a parameter-free criterion that couples average attribution strength with spatial dispersion under a semantic constraint. Together, energy density suppresses diffuse low-confidence responses, while normalized entropy discourages overly collapsed attribution patterns that often arise from speckle or isolated edges. We normalize A_k into a distribution $P_k(i) = A_k(i)/(\sum_j A_k(j) + \epsilon)$ over $N_p = h \times w$ tokens, where ϵ is a small stability constant:

$$\mathcal{S}_{spa}(k) = \underbrace{\left(\frac{1}{N_p}\sum_i A_k(i)\right)}_{\text{Energy Density}} \cdot \underbrace{\left(\frac{-\sum_i P_k(i)\log P_k(i)}{\log N_p}\right)}_{\text{Normalized Entropy}}. \tag{3}$$

Over-zoomed crops tend to penalize $\mathcal{S}_{sem}$ or collapse attribution onto local artifacts, whereas under-zoomed crops reduce attribution density by spreading evidence over irrelevant anatomy. After min–max normalization ($\hat{\mathcal{S}}$), the optimal scale is

$$k^* = \arg\max_k \ \Big(\hat{\mathcal{S}}_{sem}(k) \cdot \hat{\mathcal{S}}_{spa}(k)\Big). \tag{4}$$

The bounding box B_{k^*} centered at p_0 at scale σ_{k^*} serves as the observation window for prompt refinement.

Prompt Refinement with a VFM. The VLM attribution operates on a coarse token grid, so we refine point selection using dense features from a vision foundation model. Let $\mathcal{G}(\cdot) \in \mathbb{R}^d$ denote the VFM feature at a spatial location in I_0. We compute a dense cosine similarity map anchored at the user click:

$$\mathcal{M}(x) = \cos\big(\mathcal{G}(p_0), \mathcal{G}(x)\big), \quad \forall x \in I_0. \tag{5}$$

Within B_{k^*}, we extract up to 3 local maxima via non-maximum suppression (radius r) as auxiliary positive prompts $\{p_1, p_2, p_3\}$. This cap preserves the sparse-interaction setting while still covering elongated or non-convex anatomical structures. NMS avoids redundant peaks, and the S.E.E.D.-selected window suppresses weak similarity maxima from background tissue. Restricting refinement to this window suppresses background distractors, so explicit negative prompts are unnecessary. The final prompt set $\{p_0, p_1, p_2, p_3\}$ is passed to the video segmentor.

2.3 Stage II: Reliability-Gated Memory Update

Promptable video segmentors such as SAM 2 maintain a fixed-size memory bank and, by default, update it greedily by writing each predicted frame in temporal order. Recent work has identified this strategy as a key source of error accumulation in long or ambiguous videos [5,28]. In ultrasound, severe non-rigid deformation, acoustic shadowing, and transient artifacts often yield unreliable masks; once stored, these corrupt subsequent predictions and cause compounding drift.

We therefore introduce a training-free, conservative feature-consistency write gate that decides, per frame, whether the prediction provides sufficiently stable evidence to enter the memory bank. Specifically, the gate reuses the memory-encoder feature map $\mathcal{F}_t^{\text{mem}} \in \mathbb{R}^{h\times w\times d}$ already computed by SAM 2 at every frame, introducing *zero* additional forward passes. Given the predicted mask probabilities $M_t \in [0,1]^{h\times w}$, we extract a target-specific descriptor $\mathbf{f}_t \in \mathbb{R}^d$ via masked average pooling:

$$\mathbf{f}_t = \frac{\sum_x M_t(x)\, \mathcal{F}_t^{\text{mem}}(x)}{\sum_x M_t(x) + \epsilon}. \tag{6}$$

We then measure feature consistency with a first-frame anchor $\mathbf{f}_0$, computed once from the Stage I initialization:

$$g_t = \cos(\mathbf{f}_t,\, \mathbf{f}_0), \qquad W_t = \mathbb{I}[g_t > \tau], \tag{7}$$

where $\mathbb{I}[\cdot]$ is the indicator function and τ is a reliability threshold. If W_t=1, the frame is written into the memory bank; otherwise the write is skipped to prevent unreliable evidence from contaminating the bank. Since legitimate appearance changes from deformation, probe motion, or acoustic shadowing can lower consistency, the gate is one-sided: high consistency permits an update, while low consistency triggers a skip rather than forcing a correction—favoring stable propagation over aggressive appearance adaptation. This mechanism preserves SAM 2's architecture and weights unchanged, yielding a plug-and-play drop-in replacement for the default greedy memory update.

