ANGIOFM: A VISION FOUNDATION MODEL FOR X-RAY CORONARY ANGIOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

X-ray coronary angiography (CAG) presents complex scenes in which coronary vessels, surrounding anatomy, and interventional devices overlap in two-dimensional projections. Cardiac motion, contrast filling, and changing projection geometry further complicate representation learning. We introduce **AngioFM**, a vision foundation model that uses these structured variations as supervision through feature prediction conditioned on acquisition time, contrast state, and imaging rays. Pretraining combines 31,559 unlabeled coronary videos with projections from 1,000 coronary CT angiography (CCTA) cases, coupling real angiographic dynamics with explicit cross-view, vessel-tangent, and branch supervision. We assess frozen representations using linear probes for controlled comparisons of accessible pathological, anatomical, and temporal information, complemented by CNN decoders for anatomical and device segmentation and a direct centerline-regression readout. With matched readouts and frozen encoders, AngioFM improves coronary-segment parsing over generic image-pretrained and CAG-pretrained baselines, with additional gains in video-based catheter and guidewire segmentation. Linear probes show higher coronary-dominance AUC, more accurate cardiac-phase matching, and competitive lesion localization. Centerline regression shows improved vessel localization in single-view and multi-view settings. Branch-supervision ablations further support improved anatomical discrimination. Together, these findings support learning reusable coronary representations from acquisition physics and vascular structure, providing a foundation for tools supporting angiographic assessment and image-guided intervention.

1 INTRODUCTION

X-ray coronary angiography (CAG) observes a moving three-dimensional coronary tree through two-dimensional projections. As illustrated in Fig. 1, cardiac motion, contrast propagation, and C-arm viewpoint change vessel appearance without necessarily changing the underlying anatomy. Conversely, anatomically distinct branches can overlap in the image plane. Useful CAG representations should therefore retain local pathology, global anatomy, temporal state, and vessel continuity rather than relying only on local appearance.

Our premise is that these variations are *structured*, rather than arbitrary nuisance factors. Multiple cine frames supply complementary observations as motion and contrast change vessel visibility. CCTA-derived coronary tree masks additionally provide controlled projections with known three-dimensional branch relationships. Real cine and synthetic views thus provide complementary supervision for temporal, contrast-dependent, and cross-view representation learning.

We formulate CAG pretraining as *conditional feature prediction*. Given context observations, AngioFM predicts target features conditioned on acquisition time, contrast state, and imaging rays. An exponential-moving-average teacher supplies representation-space targets (Assran et al., 2023; Bardes et al., 2024). Synthetic projections also supervise vessel tangents and branch embeddings, encouraging anatomical discrimination despite projection overlap. Inherited self-distillation and Gram anchoring retain useful appearance structure during this adaptation (Caron et al., 2021; Siméoni et al., 2025; Zhou et al., 2021).

We evaluate **AngioFM** through frozen-feature readouts spanning lesions, coronary dominance, cardiac phase, dense masks, and direct centerline regression. Completed experiments show higher

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.

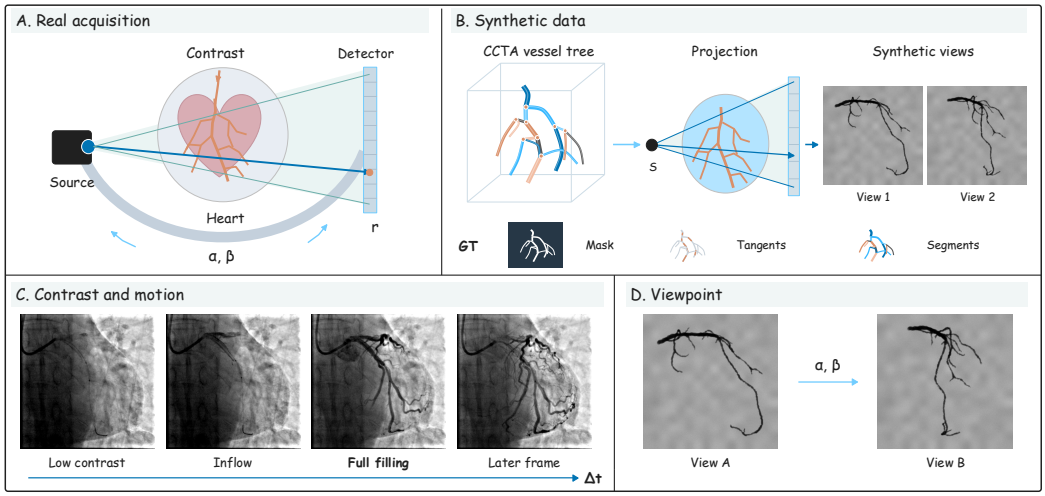

Figure 1: **Motivation of AngioFM.** CAG appearance depends on anatomy, cardiac motion, contrast filling, and projection geometry. Real cine supplies temporal and contrast variation; CCTA-derived trees supply controlled views with known vessel structure. AngioFM uses both for conditional feature prediction and structural supervision.

video CARDIAG segment Dice than DINO-world (44.01 versus 41.91) and lower phase-matching error (1.38 versus 1.50 frames). These same-setting comparisons complement single-frame evaluation; additional observations are not treated as proof of temporal reasoning. The initialization control is in Appendix D.1, and component ablations are in Sec. 6. Dominance and phase findings retain the split and study-composition qualifications reported below; centerline results are reported in Sec. 5.3.

