NEGATIVE MEMORY SUPPRESSION: REJECTING DISTRACTORS IN CORONARY ANGIOGRAPHY VESSEL SEGMENTATION USING AN ANCHOR FRAME

Anonymous authors
Paper under double-blind review

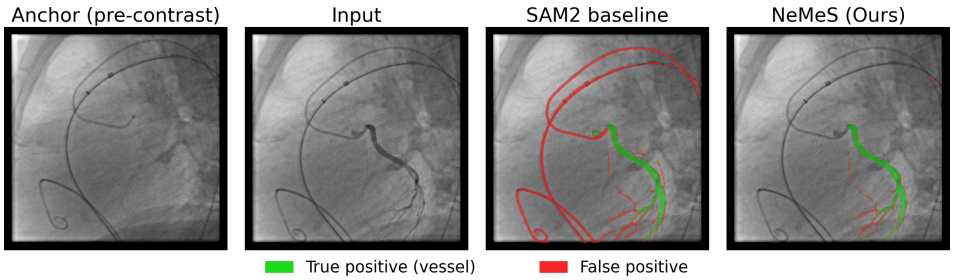

Figure 1: **Vessel-like distractors fool strong segmenters; NeMeS suppresses them with a pre-contrast anchor.** A device-cluttered coronary angiography frame (*Input*) shows a contrast-filled vessel alongside vessel-like distractors (a catheter arc and pigtail loop). A strong segmentation baseline (*SAM2*) mistakes these distractors for vessel (red), whereas NeMeS (*Ours*) keeps only the true vessel. NeMeS reuses the annotation free *pre-contrast anchor* which is already acquired at the start of every acquisition (*Anchor*) as *negative memory* to suppress against hard negatives.

ABSTRACT

Accurate vessel segmentation in X-ray coronary angiography supports quantitative analysis and procedural guidance, but is complicated by vessel-like nonvascular structures, such as catheters and ribs, that resemble coronary vessels in projection and can move differently from them. A pre-contrast frame provides a reference for identifying nonvascular structures already present before contrast injection. We introduce NeMeS, a promptless dual-branch memory model for coronary vessel segmentation in X-ray angiography that uses a pre-contrast anchor frame, typically available at the beginning of an angiographic acquisition, as annotation-free negative memory. A proposer branch segments vessel-like candidates from the video clip, while an anchor-conditioned suppressor attenuates these responses through a multiplicative gate that can suppress, but cannot increase, proposer predictions. Trained exclusively on the in-house training split, NeMeS outperforms the strongest baseline on both the full test set and a device-cluttered hard-case subset. It also transfers zero-shot to XACV and MOSXAV, achieving comparable Dice coefficient to the reported finetuned VasoMIM result on XACV and improving MOSXAV Dice coefficient by 5.0 points over TMANet. Anchor perturbation and synthetic distractor experiments support anchor-dependent rejection, while probing trained and randomly initialized, frozen networks provides evidence consistent with an architectural inductive bias for vessel-presence discrimination.

1 INTRODUCTION

Coronary angiography, which uses X-ray fluoroscopy following the injection of a contrast agent to visualize coronary anatomy and identify pathological changes, is the primary imaging modality for diagnosing coronary artery disease and guiding its treatment. Automatic segmentation of the opacified vessel tree Wu et al. (2025); Huang et al. (2026) is prerequisite for downstream tasks

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.

such as coronary roadmapping Behnoush et al. (2024); Ma et al. (2020), quantitative coronary angiography Garrone et al. (2009); de Feyter et al. (1991), and virtual fractional flow reserve Morris et al. (2013). The task is challenging. Coronary angiograms contain a complex mixture of thin, low-contrast, non-rigidly moving vessels. They also include overlapping interventional devices, such as catheters, guidewires, and pacing leads, and anatomical structures like the ribs, spine, and diaphragm. All of these elements exhibit independent or partially correlated motion driven by cardiac contraction and respiratory cycles. At a local scale, many of these distractors resemble contrast-filled vessels, as they are similarly thin, dark, and tubular. Across CNN, ViT, and memory-based segmenters, errors on heavily cluttered frames concentrate on these vessel-like devices and anatomical structures (Fig. 1; Appendix E), and the standard remedy of hard-example mining yields only marginal gains (Sec. 4.4). The difficulty is further compounded by expensive pixel-level annotations and the need to generalize across scanners, acquisition protocols, projection angles, contrast-injection conditions, and patient anatomies. Device-cluttered cases pose a particularly difficult generalization challenge, yet they remain underexplored as a dedicated evaluation setting, and device-related false positives can be masked by aggregate metrics.

In routine X-ray coronary angiography clinical acquisitions, one or more pre-contrast frames are often available before substantial coronary opacification. Di Mario & Sutaria (2005); Buytaert et al. (2020). These pre-contrast frames provide an intrinsic reference for distinguishing contrast-filled vessels from nonvascular clutter. A pre-contrast *anchor* frame captures essentially the same scene that includes catheters, guidewires, and anatomical structures but little to no contrast opacification of the coronary vessels, providing an annotation-free reference at the start of each run. Therefore, we use this frame as a *negative memory anchor*. As illustrated in Fig. 2, a *suppressor* branch conditioned on the anchor learns to veto precisely the structures present before contrast, while a *proposer* branch segments vessel-like candidates from the clip. The two are combined multiplicatively, $P = \sigma(L_A)\,(1 - \sigma(L_B))$, so a pixel is labeled vessel only if it is a tubular candidate *and* not consistent with the pre-contrast scene. The resulting dual-branch memory model is trained end-to-end and requires neither manual prompts nor a first-frame segmentation mask, enabling automatic segmentation of an angiographic sequence.

The dual-branch architecture introduces an inductive bias that mitigates shortcut learning from more available but less predictive features Hermann et al. (2024), such as dark, tubular appearance. Specifically, cross-attention between the anchor and target frames makes temporal novelty (whether a structure is present in the angiographic frame but absent from the reference frame) a more available feature for segmentation. This architectural bias is supported by the probing experiment in Sec. 4.5, in which MLP probing heads trained on frozen features from randomly initialized networks show that the proposed architecture makes temporal novelty available even before task-specific training. This inductive bias translates into improved performance both in-domain and, more prominently, under distribution shift. On the out-of-domain MOSXAV benchmark, it outperforms the CNN- and ViT-based baselines, improves vessel J&F from 0.558 to 0.664 over SAM2, and reduces the device false-positive rate from 0.216 to 0.019.

In summary, our contributions are threefold:

- We introduce NeMeS, a negative-memory architecture that uses an annotation-free pre-contrast anchor to suppress vessel-like distractors by making contrast-induced temporal novelty more available for segmentation.
- We highlight device-cluttered hard cases as an underexplored evaluation setting in coronary vessel segmentation and introduce targeted hard-case and device-specific false-positive analyses.
- We conduct extensive mechanistic and quantitative evaluations on in-domain and zero-shot external datasets, demonstrating improved robustness and generalization over CNN-, ViT-, and memory-based baselines.