3 Experiments and Results

Datasets and Evaluation Metrics. We evaluate on three ultrasound video segmentation datasets: *CAMUS* [11], *Breast Lesion* [12], and *Placenta*, a newly collected fetal placenta VOS dataset. We report Dice coefficient ($\mathcal{D}$), Average Surface Distance ($\mathcal{A}$), and mean F-boundary score ($\mathcal{F}$) [19] as evaluation metrics.
Implementation Details. All experiments are conducted in a training-free setting. To simulate clinical point prompts, we randomly sample a single positive point from the ground-truth mask on the first frame of each video; we fix random seeds to $\{2025, 2026, 42\}$ and report results averaged across all seeds. For VLM, we adopt BioMedCLIP (ViT-B) [29], a biomedical vision–language foundation model pretrained on large-scale diverse biomedical image–text pairs. For VFM,

Table 1. Quantitative comparison on three ultrasound VOS datasets. Methods are grouped by pretrained backbone weights (SAM2 and MedSAM2). $\mathcal{P}$ and $\mathcal{T}$ denote point and text prompts, respectively. **Bold** and underline mark the best/second-best results within each weight group. Percentages below EchoPilot entries indicate relative improvement over the corresponding base segmentor (SAM2 or MedSAM2).

Method	Venue	Prompt	CAMUS			Breast Lesion			Placenta		
			$\mathcal{D}\uparrow$ (%)	$\mathcal{A}\downarrow$	$\mathcal{F}\uparrow$ (%)	$\mathcal{D}\uparrow$ (%)	$\mathcal{A}\downarrow$	$\mathcal{F}\uparrow$ (%)	$\mathcal{D}\uparrow$ (%)	$\mathcal{A}\downarrow$	$\mathcal{F}\uparrow$ (%)
MedSAM 3	arXiv'25	$\mathcal{P}$+$\mathcal{T}$	67.15	13.98	15.24	56.93	43.31	16.58	20.69	140.16	5.39
Pretrained Weights: SAM2											
SAM2	ICLR'25	$\mathcal{P}$	28.41	56.84	0.80	<u>56.00</u>	56.54	<u>19.83</u>	16.74	<u>267.24</u>	0.89
MA-SAM2	MICCAI'25	$\mathcal{P}$	28.43	56.81	<u>0.81</u>	55.76	56.61	19.67	16.75	267.26	<u>0.90</u>
SAM2Long	ICCV'25	$\mathcal{P}$	<u>28.44</u>	<u>56.65</u>	0.80	55.41	<u>54.96</u>	19.56	<u>16.76</u>	268.99	0.89
EchoPilot	-	$\mathcal{P}$+$\mathcal{T}$	**34.09** ↑19.99%	**28.86** ↓49.23%	**3.21** ↑301.25%	**63.38** ↑13.18%	**22.98** ↓59.36%	**21.35** ↑7.67%	**39.74** ↑137.40%	**84.96** ↓68.21%	**5.92** ↑565.17%
Pretrained Weights: MedSAM2											
MedSAM2	arXiv'25	$\mathcal{P}$	90.83	2.29	77.34	61.24	86.12	23.63	33.52	129.56	3.27
MA-SAM2	MICCAI'25	$\mathcal{P}$	90.85	2.28	<u>77.37</u>	61.15	86.14	23.25	33.54	129.54	<u>3.29</u>
SAM2Long	ICCV'25	$\mathcal{P}$	<u>91.01</u>	<u>2.23</u>	**78.91**	<u>66.73</u>	<u>47.00</u>	**26.10**	<u>34.92</u>	<u>120.38</u>	**3.48**
EchoPilot	-	$\mathcal{P}$+$\mathcal{T}$	**95.31** ↑4.93%	**0.83** ↓63.76%	77.35 ↑0.01%	**68.44** ↑11.76%	**28.20** ↓67.25%	<u>23.91</u> ↑1.18%	**38.87** ↑15.96%	**63.40** ↓51.07%	3.28 ↑0.31%