Our contributions are threefold:

- Conditional CAG representation prediction using acquisition time, contrast state, and imaging-ray geometry.
- Complementary real-cine and CCTA-derived supervision for temporal, cross-view, tangent, and branch structure.
- Frozen-feature evaluation across local pathology, global anatomy, temporal correspondence, and dense structure, separating encoder comparisons from initialization, component, and input-context ablations.

2 Related Work

Coronary angiography representation learning. CathAI combines projection recognition, anatomical identification, and stenosis analysis (Avram et al., 2023). VasoMIM introduces vascular priors into masked image modeling (Huang et al., 2026), DeepCORO-CLIP learns from paired angiographic videos and reports (Harrabi et al., 2026), and FIMAE exploits temporal interpolation in interventional X-ray video (Islam et al., 2024). AngioFM complements these approaches by conditioning dense feature prediction explicitly on acquisition state and geometry, with transfer assessed across pathology, anatomy, time, and devices.

Predictive and temporal representation learning. MAE reconstructs image content (He et al., 2022), while DINO-style self-distillation learns transferable visual features (Caron et al., 2021; Siméoni et al., 2025). I-JEPA, V-JEPA, and image world models predict representations from context (Assran et al., 2023; Bardes et al., 2024; Garrido et al., 2024). VideoMAE, Echo-SyncNet, and EchoFM exploit temporal or physiological structure (Tong et al., 2022; Dezaki et al., 2021; Kim

et al., 2025). AngioFM separates relative acquisition time from contrast opacification and combines both with viewing-ray geometry, making motion, visibility, and projection explicit conditioning variables.

Geometry-aware coronary learning. Projection geometry supports coronary reconstruction and correspondence (Çimen et al., 2016). NeRF-CA reconstructs dynamic coronary structure from sparse views (Maas et al., 2025); multi-view and CCTA-based methods establish cross-projection correspondences (Wu et al., 2023; Zhao et al., 2025). AngioFM instead uses geometry as a representation-learning prior: rays condition feature prediction, and synthetic projections supervise cross-view features, vessel tangents, and branches.

3 METHOD

3.1 CONDITIONAL FEATURE PREDICTION

AngioFM models angiography as observations of the same underlying coronary structure under changing acquisition state and projection geometry. Let $I_{1:M}$ denote M context frames and $I^\star$ a target frame. A student ViT encoder f_θ (Dosovitskiy, 2020) maps each context frame to patch tokens, while an exponential-moving-average (EMA) encoder $f_{\bar\theta}$ provides stop-gradient target features. A geometry-aware predictor g_ψ estimates target tokens from context features and target queries:

$$Z_m = f_\theta(\widetilde{I}_m), \qquad \bar{Z}^\star = \mathrm{LN}(f_{\bar\theta}(I^\star)), \qquad \widehat{Z}^\star = g_\psi(\{Z_m, \mathcal{C}_m\}_{m=1}^M, \mathcal{Q}^\star). \tag{1}$$

Here, $\widetilde{I}_m$ may contain masked patches, $\mathcal{C}_m$ describes the context coordinates, and $\mathcal{Q}^\star$ describes the target queries. The student encoder and predictor are optimized jointly, while the teacher follows the EMA update.

Imaging rays. For each patch center, the known C-arm pose and acquisition geometry define a 3D line in patient coordinates. We represent this ray by

$$r = (d, p/s_p) \in \mathbb{R}^6, \tag{2}$$

where d is its unit direction, p is the point on the line closest to the isocenter, and $s_p = 100\,\mathrm{mm}$. Context and query rays are expressed relative to the first context view, so the predictor receives relative rather than absolute viewing geometry.

Acquisition state. Each frame is associated with

$$a = (\Delta t, \kappa), \tag{3}$$

where Δt denotes normalized time relative to the target and $\kappa \in [0, 1]$ denotes continuous contrast opacification. Separating Δt and κ allows the predictor to distinguish temporal displacement from contrast-dependent visibility.

Conditioned queries and geometric attention. Ray and acquisition-state variables are encoded by Fourier-feature MLPs (Tancik et al., 2020). Context tokens and target queries are

$$h_{m,i} = W_c Z_{m,i} + E_r(r_{m,i}) + E_a(a_m), \tag{4}$$

$$q_j = q_0 + E_r(r_j^\star) + E_a(a^\star). \tag{5}$$

A 12-layer cross-attention transformer (Vaswani et al., 2017) predicts target features from context tokens. The geometric bias uses a *query-segment-to-context-line distance*, not the shortest distance between two unrestricted lines. Let $\Gamma_j^\star$ be the finite query-ray segment in the $\pm 100\,\mathrm{mm}$ isocenter slab, with the X-ray source excluded. In physical millimetres,

$$D(r_j^\star, r_{m,i}) = \min_{\substack{x \in \Gamma_j^\star \\ u \in \mathbb{R}}} \left\| x - (p_{m,i} + u d_{m,i}) \right\|_2 . \tag{6}$$

The attention logit is

$$A_{j,mi}^{(h)} = \frac{\langle W_q^{(h)} q_j, W_k^{(h)} h_{m,i} \rangle}{\sqrt{d_h}} - \lambda_h \left(\frac{D(r_j^\star, r_{m,i})}{20\,\mathrm{mm}} \right)^2, \qquad \lambda_h = \mathrm{softplus}(\alpha_h). \tag{7}$$

Unlike unrestricted line-to-line distance, the finite-segment distance excludes the common source and preserves same-view locality. Across views it gives a soft depth-restricted epipolar prior. Segment construction, the closed form, and the same-view derivation are provided in Appendix A.2.