2 Related Work

Video semantic segmentation. Prompt-free video segmentation predicts masks without a first-frame annotation, click, or box. Object-centric variants discover targets without a reference mask Ventura et al. (2019); Wang et al. (2019); Lee et al. (2023); category-specific methods aggre-

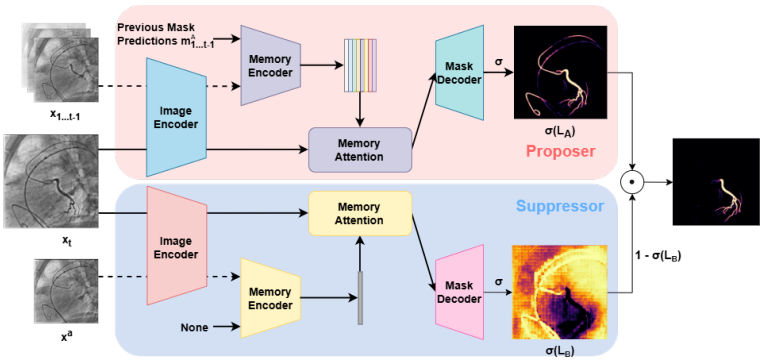

Figure 2: **NeMeS suppresses device false positives using a pre-contrast anchor.** *Top:* The proposer branch responds to both vessels and vessel-like non-vessel structures. The suppressor branch, conditioned on the pre-contrast anchor frame, identifies anchor-consistent distractors, and the gated output $P = \sigma(L_A)(1 - \sigma(L_B))$ retains the vessel response while suppressing these distractors.

gate temporal evidence through propagation, feature fusion, or memory attention Nilsson & Sminchisescu (2018); Jain et al. (2019); Hu et al. (2020); Paul et al. (2021); Wang et al. (2021). SAM2's streaming memory can also be adapted to this setting Ravi et al. (2025). NeMeS instead uses a pre-contrast frame as negative memory to identify structures excluded from vessel predictions.

Vessel segmentation in X-ray angiography. Frame-based coronary segmenters include AngioNet, DBCU-Net, and nnU-Net Iyer et al. (2021); Shen et al. (2023); Isensee et al. (2021); annotation-efficient methods use self-supervision or masked pretraining Ma et al. (2021); Huang et al. (2026). Video methods use spatiotemporal encoding, attention, or deformable representations Wang et al. (2020); Hao et al. (2020); He et al. (2025); Wu et al. (2025). These approaches strengthen vessel representation but do not explicitly suppress vessel-like devices and anatomy using a pre-contrast reference.

Shortcut learning and vessel-like distractor suppression. Our design is motivated by the shortcut-learning literature Geirhos et al. (2020); Hermann & Lampinen (2020); Hermann et al. (2024), which shows that neural networks often rely on predictive but readily available spurious cues. In X-ray angiography, such cues include the thin, dark, tubular appearance shared by contrast-filled vessels and vessel-like nonvascular structures. Using a pre-contrast reference to remove structures present before vessel opacification has a long history in angiographic imaging Meng et al. (2020). Classical digital subtraction angiography subtracts a pre-contrast mask frame, while layer-separation methods decompose angiographic sequences into respiratory, quasi-static, and vessel components using robust PCA, morphological filtering, or online low-rank and tensor models Ma et al. (2017). However, angiograms are 2D projections of superimposed anatomical layers that undergo different cardiac and respiratory motions. A single image-space deformation field therefore cannot generally align all structures simultaneously, leaving pixel-level subtraction and layer-separation methods vulnerable to residual misregistration artifacts. NeMeS instead uses the pre-contrast anchor as negative memory and learns content-based correspondence through SAM-like memory attention, enabling motion-tolerant, registration-free suppression of vessel-like devices and anatomical distractors.

3 METHOD

Local appearance alone cannot separate contrast-filled vessels from vessel-like distractors, but whether a structure was already present before contrast injection can. NeMeS makes this cue explicit by conditioning segmentation on a pre-contrast anchor frame used as negative memory (Fig. 2). We first formalize promptless causal segmentation with an anchor (Sec. 3.1) and then describe the dual-branch model (Sec. 3.2): a proposer that segments vessel-like candidates using memory of preceding predictions, and a suppressor that attends to the fixed anchor memory. The two branches are combined by a multiplicative gate and trained with a single segmentation loss.

3.1 Problem Formulation

Given a grayscale fluoroscopy sequence $X = (x_1, \ldots, x_T)$, with $x_t \in \mathbb{R}^{H\times W}$, the model segments the frames causally. At step t, it predicts a vessel-probability map $P_t \in [0,1]^{H\times W}$ using the current frame, the memory accumulated from frames $x_1, \ldots, x_{t-1}$, and an annotation-free pre-contrast anchor $x^{\mathrm{a}} \in \mathbb{R}^{H\times W}$. The prediction for frame t is then encoded into memory before processing frame $t+1$, enabling the entire sequence to be segmented through sequential rollout. The task is binary vessel segmentation, with all non-vessel structures treated as background. The setting is promptless: no mask, point, or box annotation is provided at inference. The anchor is acquired before coronary opacification and depicts the anatomy and radiopaque devices without contrast-filled vessels (Fig. 5 in the Appendix). Thus, the model implements the causal mapping

$$P_t = f_\theta(x_t, \mathcal{M}_{t-1}, x^{\mathrm{a}}),$$

where $\mathcal{M}_{t-1}$ denotes the memory accumulated from preceding frames.

3.2 Dual-branch Memory-based Model

Both branches use the SAM2 Hiera-T architecture Ravi et al. (2025), excluding the prompt encoder because no prompts are provided. The image encoder, memory-attention module, memory encoder, and mask decoder are initialized from the public checkpoint; the branches have independent weights, differ only in memory construction, and each produces a per-frame logit map at the input resolution.

Proposer branch (L_A). The proposer performs causal segmentation using a FIFO memory bank of capacity N_{mem}=7. For each frame x_t, the Hiera encoder extracts multi-resolution features

$$(f^A_{ta}, f^A_{tb}, f^A_{tc}) = \mathrm{Hiera}_A(x_t), \tag{1}$$

where a, b, and c denote high-, intermediate-, and low-resolution features, respectively. For the first frame, no preceding memory is available, so the deepest feature is conditioned by a learned no-memory token f_0. Subsequent frames attend to the memories accumulated from earlier predictions:

$$\tilde{f}^A_t = \begin{cases} f^A_{1c} + f_0, & t = 1, \\ \mathrm{MemAttn}_A(f^A_{tc}, \mathcal{B}_{t-1}), & t > 1, \end{cases} \tag{2}$$

where $\mathcal{B}_{t-1}$ is the current FIFO memory bank. The memory-attention module follows the standard SAM2 design, comprising self-attention, cross-attention to the stored memories, and an MLP.

The conditioned feature is decoded together with the higher-resolution features:

$$L_{tA} = \mathrm{MaskDec}_A(f^A_{ta}, f^A_{tb}, \tilde{f}^A_t), \qquad m^A_t = \sigma(L_{tA}). \tag{3}$$

The resulting mask is encoded with the deep image feature and appended to the memory bank:

$$\begin{aligned} \mathcal{M}^A_t &= \mathrm{MemEnc}_A(f^A_{tc}, m^A_t), \\ \mathcal{B}_t &= \mathrm{FIFO}_{N_{\mathrm{mem}}}(\mathcal{B}_{t-1}, \mathcal{M}^A_t). \end{aligned} \tag{4}$$

The proposer thus propagates positive temporal evidence and identifies vessel-like candidates throughout the clip.