we use DINOv3 (ViT-Plus) [22]. The NMS radius for auxiliary prompt selection is r=6 pixels, and the memory-gate threshold in Stage II is τ=0.5 as default.

Results and visualization. We compare EchoPilot against training-free prompting strategies (SAM2 [20], MA-SAM2 [28], SAM2Long [5]) and finetuned medical segmentors (MedSAM2 [15], MedSAM 3 [13]). The prompt column reflects each method's native interface: SAM2-family baselines are point-guided, whereas EchoPilot and MedSAM 3 can consume the anatomical category name at initialization. We therefore interpret Tab. 1 as an end-to-end comparison under the target sparse-interactive setting, and use the ablations below to separate the effects of VLM-guided initialization, VFM-based auxiliary prompts, and memory gating under matched segmentor weights.

As shown in Tab. 1, EchoPilot achieves the best Dice and ASD across all three datasets under both weight families. The gains are most pronounced on the challenging Placenta dataset, where EchoPilot improves Dice by +23.0% absolute over the SAM2 baseline (16.74→39.74) and reduces ASD by 68%. With MedSAM2 weights, EchoPilot reaches 95.31% Dice on CAMUS with an ASD of only 0.83, approaching clinical-grade accuracy. Notably, despite being fully training-free, EchoPilot consistently outperforms the finetuned MedSAM 3 by a large margin on all datasets, confirming that scale-space semantic prompting and reliability-gated memory update can effectively substitute for task-specific supervision. Under MedSAM2 weights, the boundary $\mathcal{F}$-score of EchoPilot still surpasses the MedSAM2 baseline on all datasets, while SAM2Long achieves marginally higher $\mathcal{F}$ on some splits; this is expected because the reliability gate favors conservative memory writes that stabilize region overlap (Dice/ASD) at the cost of slightly smoother boundaries.

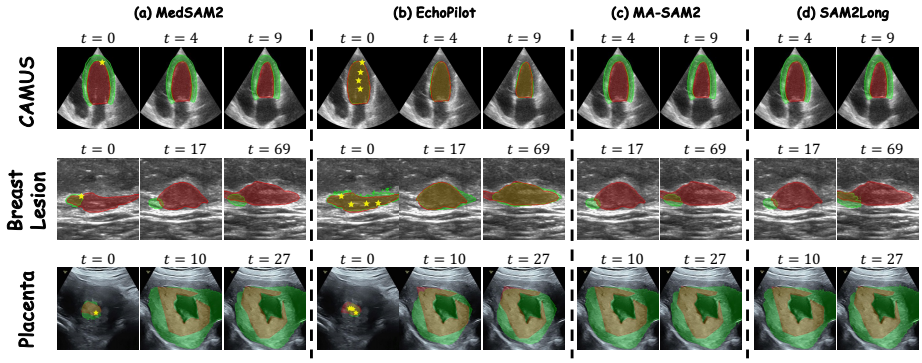

Fig. 3. Qualitative comparison across time steps on three ultrasound VOS datasets. Each row shows segmentation masks at selected frames for (a) MedSAM2, (b) EchoPilot (Ours), (c) MA-SAM2, and (d) SAM2Long. Red contours indicate ground-truth boundaries and green contours indicate predictions. Yellow stars in (b) mark the auxiliary point prompts synthesized by Stage I. Baseline methods exhibit progressive mask drift and boundary leakage over time, whereas EchoPilot maintains stable and accurate delineation throughout the sequence.