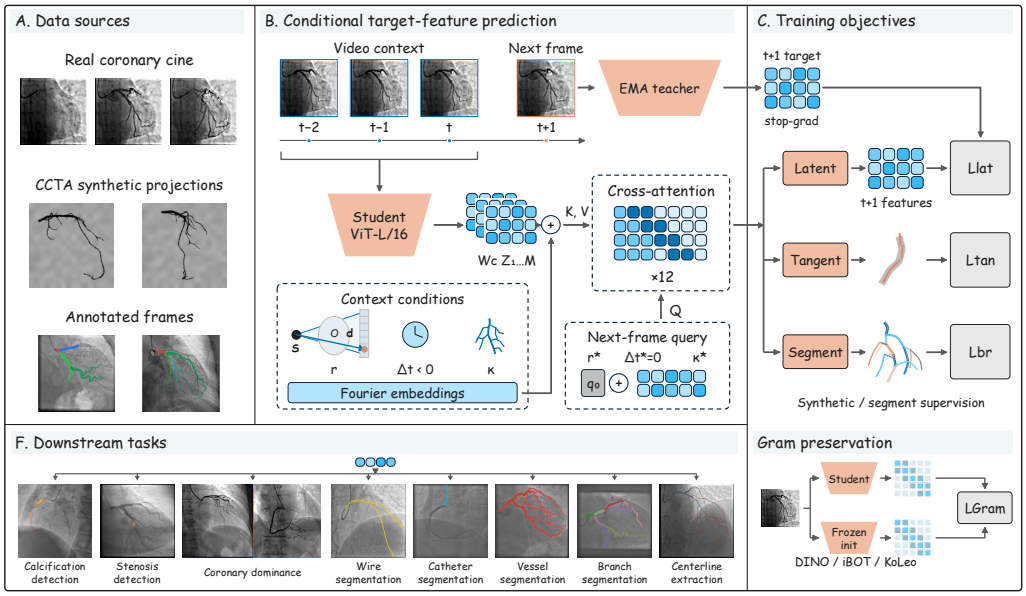

Figure 2: **Overview of AngioFM pretraining and downstream evaluation.** (A) Real coronary cine and CCTA-derived synthetic projections provide pretraining data; annotated angiographic frames support downstream evaluation. (B) A student ViT encodes context observations, and a cross-attention predictor estimates EMA-teacher target features conditioned on acquisition time, contrast state, and imaging-ray geometry. (C) Feature prediction comprises masked-anchor prediction, contrast translation, and synthetic cross-view prediction, complemented by vessel-tangent and branch supervision, inherited DINO/iBOT/KoLeo objectives, and Gram anchoring. (D) The frozen representation is evaluated through lesion localization, coronary dominance, device and vessel segmentation, branch segmentation, and direct centerline extraction.

3.2 Pretraining Tasks

The feature-prediction tasks use task-specific targets: masked patches within the same frame, an opacified frame from the same acquisition, or a held-out synthetic view. These tasks share the encoder, predictor, and latent feature objective, and are complemented by vessel-tangent and branch supervision and inherited self-distillation. Pretraining data and task-sampling probabilities are provided in Appendices A.1 and A.3.

Masked anchor prediction. For a real angiographic frame, 50% of context patches are block-masked and the model predicts EMA target features at the hidden locations. This provides an appearance-preserving within-frame objective without requiring pixel reconstruction.

Contrast translation. The context contains up to four contrast-free frames and the target is an opacified frame from the same acquisition. The model receives relative time, contrast state, and imaging rays and predicts the target representation. The task therefore asks the model to infer where the coronary tree becomes visible when contrast arrives, encouraging vessel-aware features rather than features tied only to observed lumen intensity.

Cross-view prediction. CCTA-derived coronary trees are analytically rendered from multiple C-arm views with known camera geometry, vessel masks, and tangent fields. The model observes one to four context projections and predicts a held-out view; the view difference is progressively increased during training. Because the same 3D tree is observed under controlled projection changes, this task supplies cross-view geometric supervision unavailable from an isolated 2D angiogram.

Branch-level structural supervision. Synthetic centerline trees are divided into branch segments at bifurcations. A branch embedding head is trained with supervised contrastive learning (Khosla

et al., 2020), treating patches from the same segment as positives and different segments as negatives within a consistent annotation scheme. The model is not required to recover a semantic branch name. Instead, this objective encourages features to distinguish vessel branches that may be adjacent or superimposed after projection. Synthetic data also supervise an unoriented vessel-tangent field, providing local orientation information.

3.3 Representation Objectives

All conditional-prediction tasks use a scale-free cosine loss,

$$\mathcal{L}_{\text{lat}} = \frac{1}{|\Omega|} \sum_{j \in \Omega} \left(1 - \frac{\widehat{z}_j^\top \bar{z}_j}{\|\widehat{z}_j\|_2 \|\bar{z}_j\|_2} \right), \tag{8}$$

where Ω denotes the task-specific set of supervised patches. Coronary-specific prediction is interleaved with the inherited DINO/iBOT self-distillation objectives (Caron et al., 2021; Zhou et al., 2021), KoLeo regularization (Sablayrolles et al., 2018), and multi-resolution Gram anchoring to the initialization encoder. These terms reduce representational drift while the model acquires angiography-specific temporal and geometric structure.