Suppressor branch (L_B). The suppressor uses the same modules, with independent parameters, but replaces the evolving memory bank with a single fixed negative memory. The pre-contrast anchor x^{a} is encoded once:

$$\begin{aligned} (f^{\mathrm{a}}_a, f^{\mathrm{a}}_b, f^{\mathrm{a}}_c) &= \mathrm{Hiera}_B(x^{\mathrm{a}}), \\ \mathcal{M}^{\mathrm{a}} &= \mathrm{MemEnc}_B(f^{\mathrm{a}}_c, \varnothing). \end{aligned} \tag{5}$$

where $\varnothing$ denotes the absence of a mask input. This anchor memory remains fixed throughout the clip and is never overwritten.

For each frame x_t, the suppressor attends only to the fixed anchor memory:

$$(f^B_{ta}, f^B_{tb}, f^B_{tc}) = \mathrm{Hiera}_B(x_t), \tag{6}$$

$$\tilde{f}^B_t = \mathrm{MemAttn}_B(f^B_{tc}, \{\mathcal{M}^{\mathrm{a}}\}), \tag{7}$$

$$L_{tB} = \mathrm{MaskDec}_B(f^B_{ta}, f^B_{tb}, \tilde{f}^B_t). \tag{8}$$

No per-frame memory is written back in this branch. Under the gated segmentation objective, the suppressor learns to respond to structures consistent with the pre-contrast scene, including devices and anatomical background. Because matching is performed through feature-level attention rather than pixelwise subtraction, it remains tolerant to cardiac, respiratory, table, and device motion without explicit registration.

Anchor-gated output and loss. The per-pixel vessel probability is $P = \sigma(L_A)\big(1 - \sigma(L_B)\big)$. This multiplicative gate allows the suppressor to attenuate, but not introduce, proposer responses. We convert P to logits and optimize binary cross-entropy against the expert vessel mask on the query frame x_T. No auxiliary, Dice, boundary, or deep-supervision losses are used; the proposer–suppressor decomposition is learned solely through the gated segmentation objective.

4 Experiments

4.1 Datasets

Training and in-domain test data. We use an in-house coronary X-ray angiography (XCA) corpus comprising 5,621 sequences and 30,535 expert-annotated binary vessel frames. Using a fixed split, 5,058 sequences are assigned to training, 263 to validation, and 300 to testing; the validation set is used exclusively for checkpoint selection. All frames are converted to grayscale and resized to 512×512. Within the dataset, we further identified a *hard-case subset* of 190 sequences in which the coronary vessels substantially overlap with vessel-like distractors, including the spine, ribs, catheters, and other tubular or dark devices (Fig. 3). All models are trained and evaluated using the same data splits.

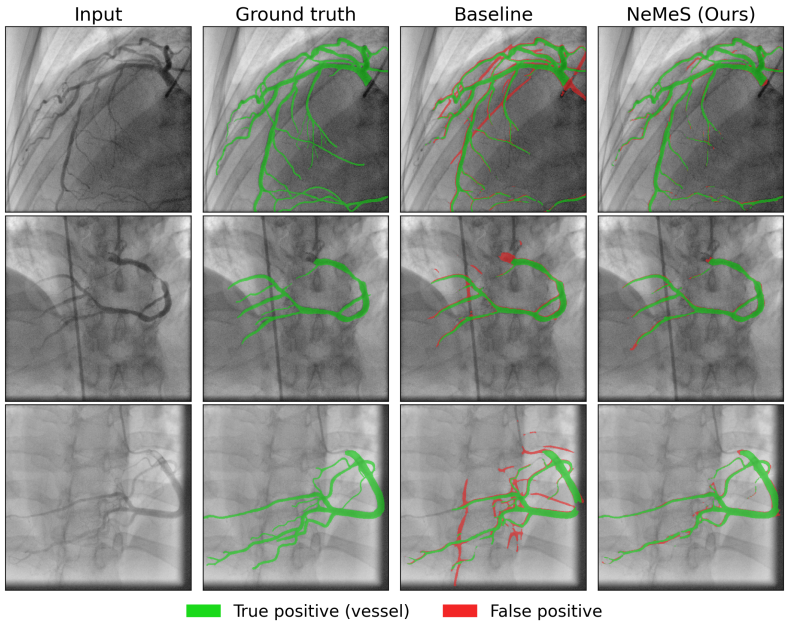

Figure 3: **Qualitative comparison on in-house hard cases.** Each row shows a test frame in which the coronary tree overlaps ribs, spine, diaphragm, or radiopaque devices. Columns: input, expert ground truth (green), SAM2, and NeMeS; green marks true positives and red false positives. SAM2 produces false positives along vertebral and diaphragm edges and on devices, whereas NeMeS suppresses them and preserves the vessel tree.

Zero-shot external test data We evaluate cross-dataset generalization on two public XCA datasets by applying the in-house-trained models without target-domain training, fine-tuning, or prompts.

MOSXAV. MOSXAV Xi et al. (2026) contains 12 densely annotated videos and 488 frames with pixel-level labels for vessels and multiple device classes. Because three ventriculography videos

Method	Cardiac (full test, n=300)			Cardiac (hard test, n=190)			XACV (zero-shot)			MOSXAV (zero-shot)		
	Dice	clDice	J&F	Dice	clDice	J&F	Dice	clDice	J&F	Dice	clDice	J&F
ViT-CoMer	67.5	67.3	64.8	55.8	55.2	54.3	25.4	22.8	24.3	57.6	55.0	53.9
TVS-Net	70.6	70.8	68.0	59.2	59.0	57.3	57.8	51.2	52.4	40.3	35.0	33.7
nnU-Net	72.7	73.7	71.7	62.4	62.4	61.9	72.7	69.4	72.2	43.9	39.1	36.8
TMANet	74.3	75.2	75.2	65.1	64.5	66.5	74.6	71.9	74.7	64.3	64.0	60.9
SAM2	78.5	80.3	77.4	69.5	69.9	68.2	76.6	75.6	76.1	60.0	60.6	55.8
VasoMIM	–	–	–	–	–	–	**77.5**	74.2	–	–	–	–
NeMeS (Ours)	**80.5**	**82.1**	**79.2**	**72.0**	**72.0**	**70.2**	**77.5**	**77.1**	**77.1**	**69.3**	**71.4**	**66.4**

Table 1: **Main results across datasets** (binary vessel segmentation, promptless; Dice / clDice / J&F in %, threshold 0.5, common 512 grid; higher is better). *Cardiac* is our in-house test set, reported on the full test set (n=300) and the hard-case test set (n=190). *XACV* follows the VasoMIM patient-level protocol. *MOSXAV* is evaluated *zero-shot* on the cleaned coronary subset (n=317; §4.1). Values are means; XACV values are averaged over three patient-level test splits. VasoMIM is a reported external baseline; "–" = not evaluated. NeMeS outperforms every directly evaluated baseline on all datasets (two-sided paired t-test, $P < 0.01$) and ties the reported VasoMIM Dice on XACV.

label opacified cardiac chambers as vessels, we report primary results on a coronary subset of 9 videos and 372 frames, including 317 vessel-present frames. Device annotations are used for device-specific false-positive analysis. Results on the full benchmark are reported in Appendix B.

XACV. XACV Wu et al. (2025) contains 59 patients, 111 sequences, and 221 sparsely annotated vessel frames. To enable comparison with VasoMIM Huang et al. (2026), we follow its patient-level evaluation protocol using three random splits of 12 test patients and report the mean across splits. Further details are provided in Appendix B.