Fig. 3 reveals two consistent advantages of EchoPilot: (i) the auxiliary prompts from Stage I yield well-aligned masks from the very first frame, and (ii) on longer sequences (Breast Lesion, Placenta), the predicted contours remain anchored to the target with markedly less boundary leakage and shape drift than all baselines.

Table 2. Ablation of Stage I initialization on all three datasets. We compare the base segmentor (single random click), VFM-only prompt selection (without VLM scale guidance), and the full EchoPilot pipeline instantiated with two different medical VLM backbones. Δ denotes the change relative to the base.

Stage-I Variant	CAMUS				Breast Lesion				Fetal Placenta			
	$\mathcal{D}\uparrow$ (%)	$\mathcal{A}\downarrow$	$\Delta\mathcal{D}$ (%)	$\Delta\mathcal{A}$	$\mathcal{D}\uparrow$ (%)	$\mathcal{A}\downarrow$	$\Delta\mathcal{D}$ (%)	$\Delta\mathcal{A}$	$\mathcal{D}\uparrow$ (%)	$\mathcal{A}\downarrow$	$\Delta\mathcal{D}$ (%)	$\Delta\mathcal{A}$
Pretrained Weights: SAM2												
SAM2 (Base; 1-click random)	28.41	56.84	–	–	56.00	56.54	–	–	16.74	267.24	–	–
VFM-only	30.52	30.82	+2.11	-26.02	54.87	29.38	-1.13	-27.16	39.73	87.34	+22.99	-179.90
UniMedCLIP+VFM (EchoPilot)	29.98	31.16	+1.57	-25.68	57.78	27.61	+1.78	-28.93	39.04	88.33	+22.30	-178.91
BioMedCLIP+VFM (EchoPilot)	**34.09**	**28.86**	+5.68	-27.98	**63.38**	**22.98**	+7.38	-33.56	**39.74**	**84.96**	+23.00	-182.28
Pretrained Weights: MedSAM2												
MedSAM2 (Base; 1-click random)	90.83	2.29	–	–	61.24	86.12	–	–	33.52	129.56	–	–
VFM-only	78.32	2.65	-12.51	+0.36	67.72	24.92	+6.48	-61.20	37.81	64.50	+4.29	-65.06
UniMedCLIP+VFM (EchoPilot)	84.69	2.08	-6.14	-0.21	68.64	21.22	+7.40	-64.90	38.54	63.53	+5.02	-66.03
BioMedCLIP+VFM (EchoPilot)	**95.31**	**0.83**	+4.48	-1.46	**68.44**	28.20	+7.20	-57.92	**38.87**	**63.40**	+5.35	-66.16

Ablation Studies. Table 2 shows that VFM-only prompting is unstable (-12.5% Dice on CAMUS, MedSAM2), whereas adding a frozen VLM as a semantic anchor yields consistent gains with both UniMedCLIP [9] and BioMedCLIP [29],

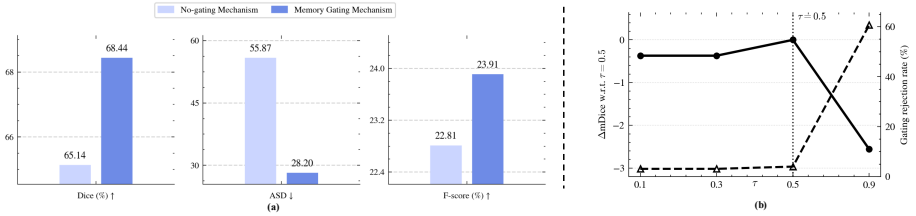

Fig. 4. Ablation of the Reliability-Gated Memory Update on the Breast Lesion dataset (MedSAM2 weights). (a) Enabling the gating mechanism substantially reduces ASD from 55.87 to 28.20 by preventing artifact-corrupted predictions from entering the memory bank. **(b)** Sensitivity analysis of the threshold τ: performance remains stable across $\tau \in [0.1, 0.5]$, while an overly aggressive threshold (τ=0.9) causes a rejection rate exceeding 60%, discarding valuable temporal context and degrading Dice.

confirming that the improvement stems from the scale-space prompting formulation rather than a specific VLM. Fig. 4 validates the reliability gate: it halves ASD on Breast Lesion (55.87→28.20) and is robust for $\tau \in [0.1, 0.5]$, while τ=0.9 over-rejects frames and degrades Dice.