Vessel structural objectives. Synthetic vessel tangents are represented by $u_j = (\cos 2\vartheta_j, \sin 2\vartheta_j)$, making the target invariant to a 180° reversal; normalized predictions use squared error. For branch embeddings e_j, let $P(j)$ denote positive patches from the same branch segment and $A(j)$ the valid contrast set under the same annotation scheme. Supervised contrastive learning (Khosla et al., 2020) gives

$$\mathcal{L}_{\text{br}} = -\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{1}{|P(j)|} \sum_{p \in P(j)} \log \frac{\exp(e_j^\top e_p / \tau)}{\sum_{a \in A(j)} \exp(e_j^\top e_a / \tau)}, \qquad \tau = 0.1. \tag{9}$$

Only patches with sufficient vessel occupancy and branch purity contribute, preventing predominantly background or ambiguous projected regions from defining branch targets. For row-normalized student and frozen-initialization tokens U and U_0, Gram anchoring preserves patch-wise similarity:

$$\mathcal{L}_{\text{Gram}} = \frac{1}{N^2} \left\| UU^\top - U_0 U_0^\top \right\|_F^2. \tag{10}$$

The structural losses supervise vessel organization, whereas Gram anchoring limits changes to the initialization's pairwise feature relationships.

4 Evaluation

We evaluate AngioFM on downstream CAG tasks spanning lesion analysis, anatomical characterization, temporal correspondence, and interventional-device analysis. The encoder remains frozen across all tasks, allowing us to evaluate transfer without task-specific encoder adaptation.

We use readouts of varying capacity. *Linear Probing* uses lightweight heads to measure whether lesion, temporal, and study-level information is directly accessible. *CNN Decoding* assesses whether the features support pixel-level anatomical and device predictions. *Centerline Regression* tests whether frozen features encode vascular connectivity strongly enough for direct vessel centerline prediction. Conventional pipelines typically derive centerlines indirectly from segmentation masks without an explicit representation of centerline which favors its connectivity.

We summarize the evaluation protocols below; Appendix B.1 describes the downstream datasets, and Appendix B provides task-specific settings.

Linear Probing We first test what information is directly accessible from the frozen representation using lightweight linear readouts. The probes span three complementary aspects of angiographic understanding: lesion localization tests whether local pathological features are spatially encoded; coronary-dominance classification tests whether the representation captures global coronary anatomy; and cardiac-phase matching tests whether it preserves temporal information across the cardiac cycle. We evaluate both single-frame and multi-frame representations where applicable.