4.2 Evaluation Protocol

Benchmarked models. We compare NeMeS with representative per-frame, temporal, and memory-based segmentation models. The per-frame baselines include nnU-Net Isensee et al. (2021) and ViT-CoMer Xia et al. (2024). Temporal XCA baselines include the model of He et al. He et al. (2025), while the memory-based baselines include TMANet Wang et al. (2021) and a prompt-free adaptation of SAM2 Ravi et al. (2025). All trainable baselines use the same in-house training and validation splits and are evaluated without prompts. External datasets are evaluated zero-shot. On XACV, we additionally include the published VasoMIM result Huang et al. (2026) as an external reference.

Benchmark metrics. We report Dice similarity coefficient (Dice) for region overlap, centerline Dice (clDice) Shit et al. (2021) for vascular topology, and DAVIS-style J&F Perazzi et al. (2016) for combined region and boundary accuracy. For the sparsely annotated in-house and XACV datasets, J&F is computed independently on each annotated frame; for densely annotated MOSXAV, it is averaged over the complete annotated sequences. Predictions are thresholded at 0.5 on a common 512×512 grid. All results are reported as means.

Ablation studies and additional analyses. We isolate the contributions of the suppressor and pre-contrast anchor through suppressor removal, a parameter-matched random-anchor control, and inference-time anchor perturbations. We further examine the mechanism using a synthetic pasted-distractor experiment and MLP probing of frozen trained and randomly initialized features. Device-specific false positives, temporal stability, and computational cost are analyzed separately. Full training details, metric, and ablation protocols are provided in Appendix C.

4.3 Main Results

We evaluate all baselines and NeMeS on the in-house cardiac test set, XACV, and MOSXAV in the promptless regime (Table 1). We report Dice for region overlap, centerline-Dice (clDice) Shit et al. (2021) for topological fidelity, and the DAVIS $J\&F$ score Perazzi et al. (2016). For the in-house data, we report results on the full test set (n=300) and a hard-case subset (n=190) characterized by substantial overlap between coronary vessels and vessel-like distractors.

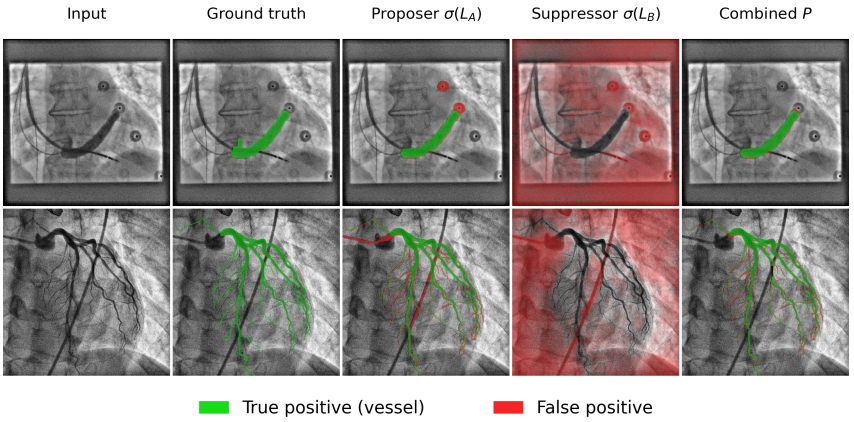

Figure 4: **Per-branch outputs of NeMeS (row 1: MOSXAV; row 2: XACV; both zero-shot).** Left to right: input frame, expert ground truth (green), proposer response $\sigma(L_A)$, suppressor response $\sigma(L_B)$, and combined prediction $P = \sigma(L_A)\,(1-\sigma(L_B))$; green marks true positives and red false positives. The proposer responds to the vessel and to distractors (a catheter arc, pigtail loop, and ring markers in row 1; a catheter in row 2). The suppressor activates on the anchor-consistent scene but not on the contrast-filled vessel, so the product retains the vessel and removes the distractors.

NeMeS achieves the best performance on both in-house test sets, outperforming SAM2, the strongest baseline, by 2.0 Dice points on the full test set and 2.5 points on the hard-case test set. The improvements are larger on the hard-case test set, consistent with the suppressor removing device-related distractors.

Our method also generalizes strongly across domains. On XACV and MOSXAV, we apply the in-house-trained model directly, without target-domain training or prompts. On XACV, NeMeS achieves 77.5 Dice, matching the finetuned VasoMIM baseline Huang et al. (2026), while attaining higher clDice (77.1 vs. 74.2). Among the directly evaluated zero-shot models, NeMeS performs best, improving over the same-backbone SAM2 baseline by 0.9 Dice and 1.5 clDice. The advantage is substantially larger on MOSXAV, where NeMeS achieves 69.3 Dice, compared with 64.3 for TMANet and 60.0 for SAM2. Relative to SAM2, this corresponds to gains of 9.3 Dice and 10.8 clDice while reducing the device false-positive rate from 0.216 to 0.019, an $11\times$ reduction (see Appendix E). Although this transfer may partly benefit from the scale of the in-house training set, the comparison with SAM2 isolates the contribution of the anchor-conditioned suppressor under an identical backbone. As illustrated in Fig. 4, the proposer responds to both vessels and distractors, whereas the suppressor removes anchor-consistent background structures and devices from the final prediction. Together, these results demonstrate that the suppressor is particularly beneficial under domain shift involving abundant non-vessel distractors.

4.4 Ablations

Suppressor and capacity control. Removing the anchor-conditioned suppressor recovers the prompt-free SAM2 baseline (Table 1). To separate the effect of anchor information from that of additional capacity, we train a parameter-matched random-anchor variant with the same two-branch architecture, multiplicative gate, and 77.9M parameters. On the hard-case test set, it achieves 0.709 Dice, between SAM2 (0.695) and the pre-contrast-anchor model (0.720); the same ordering holds on the full test set ($0.785 < 0.797 < 0.805$). Thus, the additional branch provides a modest gain, while the informative pre-contrast anchor accounts for the remaining improvement.

Anchor-frame choice. We replace the pre-contrast anchor at inference with frames of progressively greater vessel opacification (Tab. 2). Performance generally decreases as anchor contrast increases, with the pre-contrast anchor performing best or tying for best on all three datasets. The decline is negligible on XACV ($\Delta = -0.002$, due to the simplicity of the data), larger on MOSXAV ($\Delta = -0.036$), and greatest on the in-house hard-case test set ($\Delta = -0.056$), indicating that the

anchor is most beneficial when vessel-like clutter is complex. These results support the pre-contrast frame as the appropriate anchor while showing tolerance to mild opacification.

Hard-example mining. A natural alternative remedy for hard-case failures is to oversample difficult training data Shrivastava et al. (2016); Lin et al. (2017). We continue training the SAM2 baseline for 30 epochs while oversampling low-Dice training videos, and compare it with a control continued for the same number of epochs under uniform sampling. Relative to this control, resampling improves the hard-case test set by only 0.9 Dice and the full test set by 0.2 (Appendix **??**). Neither continued variant reaches NeMeS, which is trained for 100 epochs without additional steps: additional exposure to difficult cases yields modest gains, whereas the pre-contrast anchor provides a larger improvement on both test sets.