4 Conclusion

We presented **EchoPilot**, a training-free framework for ultrasound video object segmentation that requires only a single point click and an anatomical category name. By coupling Scale-Space Semantic Prompting with Reliability-Gated Memory update, EchoPilot resolves initialization ambiguity and suppresses propagation drift without any task-specific fine-tuning. Experiments on CAMUS, Breast Lesion, and a newly curated fetal placenta dataset show that EchoPilot consistently outperforms both training-free baselines and finetuned specialists under two video segmentor models, while the VLM and VFM priors are queried only once on the first frame, preserving practical throughput.

Acknowledgments. This work was supported by the Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007), Guangdong Science and Technology Department (2024ZDZX2004), the National Natural Science Foundation of China (NSFC) Grant (No. 62472366), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (No. UGC/FDS16/E02/23), and AI Research and Learning Base of Urban Culture under Project 2023WZJD008.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
2. Chen, Y., Yildiz, Z., Li, Q., Chen, Y., Dong, H., Gu, H., Konz, N., Mazurowski, M.A.: Accelerating volumetric medical image annotation via short-long memory sam 2. IEEE Transactions on Medical Imaging (2025)
3. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
4. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20224–20234 (2023)
5. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13614–13624 (2025)
6. Gao, M., Zheng, F., Yu, J.J., Shan, C., Ding, G., Han, J.: Deep learning for video object segmentation: a review. Artificial Intelligence Review **56**(1), 457–531 (2023)
7. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE transactions on pattern analysis and machine intelligence **34**(7), 1409–1422 (2011)
8. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC medicine **17**(1), 195 (2019)
9. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv preprint arXiv:2412.10372 (2024)
10. Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., Li, Q.: Echofm: Foundation model for generalizable echocardiogram analysis. IEEE transactions on medical imaging (2025)
11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
12. Li, J., Zheng, Q., Li, M., Liu, P., Wang, Q., Sun, L., Zhu, L.: Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–400. Springer (2022)
13. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)
14. Liu, X., Zhou, T., Lu, M., Yang, Y., He, Q., Luo, J.: Deep learning for ultrasound localization microscopy. IEEE transactions on medical imaging **39**(10), 3064–3078 (2020)
15. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
16. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Transactions on medical imaging **25**(8), 987–1010 (2006)

17. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9226–9235 (2019)
18. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2663–2672 (2017)
19. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
20. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
21. Salti, S., Cavallaro, A., Di Stefano, L.: Adaptive appearance modeling for video tracking: Survey and evaluation. IEEE Transactions on Image Processing **21**(10), 4334–4348 (2012)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
23. Wagner, R.F., Smith, S.W., Sandrik, J.M., Lopez, H.: Statistics of speckle in ultrasound b-scans. IEEE Transactions on sonics and ultrasonics **30**(3), 156–163 (2005)
24. Wang, W., Shen, J., Porikli, F., Yang, R.: Semi-supervised video object segmentation with super-trajectories. IEEE transactions on pattern analysis and machine intelligence **41**(4), 985–998 (2018)
25. Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. Radiology **295**(1), 4–15 (2020)
26. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
27. Yan, Z., Song, S., Song, D., Li, Y., Zhou, R., Sun, W., Chen, Z., Kim, S., Ren, H., Liu, T., et al.: Samed-2: Selective memory enhanced medical segment anything model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 540–550. Springer (2025)
28. Yin, M., Wang, F., Ye, X., Meng, Y., Fu, Z.: Memory-augmented sam2 for training-free surgical video segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 328–337. Springer (2025)
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
30. Zhao, C., Wang, K., Zeng, X., Zhao, R., Chan, A.B.: Gradient-based visual explanation for transformer-based clip. In: International Conference on Machine Learning. pp. 61072–61091. PMLR (2024)