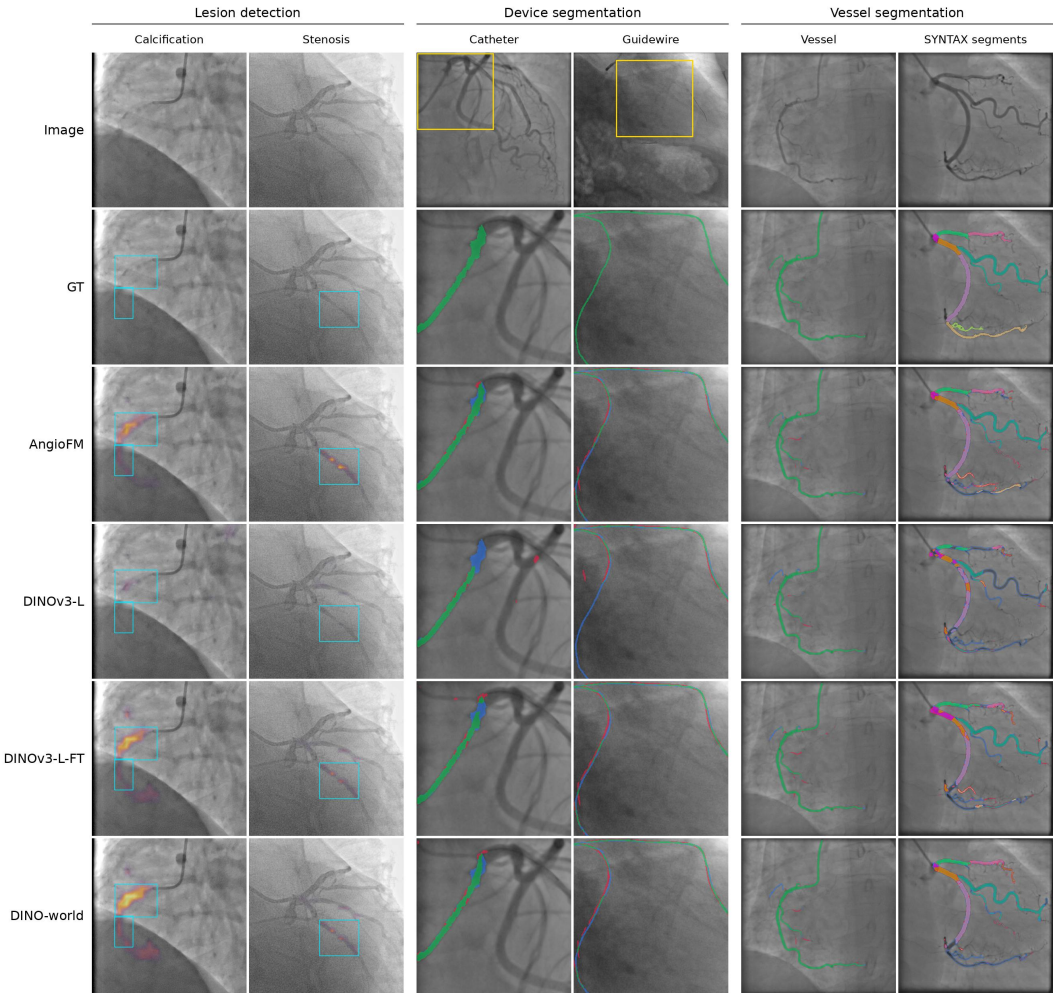

Figure 3: **Qualitative comparison of frozen encoders on six angiography tasks.** Each column shows one test target; rows show the input, ground truth (GT), and predictions from AngioFM, DINOv3-L, DINOv3-L-FT, and DINO-world. Readouts are matched within each task: 1×1 lesion probes or CNN-256 segmentation decoders. AngioFM and DINO-world use video where available (target plus three black context frames for single image); both DINOv3 variants use single frames. Cyan boxes mark lesions; probability overlays share an inferno scale from purple (low) to yellow (high), with values below 0.05 transparent. Yellow boxes indicate enlarged device regions. For binary masks, green/red/blue denote true-positive/false-positive/false-negative pixels; GT masks are green. SYNTAX colours identify segments consistently across GT and predictions; red outlines mark incorrect segment assignments, and blue outlines mark vessel pixels predicted as background.

CNN Decoding CNN decoding tests whether frozen representations retain sufficient spatial detail for pixel-level prediction when paired with a higher-capacity task head. We evaluate vessel and coronary-segment segmentation, which probe anatomical structure, together with catheter and guidewire segmentation, which probe interventional-device representation. Single-frame and multi-frame settings are evaluated separately where temporal data are available.

Centerline Regression We formulate centerline extraction as structured set prediction rather than mask segmentation. Each target is an anatomically labeled vessel path represented as an ordered sequence from the coronary root to a branch endpoint; paths sharing a parent vessel therefore share their upstream trajectory. A set of learned queries predicts the vessel identity and two-dimensional spline control points for each path, with an additional no-object class for unmatched queries. The

control points define a continuous Catmull–Rom spline, which is sampled at multiple ordered locations for supervision. Hungarian matching associates predicted and reference paths using vessel-class and ordered-point costs, followed by pointwise regression of the matched curves.

5 RESULTS

Baseline pretraining. DINO-world (Baldassarre et al., 2025) and LeVJEPA (Kuhn et al., 2026) are pretrained on our private CAG data before downstream evaluation. DINOv3-L (Siméoni et al., 2025) refers to the original pretrained model without additional CAG-specific pretraining. Checkpoint provenance and initialization controls are detailed in Appendices A.4 and D.1.

5.1 LINEAR PROBING

Table 1: **Frozen-feature linear probing.** DINO-world and LeVJEPA are CAG-pretrained. Lesions report step AP (%); phase reports anchor-weighted mean ± SD with 16-frame input and five-embedding-window matching. ↓: significantly worse than AngioFM, irrespective of metric direction, under 20,000 paired bootstrap resamples, raw two-sided $p < 0.05$, without multiplicity correction. Supplementary and initialization results are in Appendices C and D.1.

	Single-frame lesion AP		Video lesion AP
Method	In-house calcification	ARCADE stenosis	In-house calcification
AngioFM	**30.91**	**76.38**	**31.63**
DINOv3-L	20.42↓	69.99↓	—
DINO-world	27.38	75.26	30.55
LeVJEPA	12.67↓	40.99↓	14.43↓
Cardiac-phase matching (lower is better)			
Method	Dist. (frames)	Dist. (% of cycle)	
AngioFM	$\mathbf{1.38} \pm 1.57$	$\mathbf{14.18} \pm 15.58$	
DINO-world	1.50 ± 1.62↓	15.28 ± 15.98↓	
LeVJEPA	1.48 ± 1.56↓	15.34 ± 15.54↓	

Lesion localization. AngioFM reaches 30.91 AP on in-house calcification and 76.38 on ARCADE (Table 1). Generic DINOv3-L and the evaluated LeVJEPA readouts score significantly lower; DINO-world differences are not significant. Video calcification AP is 31.63, versus 30.55 for DINO-world (nonsignificant) and 14.43 for LeVJEPA (significantly lower). These LeVJEPA results use the CAG-pretrained checkpoint and downstream resolution specified in Appendix B.4. Box-trained detector references are reported separately in Appendix C.3: their box centers need not satisfy the point-localization rule even when a box covers a lesion.

Cardiac-phase matching. AngioFM gives 1.38 ± 1.57 frames ($14.18 \pm 15.58\%$ of a cycle), compared with 1.50 ± 1.62 for DINO-world and 1.48 ± 1.56 for LeVJEPA. Both comparisons are significant in both units under study resampling. However, two studies contribute 58% of anchors; excluding them preserves the direction but not significance ($p \geq 0.09$). The quoted DCR/AIT expert distances use a different test list and matching task and are not paired comparisons (Appendix B.6).