	Anchor contrast ratio					
Dataset	0	0.25	0.5	0.75	1.0	Δ
In-house hard	**72.0**	72.0	71.8	71.3	66.5	-5.6
MOSXAV zero-shot	**67.4**	65.3	63.8	61.5	63.8	-3.6
XACV zero-shot	**77.9**	77.9	77.9	77.7	77.7	-0.2

Table 2: **Anchor-frame ablation.** Mean Dice (%) of the trained NeMeS with its anchor swapped *at inference* from a (near-)pre-contrast frame (ratio 0) to a full contrast frame (ratio 1); $\Delta = \mathrm{Dice}(1) - \mathrm{Dice}(0)$. A pre-contrast anchor is best everywhere ($\Delta \leq 0$), and the drop grows with the amount of static device clutter—largest on the in-house hard-case set (n=190), moderate on device-heavy MOSXAV (n=317), negligible on clean XACV. Frames are binned by contrast using a vessel-detector proxy (in-house and XACV; XACV's Δ is robust to the detector used) or GT vessel area (MOSXAV, densely annotated).

4.5 Shortcut-learning analysis

Shortcut learning occurs when a model favors features that are easier to extract over those that are more predictive Hermann & Lampinen (2020); Hermann et al. (2024). In coronary angiography, conventional architectures are readily attracted to local appearance cues, such as dark, thin, and tubular structures, causing vessel-like devices and anatomy to be mistaken for vessels. We hypothesize that the pre-contrast anchor provides a direct architectural shortcut to the more discriminative cue of contrast-induced temporal novelty. We examine this hypothesis through synthetic distractor insertion and probing of the frozen dual-branch representation.

Distractor robustness: paste-in-a-rib.

To test whether the models rely on the appearance shortcut of dark, tubular structure, we paste a real radiopaque distractor into a vessel-free, no-contrast clip and measure *coverage*, the fraction incorrectly labeled as vessel. We vary the distractor thickness and compare two NeMeS settings: the distractor appears in both the anchor and clip (**anchor+clip**) or in the clip only.

All models become more susceptible as the distractor approaches vessel thickness, but NeMeS with the anchor consistently yields the lowest coverage among the SAM2-family models (Table 3). The reduction from clip-only to anchor+clip, from 0.18 to 0.02 at baseline thickness and from 0.79 to 0.57 at the thickest setting, shows that the suppressor rejects the same vessel-like structure when it is identified in the anchor. This supports anchor-relative temporal novelty, rather than tubular appearance alone, as the operative cue.

Inductive bias of the dual-branch representation

To test whether NeMeS makes temporal novelty more readily decodable, we freeze the image encoders and memory modules of NeMeS and the prompt-free SAM2 baseline, and train only an MLP probing head to predict whether a frame contains a newly opacified vessel absent from the pre-contrast scene. We probe both the trained backbones and randomly initialized backbones; the latter separates the contribution of the architecture from that of the learned weights. With trained backbones on MOSXAV, the proposer$-$suppressor features achieve 0.949 AUC, compared with 0.885 using the proposer alone and 0.872 using SAM2 (Table 4). The similar performance of the proposer

			NeMeS	
Rib thickness	TMANet	SAM2	clip-only	anchor+clip
thinner (-1)	0.14	0.06	0.03	**0.02**
baseline (0)	0.52	0.06	0.18	**0.02**
$+1$	0.67	0.32	0.52	**0.13**
$+2$	0.77	0.64	0.75	**0.44**
thicker ($+3$)	0.81	0.76	0.79	**0.57**

Table 3: **Distractor robustness (paste-in-a-rib).** Coverage—the fraction of a real tubular distractor pasted into a vessel-free, no-contrast clip that the model wrongly labels vessel (lower is better; ideal 0)—as the distractor is dilated from thinner-than-vessel (-1) to thicker ($+3$) at fixed intensity. NeMeS with the anchor is the most false-positive-resistant SAM2-family model at every thickness; the clip-only vs. anchor+clip gap is the suppressor's causal false-positive rejection.

and SAM2 indicates that the gain arises from the complementary anchor-conditioned suppressor features. With randomly initialized backbones, averaged over five initialization seeds, the dual-branch features reach 0.621 AUC compared with 0.378 for SAM2, indicating that temporal novelty is already more decodable from the dual-branch architecture before any training and supporting it as an architectural inductive bias.

Probe features	MOSXAV AUC	XACV AUC
Trained backbone		
SAM2 (single branch)	0.872	1.000
NeMeS proposer only	0.885	1.000
NeMeS proposer$-$suppressor	**0.949**	1.000
Randomly initialized backbone (untrained)		
SAM2 (single branch)	0.378	0.938
NeMeS (dual-branch)	**0.621**	0.958

Table 4: **Inductive bias of the dual-branch representation.** Frame-level "new vessel present" AUC of an MLP probing head trained on *frozen* features (higher is better). *Top (trained backbone):* on MOSXAV the proposer$-$suppressor features are the most novelty-decodable, while the proposer alone only matches SAM2, so the gain comes from the anchor-conditioned suppressor rather than added capacity. *Bottom (randomly initialized backbone; no pretraining, both branches tied to identical random weights):* the dual-branch features already decode novelty far better than the single-branch baseline on MOSXAV ($+0.243$; SAM2 is below chance), indicating the bias is architectural. **XACV saturates for all settings** (trained 1.000; random-init $\approx$0.94–0.96) because its negatives are pure pre-contrast frames, trivially separable; MOSXAV is the discriminative benchmark, as its negatives contain devices (vessel-novelty vs. device-presence). Randomly initialized rows are averaged over five backbone initialization seeds.

5 Conclusion

We presented NeMeS, a promptless dual-branch model that uses a pre-contrast frame as negative memory for coronary vessel segmentation. The proposer identifies vessel-like candidates, while the anchor-conditioned suppressor removes structures consistent with the pre-contrast scene through the gated output $P = \sigma(L_A)(1 - \sigma(L_B))$. NeMeS outperforms strong CNN, ViT, and memory-based baselines in-domain and transfers zero-shot to external datasets. Anchor perturbation, synthetic distractor experiments, and representation probing further show that the improvement arises from anchor-dependent distractor rejection and a more accessible representation of contrast-induced temporal novelty thanks to the architecture design. These results suggest that pre-contrast negative memory is a simple and effective mechanism for robust angiographic vessel segmentation.

AI use statement

The authors used a large language model (Claude, Anthropic) as a writing and research assistant: to edit and restructure prose, check consistency of numbers between text, tables, and figures, and assist with evaluation scripts and code for the experiments. All experimental design, results, figures, and scientific claims were produced, verified, and are the responsibility of the authors.

Reproducibility statement

The architecture, training setup, datasets, evaluation protocol, and ablation procedures are described in Sections 4 and 4.4 and in the appendix. The appendix also provides metric definitions and computational cost details.

References

Amir Hossein Behnoush, Alireza Ramandi, Sugandhi Mahajan, Ahmed Altibi, Amirsaeed Samavarchitehrani, and Rahul Gupta. Dynamic coronary roadmap in percutaneous coronary intervention: a systematic review and meta-analysis. *BMC Cardiovascular Disorders*, 24(1):681, 2024.

Dimitri Buytaert, Benny Drieghe, Frédéric Van Heuverswyn, Jan De Pooter, Peter Gheeraert, Daniël De Wolf, Yves Taeymans, and Klaus Bacher. Combining optimized image processing with dual axis rotational angiography: toward low-dose invasive coronary angiography. *Journal of the American Heart Association*, 9(13):e014683, 2020.