Coronary dominance. Video AngioFM reaches 91.05% accuracy, 91.85% AUC, and 0.622 MCC on the real-distribution set; DINO-world has significantly lower AUC and MCC, with the AUC test narrowly below threshold ($p = 0.045$). Under domain shift, AUC is 67.40% versus 60.91% for DINO-world (significant), whereas accuracy and MCC differences are not. These released splits contain acquisition overlap and duplicate or contradictory studies, so the results do not establish independently verified out-of-domain generalization. The complete comparison and audit are in Table 4 and Appendix B.5.

5.2 CNN DECODING

We next assess dense anatomical and device segmentation using the same baseline checkpoints, with frozen encoders and matched CNN-256 readouts.

Table 2: **Single-frame CNN-256 decoding: Dice (%).** All encoders are frozen; DINO-world and LeVJEPA are CAG-pretrained. Binary Dice pools test pixels; segment Dice averages class-wise pooled scores over present classes. ↑/↓: significantly better/worse than AngioFM under paired image bootstrap, raw two-sided $p < 0.05$. Initialization controls and published references are in Appendices D.1 and C.2.

Dataset	Target	AngioFM	DINOv3-L	DINO-world	LeVJEPA
ARCADE	Coronary segments	**40.52**	34.80↓	36.42↓	32.89↓
CARDIAG	Coronary segments	**43.30**	38.98↓	41.12↓	34.22↓
CARDIAG	Vessel	**88.11**	87.42↓	88.05	86.68↓
XCAD	Vessel	**82.83**	81.37↓	82.37	79.97↓
XACV	Vessel	86.92	86.00↓	**87.10**	83.38↓
CARDIAG	Catheter	84.57	81.90↓	**84.87**	77.19↓
XACV	Catheter	**87.29**	82.73↓	86.40	74.49↓
In-house wire	Guidewire	60.43	59.73↓	**60.74**	55.84↓

Table 3: **Video CNN-256 decoding: Dice (%).** Metrics and statistical conventions follow Table 2. Bold denotes the highest available point estimate in each row.

Dataset	Target	AngioFM	DINO-world	LeVJEPA
CARDIAG	Coronary segments	**44.01**	41.91↓	33.40↓
CARDIAG	Vessel	**88.09**	87.87↓	86.45↓
XACV	Vessel	87.04	**87.10**	82.87↓
CARDIAG	Catheter	**85.04**	84.29↓	77.93↓
XACV	Catheter	**87.59**	86.13	72.19↓
In-house wire	Guidewire	**60.67**	60.34↓	55.49↓

Single-frame segment Dice reaches 40.52 on ARCADE and 43.30 on CARDIAG (Table 2), significantly above every other completed encoder in those rows. Relative to generic DINOv3-L, the differences are 5.72 and 4.32 points. XACV catheter Dice reaches 87.29, versus 82.73 for DINOv3-L and 74.49 for LeVJEPA. Gains are not uniform: DINO-world does not differ significantly on single-frame binary vessel, catheter, or guidewire tasks and has higher point estimates on several rows. Thus, the strongest dense evidence is for segment-level anatomical discrimination, rather than uniform improvement on every structure.

In the video setting (Table 3), AngioFM exceeds DINO-world on CARDIAG segments (44.01 versus 41.91), vessels (88.09 versus 87.87), catheters (85.04 versus 84.29), and in-house guidewires (60.67 versus 60.34), all at raw $p < 0.05$. XACV differences are not significant. All evaluated LeVJEPA decoders score significantly below AngioFM. These same-setting comparisons do not establish a within-model temporal gain. XCAD/XACV contain only 26/34 test images, and the 640 guidewire frames come from 11 sequences whereas the reported bootstrap resamples images. Full protocols, same-split nnU-Net comparisons, and unpaired literature references are in Appendices B.7 and C.2.

5.3 CENTERLINE REGRESSION

AngioFM achieves lower Chamfer distance than DINOv3-L and DINO-world on both in-domain ARCADE-SYNTAX (3.25 vs. 4.36/3.94 px) and out-of-domain CARDIAG (7.27 vs. 9.72/8.07 px). In multi-view evaluation, AngioFM also achieves lower Chamfer distance than DINO-world (3.87 vs. 5.39 px). These results indicate that AngioFM provides stronger geometric representations for recovering vessel routes, including under dataset shift. More details are in Appendix C.4.

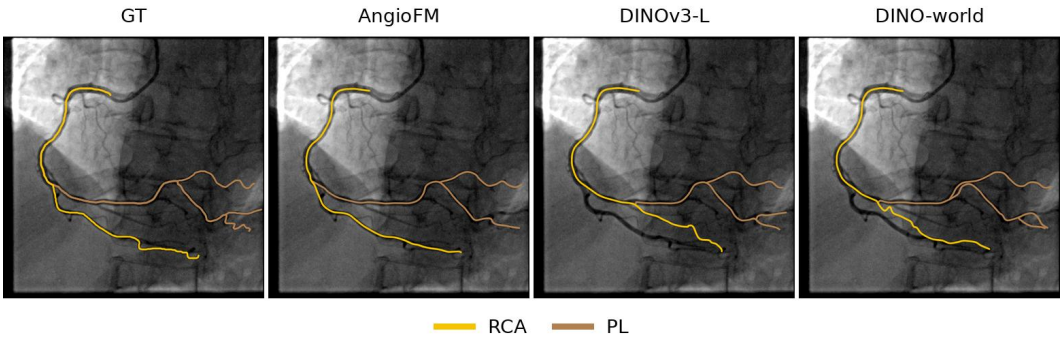

Figure 4: **Single-view centerline regression.** AngioFM better follows the annotated vessel, while DINOv3-L and DINO-world show branch drift and off-vessel trajectories.

6 Ablation Studies

Branch supervision. Removing branch supervision reduces segment Dice from 40.52 to 37.86 on ARCADE, from 43.30 to 41.47 on single-frame CARDIAG, and from 44.01 to 42.31 on video CARDIAG. All three differences are significant under paired image bootstrap (raw $p < 0.05$). This supports a task-dependent benefit for vessel identity and branch structure (Appendix D.2).

Physical conditioning. Replacing physical ray and acquisition-state conditioning with positional embeddings increases mean Chamfer distance from 3.87 to 5.43 pixels in multi-view centerline regression (paired difference $+1.56$ pixels, 95% CI $[1.36, 1.77]$, $p < 10^{-4}$) and decreases recall from 89.04% to 85.99%. Branch identity changes little (96.07% versus 96.27%), suggesting that the observed benefit primarily concerns geometric localization and route recovery rather than vessel classification. This remains a continuation comparison with different pretraining history and budget; full results are in Appendix D.4.

7 Conclusion

We presented AngioFM, a vision foundation model for coronary angiography that conditions feature prediction on acquisition time, contrast state, and imaging-ray geometry. By combining real cine with CCTA-derived projections and branch and tangent supervision, AngioFM uses structured acquisition variation and explicit vascular geometry to learn coronary representations. Linear probes provide controlled comparisons of the pathological, anatomical, and temporal information accessible in frozen representations; complementary decoder evaluations assess their utility for segmentation and centerline regression. Together, these evaluations support using the physical and anatomical structure of angiography to learn representations reusable across tasks. Clinically, this work provides a foundation for developing tools that combine lesion assessment, anatomical mapping, and device localization to support angiographic interpretation and image-guided intervention.

AI USE STATEMENT

Generative AI tools were used solely for language polishing, formatting checks, and coding assistance, including support for implementing and running experiments. The authors take full responsibility for the content of this manuscript.

REPRODUCIBILITY STATEMENT

The pretraining formulation is specified in Sec. 3 and Appendix A. Appendix B provides readout, metric, statistical, and split-audit details. Supplementary comparisons and ablation protocols are organized in Appendices C and D, with unavailable results explicitly distinguished from completed measurements.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Robert Avram, Jeffrey E Olgin, Zeeshan Ahmed, Louis Verreault-Julien, Alvin Wan, Joshua Barrios, Sean Abreau, Derek Wan, Joseph E Gonzalez, Jean-Claude Tardif, et al. Cathai: fully automated coronary angiography interpretation and stenosis estimation. *NPJ Digital Medicine*, 6(1):142, 2023.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Serkan Çimen, Ali Gooya, Michael Grass, and Alejandro F Frangi. Reconstruction of coronary arteries from x-ray angiography: A review. *Medical image analysis*, 32:46–68, 2016.

Fatemeh Taheri Dezaki, Christina Luong, Tom Ginsberg, Robert Rohling, Ken Gin, Purang Abolmaesumi, and Teresa Tsang. Echo-syncnet: Self-supervised cardiac view synchronization in echocardiography. *IEEE Transactions on Medical Imaging*, 40(8):2092–2104, 2021.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Sarra Harrabi, Yichen Wu, Geoffrey H Tison, Minhaj Ansari, Milos Vukadinovic, David Ouyang, Joshua P Barrios, Jacques Delfrate, and Robert Avram. Deepcoro-clip: A multi-view foundation model for comprehensive coronary angiography video-text analysis and external validation. *arXiv preprint arXiv:2603.17675*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

De-Xing Huang, Xiao-Hu Zhou, Xiao-Liang Xie, Shi-Qi Liu, Shuang-Yi Wang, Zhen-Qiu Feng, Mei-Jiang Gui, Hao Li, Tian-Yu Xiang, Bo-Xian Yao, et al. Spironet: Spatial-frequency learning and graph-based channel interaction network for vessel segmentation. *arXiv preprint arXiv:2406.19749*, 2024.

De-Xing Huang, Chaohui Yu, Xiao-Hu Zhou, Tian-Yu Xiang, Qin-Yi Zhang, Mei-Jiang Gui, Rui-Ze Ma, Chen-Yu Wang, Nu-Fang Xiao, Fan Wang, et al. Vascular anatomy-aware self-supervised pre-training for x-ray angiogram analysis. *arXiv preprint arXiv:2602.11536*, 2026.

Saahil Islam, Venkatesh N Murthy, Dominik Neumann, Badhan Kumar Das, Puneet Sharma, Andreas Maier, Dorin Comaniciu, and Florin C Ghesu. Self-supervised learning for interventional image analytics: toward robust device trackers. *Journal of Medical Imaging*, 11(3):035001–035001, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Sekeun Kim, Pengfei Jin, Sifan Song, Cheng Chen, Yiwei Li, Hui Ren, Xiang Li, Tianming Liu, and Quanzheng Li. Echofm: Foundation model for generalizable echocardiogram analysis. *IEEE transactions on medical imaging*, 44(10):4049–4062, 2025.