Pim J de Feyter, Patrick W Serruys, Michael J Davies, Peter Richardson, Jacobus Lubsen, and Michael F Oliver. Quantitative coronary angiography to measure progression and regression of coronary atherosclerosis. value, limitations, and implications for clinical trials. *Circulation*, 84 (1):412–423, 1991.

Carlo Di Mario and Nilesh Sutaria. Coronary angiography in the angioplasty era: projections with a meaning. *Heart*, 91(7):968–976, 2005.

Paolo Garrone, GIUSEPPE BIONDI-ZOCCAI, Ilaria Salvetti, Noemi Sina, Imad Sheiban, Peter R Stella, and Pierfrancesco Agostoni. Quantitative coronary angiography in the current era: principles and applications. *Journal of interventional cardiology*, 22(6):527–536, 2009.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Dongdong Hao, Song Ding, Linwei Qiu, Yisong Lv, Baowei Fei, Yueqi Zhu, and Binjie Qin. Sequential vessel segmentation via deep channel attention network. *Neural Networks*, 128:172–187, 2020. doi: 10.1016/j.neunet.2020.05.005.

Haorui He, Abhirup Banerjee, Robin P. Choudhury, and Vicente Grau. Deep learning based coronary vessels segmentation in x-ray angiography using temporal information. *Medical Image Analysis*, 102:103496, 2025. doi: 10.1016/j.media.2025.103496.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33: 9995–10006, 2020.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, volume 2024, pp. 43832–43868, 2024.

Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8818–8827, 2020.

De-Xing Huang, Xiao-Hu Zhou, Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu, Shuang-Yi Wang, Tian-Yu Xiang, Rui-Ze Ma, Nu-Fang Xiao, and Zeng-Guang Hou. VasoMIM: Vascular anatomy-aware masked image modeling for vessel segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4985–4993, 2026. doi: 10.1609/aaai.v40i6.42503.

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(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Kritika Iyer, Cyrus P. Najarian, Aya A. Fattah, Christopher J. Arthurs, S. M. Reza Soroushmehr, Vijayakumar Subban, Mullasari A. Sankardas, Raj R. Nadakuditi, Brahmajee K. Nallamothu, and C. Alberto Figueroa. AngioNet: A convolutional neural network for vessel segmentation in x-ray angiography. *Scientific Reports*, 11:18066, 2021. doi: 10.1038/s41598-021-97355-8.

Samvit Jain, Xin Wang, and Joseph E. Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8866–8875, 2019.

Minhyeok Lee, Suhwan Cho, Seunghoon Lee, Chaewon Park, and Sangyoun Lee. Unsupervised video object segmentation via prototype memory network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5924–5934, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Hua Ma, Ayla Hoogendoorn, Evelyn Regar, Wiro J. Niessen, and Theo van Walsum. Automatic online layer separation for vessel enhancement in x-ray angiograms for percutaneous coronary interventions. *Medical Image Analysis*, 39:145–161, 2017.

Hua Ma, Ihor Smal, Joost Daemen, and Theo van Walsum. Dynamic coronary roadmapping via catheter tip tracking in x-ray fluoroscopy with deep learning based bayesian filtering. *Medical image analysis*, 61:101634, 2020.

Yuxin Ma, Yang Hua, Hanming Deng, Tao Song, Hao Wang, Zhengui Xue, Heng Cao, Ruhui Ma, and Haibing Guan. Self-supervised vessel segmentation via adversarial learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7536–7545, 2021.

Cai Meng, Kai Sun, Shaoya Guan, Qi Wang, Rui Zong, and Lei Liu. Multiscale dense convolutional neural network for dsa cerebrovascular segmentation. *Neurocomputing*, 373:123–134, 2020.

Paul D Morris, Desmond Ryan, Allison C Morton, Richard Lycett, Patricia V Lawford, D Rodney Hose, and Julian P Gunn. Virtual fractional flow reserve from coronary angiography: modeling the significance of coronary lesions: results from the virtu-1 (virtual fractional flow reserve from coronary angiography) study. *JACC: Cardiovascular Interventions*, 6(2):149–157, 2013.

David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6819–6828, 2018.

Matthieu Paul, Martin Danelljan, Luc Van Gool, and Radu Timofte. Local memory attention for fast video semantic segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1102–1109, 2021.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 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.

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.

Yuqiang Shen, Zhe Chen, Jijun Tong, Nan Jiang, and Yun Ning. DBCU-Net: Deep learning approach for segmentation of coronary angiography images. *The International Journal of Cardiovascular Imaging*, 39(8):1571–1579, 2023. doi: 10.1007/s10554-023-02849-3.

Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16560–16569, 2021.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marqués, and Xavier Giró-i Nieto. RVOS: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5277–5286, 2019.

Hao Wang, Weining Wang, and Jing Liu. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing*, pp. 2254–2258, 2021.

Lu Wang, Dongxue Liang, Xiaolei Yin, Jing Qiu, Zhiyun Yang, Junhui Xing, Jianzeng Dong, and Zhaoyuan Ma. Coronary artery segmentation in angiographic videos utilizing spatial-temporal information. *BMC Medical Imaging*, 20:110, 2020. doi: 10.1186/s12880-020-00509-9.

Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9235–9244, 2019.

Chun-Hung Wu, Shih-Hong Chen, Chih-Yao Hu, Hsin-Yu Wu, Kai-Hsin Chen, Yu-You Chen, Chih-Hai Su, Chih-Kuo Lee, and Yu-Lun Liu. DeNVeR: Deformable neural vessel representations for unsupervised video vessel segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15682–15692, 2025.

Lin Xi, Yingliang Ma, and Xiahai Zhuang. Few-shot video object segmentation in x-ray angiography using local matching and spatio-temporal consistency loss. *arXiv preprint arXiv:2601.00988*, 2026.

Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, and Yifeng Shi. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5493–5502, 2024.

A Anchor Frame

The pre-contrast anchor provides an annotation-free reference that captures the anatomical scene and interventional devices before coronary opacification. As illustrated in Fig. 5, comparing the anchor with subsequent contrast-enhanced frames helps distinguish newly opacified vessels from pre-existing vessel-like distractors.

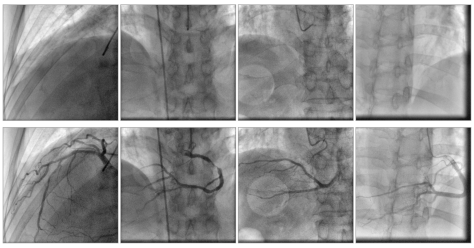

Figure 5: **Pre-contrast frames vs. contrast injected frames** *Top row:* pre-contrast frames from different videos showing the background of the videos. In an angiography video, the background typically contains catheter device, spine, ribs, etc. *Bottom row:* contrast injected frames from different videos corresponding to the top row background frames. The vessel tree can be seen clearly with the contrast injected but looking at the contrast injected frames alone can be confusing due to the vessel-like devices and background structures. The background structures can be either static, such as the spine or ribs, or dynamic, such as the diaphragm and catheter tips.

B External Datasets

All external datasets are evaluated zero-shot: models are trained exclusively on the in-house corpus and applied without target-domain training, fine-tuning, checkpoint selection, or visual prompts.

B.1 MOSXAV

MOSXAV Xi et al. (2026) is a densely annotated XCA video-segmentation benchmark containing 12 test videos and 488 annotated frames at a resolution of 512×512. Each frame includes pixel-level labels for six classes: background, vessel, contrast-filled catheter, catheter/guidewire, balloon, and other devices. Because our task is binary coronary-vessel segmentation, all non-vessel classes are treated as background.

Coronary evaluation subset. Three videos were acquired during ventriculography and label an opacified cardiac chamber as "vessel." Because these labels do not represent the coronary vessel tree targeted by our models, they constitute a task-definition confound. We therefore use the remaining 9 videos, comprising 372 frames, as the primary coronary evaluation subset. Coronary vessels are present in 317 frames.

Non-vessel device annotations occur in $250/372$ frames across $6/9$ videos. We use the union of these device classes to quantify device-related false positives:

$$\mathrm{FP}_{\mathrm{device}} = \frac{|\widehat{Y} \cap D|}{|D|},$$

where $\widehat{Y}$ is the predicted binary vessel mask and D is the device mask.

Full-benchmark analysis. For completeness, we also evaluate all 12 MOSXAV videos. NeMeS remains superior to SAM2 on the full benchmark, achieving 52.7 versus 46.7 DSC. Its improvement is larger on the coronary subset (+9.3 DSC) than on the full benchmark (+6.0 DSC). Both models are penalized by the chamber-confound videos, indicating that their exclusion aligns the evaluation with the intended coronary-segmentation task rather than selectively favoring NeMeS.

B.2 XACV

XACV, introduced with DeNVeR Wu et al. (2025), contains 59 patients, 111 angiographic sequences, and 221 sparsely annotated frames with expert binary coronary-vessel masks. Catheter-only frames are excluded following the dataset protocol.

Because the exact VasoMIM Huang et al. (2026) partition is not publicly available, we follow its reported patient-level protocol by constructing three random splits, each containing 12 held-out test patients. The in-house-trained models are applied directly to each test split. No XACV data are used for training, fine-tuning, or checkpoint selection. DSC, clDice, and per-frame J&F are aggregated by sequence, and the mean and standard deviation are reported across the three splits.

C Implementation and Evaluation Details

C.1 Model Implementation

Architecture and initialization. Both branches are based on SAM2 Hiera-T and retain its image encoder, memory-attention module, mask decoder, and memory encoder. The prompt encoder is removed, and the mask decoder is invoked without prompt embeddings. Both branches are initialized from the public SAM2 Hiera-T checkpoint but use independent weights. For each frame, the image encoder produces a low-resolution feature for memory attention and higher-resolution features for the mask decoder, which outputs a logit map at the input resolution.

Memory construction. The proposer follows the prompt-free SAM2 propagation procedure. Its memory bank is initially empty and is updated causally by encoding each processed frame and its predicted mask into a memory token. Tokens are stored in a bounded FIFO memory bank with capacity $N_{\mathrm{mem}} = 7$, allowing each frame to attend to the preceding frames in the clip. In contrast, the suppressor uses a single fixed memory generated from the pre-contrast anchor. The anchor is encoded once and retained in memory slot 0, with no write-back from subsequent clip frames.

Gated prediction and supervision. The branches produce proposer and suppressor logits L_A and L_B, respectively, which are combined as $P = \sigma(L_A)\big(1 - \sigma(L_B)\big)$. The gated probability is

converted to a single logit and supervised against the expert vessel mask on the query frame using binary cross-entropy. No auxiliary, Dice, boundary, or deep-supervision losses are applied to either branch.

Parameters. The two branches are architecturally identical, with approximately 39M parameters each and 77.9M parameters in the complete dual-branch model—approximately twice the size of the single-branch prompt-free SAM2 baseline. The additional computational cost therefore corresponds to one independently parameterized branch. Its accuracy and robustness benefits are examined in main paper **Ablations**, and its runtime cost is reported in Tab. 5.

C.2 Training Setup

Both NeMeS branches are initialized from the public SAM2 Hiera-T checkpoint and optimized with independent weights. Input frames are converted to grayscale, resized to 512×512, and normalized independently using their per-frame mean and standard deviation. NeMeS processes causal clips of $T = 8$ frames, with the annotated query frame placed last. To expose the model to different temporal intervals and motion magnitudes, the inter-frame stride is sampled independently for each clip as

$$S \sim \mathcal{U}\{1, \ldots, 7\}.$$

The first frame of the corresponding angiographic sequence is used as the pre-contrast anchor.

We optimize NeMeS using AdamW with a learning rate of 10^{-5}, weight decay of 10^{-4}, $\beta = (0.9, 0.999)$, and gradient-norm clipping at 1.0. Training is performed for 100 epochs using distributed data parallelism on three GPUs, with a batch size of one clip per GPU and bfloat16 mixed precision. The checkpoint with the highest validation DSC is selected.

All learned baselines use the same training and validation partitions, image resolution, preprocessing, and checkpoint-selection criterion. Per-frame models receive the annotated query frame, whereas temporal models use the input window required by their respective architectures. Architecture-specific optimization settings follow the original publications when applicable.

C.3 Baseline Implementations

Per-frame models. nnU-Net Isensee et al. (2021) and ViT-CoMer Xia et al. (2024) predict the vessel mask from the annotated query frame without temporal context.

Temporal XCA models. The temporal model of He et al. He et al. (2025) receive preceding angiographic frames together with the query frame and produce a vessel mask for the final frame. Their temporal windows and feature-fusion modules follow the corresponding model definitions.

Memory-based models. TMANet Wang et al. (2021) is adapted to binary vessel segmentation using its temporal memory-attention mechanism. For SAM2 Ravi et al. (2025), we remove the prompt encoder and invoke the mask decoder without point, box, or mask prompts. Its memory bank is populated causally using predictions from preceding frames, matching the proposer branch of NeMeS.

External reported baseline. VasoMIM Huang et al. (2026) is included only on XACV using the values reported in its original publication. Because its exact patient partition and trained checkpoint are unavailable, it is presented as an external reference rather than as a directly reimplemented baseline.

C.4 Metric Definitions

Let $\widehat{Y}$ and Y denote the thresholded prediction and ground-truth vessel mask, respectively. Region overlap is measured using

$$\mathrm{DSC} = \frac{2|\widehat{Y} \cap Y|}{|\widehat{Y}| + |Y|}.$$

We additionally report clDice Shit et al. (2021), which measures agreement between the predicted and ground-truth vessel centerlines and is therefore sensitive to vascular connectivity.

The DAVIS-style J&F score Perazzi et al. (2016) is

$$\mathrm{J\&F} = \frac{J + F}{2},$$

where J is region intersection-over-union and F is the boundary F-measure. On sparsely annotated datasets, J&F is computed per annotated frame and then aggregated by sequence. On MOSXAV, it is averaged over all annotated video frames.

Device-specific false positives. Let D denote the ground-truth device mask. We measure the device misfire rate as

$$\mathrm{FP}_{\mathrm{device}} = \frac{|\widehat{Y} \cap D|}{|D|},$$

which estimates the probability that a labeled device pixel is incorrectly predicted as vessel.

We also report the enrichment of false positives on devices relative to ordinary background:

$$E_{\mathrm{device}} = \frac{|\widehat{Y} \cap D|/|D|}{|\widehat{Y} \cap B|/|B|},$$

where B contains non-vessel, non-device background pixels. Values greater than one indicate that false positives preferentially concentrate on devices.

Temporal-change accuracy. Per-frame accuracy does not verify whether a model captures the correct temporal dynamics. We therefore compare the predicted frame-to-frame change to the ground-truth change. For consecutive vessel-annotated frames, the signed mask difference $\Delta_t = Y_t - Y_{t-1} \in \{-1, 0, +1\}$ labels each pixel as newly appearing ($+1$, a vessel filling with contrast), disappearing (-1, washout), or unchanged (0). We compute the same signed difference on the predictions, $\widehat{\Delta}_t = \widehat{Y}_t - \widehat{Y}_{t-1}$, and score only the two *change* classes: for each of appear ($+1$) and disappear (-1) we take the pixel-exact intersection-over-union between the predicted and ground-truth change pixels (one-vs-rest), and report their mean,

$$\mathrm{mIoU}_{\mathrm{change}} = \tfrac{1}{2}(\mathrm{IoU}_{+1} + \mathrm{IoU}_{-1}),$$

pooled over all consecutive annotated pairs (pairs with no vessel in either frame are skipped). This metric is *ground-truth-referenced*: it credits legitimate contrast-driven change and penalizes temporal jitter. Two design choices are deliberate. First, we exclude the unchanged class (0); it accounts for over 99% of pixels and is trivially matched, so including it would saturate the score and hide differences between methods. Second, we use a *pixel-exact* IoU rather than a tolerance-based surface measure such as normalized surface Dice or a dilated change F-measure: coronary vessels are only a few pixels wide, so a matching tolerance comparable to or larger than the vessel would credit change predicted on the background adjacent to—but not on—the vessel, collapsing discrimination precisely on the thin, small-caliber vessels that dominate the hard cases. We therefore report the strict change IoU together with its appear and disappear components.

C.5 Ablation Protocols

Unless otherwise stated, retrained ablation variants use the same data partitions, preprocessing, optimization, and checkpoint-selection protocol as the main NeMeS model. Inference-time perturbations are applied to the fixed trained NeMeS checkpoint without further optimization.

Suppressor removal. Removing the anchor-conditioned suppressor yields the prompt-free SAM2 baseline, which retains only the causal video-memory branch.

Random-anchor capacity control. To separate the effect of the negative-memory mechanism from the additional model capacity, we retrain a parameter-matched two-branch variant in which the suppressor receives a randomly selected frame rather than the corresponding pre-contrast anchor.

Anchor-frame choice. Using the trained NeMeS model, we replace the pre-contrast anchor at inference with frames containing progressively greater vessel opacification. This experiment tests whether the benefit depends specifically on using a contrast-free reference.

Anchor perturbations. We test whether the suppressor uses the spatial content of the anchor by dividing it into an $N \times N$ grid and randomly permuting the resulting patches. We additionally evaluate an anchor selected from another sequence and a structureless random-noise anchor.

Pasted-distractor analysis. A radiopaque tubular distractor is extracted from a pre-contrast frame and pasted into a vessel-free, no-contrast clip. We measure the fraction of the pasted structure incorrectly predicted as vessel while varying its thickness. For NeMeS, we compare placing the distractor in both the anchor and clip with placing it in the clip alone.

Random-head linear probe. We freeze the image encoders and memory modules of NeMeS and SAM2, and randomly initialize the MLP output projections. The MLP output projection is trained to predict whether a frame contains a newly appearing vessel relative to the pre-contrast scene. This setup evaluates whether temporal novelty is accessible in the learned representation independently of the trained segmentation readout.

D Computational Cost

Model	Params (M)	Latency (ms)	Throughput (clips/s)
nnU-Net	10.9	22	44.6
ViT-CoMer	37.5	155	6.5
TMANet	58.3	77	13.0
SAM2	38.9	165	6.1
NeMeS (Ours)	77.9	343	2.9

Table 5: **Model cost** on an 8-frame 512×512 clip (single GPU, bfloat16): total parameters, median forward latency, and throughput. NeMeS is $\approx 2\times$ SAM2, reflecting its two branches.

We report the total parameter count, median forward-pass latency, and throughput for an 8-frame 512×512 clip with batch size one and bfloat16 inference on a single NVIDIA A100 (40GB) GPU. Throughput is reported in clips per second. NeMeS contains two independently parameterized SAM2 branches; therefore, its parameter count and inference cost are expected to be approximately twice those of the single-branch SAM2 baseline.

E Supplementary Results

Device false positives. Because our binary-vessel models treat every non-vessel class as background, any vessel prediction landing on a labeled device is a pure out-of-distribution false positive. Tab. 6 reports the device misfire rate $\mathrm{FP}_{\mathrm{device}}$ and the device-vs-background enrichment E_{device} (defined in Sec. C.4) on the MOSXAV coronary subset (n=317, zero-shot). NeMeS attains the lowest device false-positive rate and the least enrichment, reducing the SAM2 misfire rate from 0.216 to $\mathbf{0.019}$ (an $\approx 11\times$ reduction) and the enrichment from $21.9\times$ to $3.0\times$.

Model	$\mathrm{FP}_{\mathrm{device}} \downarrow$	$E_{\mathrm{device}} \downarrow$
ViT-CoMer	0.164	35.9×
nnU-Net	0.275	18.5×
TVS-Net	0.352	23.5×
TMANet	0.035	4.6×
SAM2	0.216	21.9×
NeMeS (Ours)	**0.019**	**3.0**×

Table 6: **Device false positives on MOSXAV** (cleaned coronary subset, n=317; zero-shot). $\mathrm{FP}_{\mathrm{device}} = |\widehat{Y} \cap D|/|D|$ is the fraction of labeled device pixels predicted as vessel; E_{device} is the device-vs-background FP enrichment (both defined in the Metric Definitions above). NeMeS has the lowest device misfire rate and the least enrichment.

Temporal-change accuracy. We assess temporal behavior with a ground-truth-referenced change metric on the densely annotated MOSXAV coronary subset (zero-shot). Comparing the predicted

signed frame-to-frame mask difference against the ground-truth change (Tab. 7), NeMeS predicts which vessels newly appear or disappear far more accurately than every baseline, reaching $\mathrm{mIoU}_{\mathrm{change}} = 0.348$ versus 0.285 for the next-best TMANet and 0.247 for SAM2. Because the metric is ground-truth-referenced and scored at pixel-exact tolerance, this reflects correct localization of contrast-induced novelty *on* the thin vessels themselves, rather than temporally smooth or approximately placed masks.

Model	$\mathrm{mIoU}_{\mathrm{change}} \uparrow$	IoU_{+1}	IoU_{-1}
ViT-CoMer	0.270	0.270	0.269
nnU-Net	0.192	0.193	0.191
TVS-Net	0.159	0.163	0.155
TMANet	0.285	0.288	0.282
SAM2	0.247	0.250	0.243
NeMeS (Ours)	**0.348**	**0.347**	**0.348**

Table 7: **Temporal-change accuracy on MOSXAV** (cleaned coronary subset, 316 consecutive annotated pairs; zero-shot; GT-referenced). $\mathrm{mIoU}_{\mathrm{change}}$ is the mean of the pixel-exact appear (IoU_{+1}) and disappear (IoU_{-1}) IoU of the signed mask difference $Y_t - Y_{t-1}$ (unchanged pixels excluded; no spatial tolerance—see Metric Definitions). NeMeS predicts frame-to-frame vessel change most accurately, symmetrically for appearance and disappearance.