Ivan Kruzhilov, Gleb Mazanov, Alexander Ponomarchuk, Galina Zubkova, Artem Shadrin, Ruslan Utegenov, Pavel Blinov, and Ivan Bessonov. Coronarydominance: Angiogram dataset for coronary dominance classification. *Scientific Data*, 12(1):341, 2025.

Lukas Kuhn, Lucas Maes, Giuseppe Serra, Quentin Le Lidec, Yann LeCun, Randall Balestriero, and Florian Buettner. Levjepa: Efficient & scalable video pretraining without the heuristics. *arXiv preprint arXiv:2608.27395*, 2026.

Dominik Bernard Lau, Hubert Malinowski, Jerzy Szyjut, Adam Brzeski, Tomasz Dziubich, Radosław Targoński, Tomasz Figatowski, and Natalia Zielińska. Cardiag: A dense segment classification benchmark of deep learning architectures for coronary angiography. *arXiv preprint arXiv:2607.22139*, 2026.

Yikang Liu, Lin Zhao, Eric Z Chen, Xiao Chen, Terrence Chen, and Shanhui Sun. Auxiliary input in training: Incorporating catheter features into deep learning models for ecg-free dynamic coronary roadmapping. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 67–77. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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 (ICCV)*, pp. 7536–7545, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ma_Self-Supervised_Vessel_Segmentation_via_Adversarial_Learning_ICCV_2021_paper.html.

Kirsten WH Maas, Danny Ruijters, Anna Vilanova, and Nicola Pezzotti. Nerf-ca: dynamic reconstruction of x-ray coronary angiography with extremely sparse-views. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):8782–8795, 2025.

Maxim Popov, Akmaral Amanturdieva, Nuren Zhaksylyk, Alsabir Alkanov, Adilbek Saniyazbekov, Temirgali Aimyshev, Eldar Ismailov, Ablay Bulegenov, Arystan Kuzhukeyev, Aizhan Kulanbayeva, et al. Dataset for automatic region-based coronary artery disease diagnostics using x-ray angiography images. *Scientific data*, 11(1):20, 2024.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: Neural architecture search for real-time detection transformers. In *International Conference on Learning Representations*, volume 2026, pp. 56087–56100, 2026.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. Yolo26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Dinh-Son Tran, Anh-Khoa Huynh, Anh-Duy Huynh, and Trung Nguyen-Thoi. Anatomy-specific two-stage yolov8 approach for improved coronary segmentation using the arcade dataset. *Optics Continuum*, 4(2):303–317, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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 (CVPR)*, pp. 15682–15692, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_DeNVeR_Deformable_Neural_Vessel_Representations_for_Unsupervised_Video_Vessel_Segmentation_CVPR_2025_paper.html.

Yifan Wu, Rohit Jena, Mehmet Gulsun, Vivek Singh, Puneet Sharma, and James C Gee. Towards establishing dense correspondence on multiview coronary angiography: From point-to-point to curve-to-curve query matching. *arXiv preprint arXiv:2312.11593*, 2023.

An Zeng, Chunbiao Wu, Guisen Lin, Wen Xie, Jin Hong, Meiping Huang, Jian Zhuang, Shanshan Bi, Dan Pan, Najeeb Ullah, Kaleem Nawaz Khan, Tianchen Wang, Yiyu Shi, Xiaomeng Li, and Xiaowei Xu. ImageCAS: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *Computerized Medical Imaging and Graphics*, 109:102287, 2023. doi: 10.1016/j.compmedimag.2023.102287.

Lin Zhao, Xin Yu, Yikang Liu, Xiao Chen, Eric Z Chen, Terrence Chen, and Shanhui Sun. Leveraging diffusion model and image foundation model for improved correspondence matching in coronary angiography. *IEEE Transactions on Medical Imaging*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A Additional Method and Implementation Details

A.1 Pretraining Data

AngioFM pretraining uses 31,559 unlabeled coronary CAG videos from an in-house collection and synthetic multi-view projections generated from 1,000 CCTA-derived coronary trees (in-house data). Real cines provide temporal and contrast variation for masked prediction, contrast translation, and self-distillation. For synthetic data, known 3D coronary trees are projected from controlled C-arm viewpoints, providing camera geometry, vessel masks, centerlines, branch structure, and tangent fields without additional annotation of the rendered images.

A.2 Geometric Attention Distance

The geometric bias in Eq. 6 uses the distance from a finite query-ray segment in the patient region to a context ray, rather than unrestricted line-to-line distance. Let the retained query segment be

$$\Gamma_j = \{p_j + t d_j : t \in [t_j^-, t_j^+]\}, \tag{11}$$

where the interval is obtained by intersecting the query ray with the $\pm 100\,\mathrm{mm}$ isocenter slab and excludes the X-ray source. For unit context direction d_i, define

$$P_i = I - d_i d_i^\top, \qquad w = P_i(p_j - p_i), \qquad v = P_i d_j. \tag{12}$$

The distance is

$$D^2(r_j, r_i) = \min_{t \in [t_j^-, t_j^+]} \|w + tv\|_2^2, \tag{13}$$

with

$$\hat{t} = \mathrm{clip}\left(-\frac{v^\top w}{v^\top v}, t_j^-, t_j^+ \right) \tag{14}$$

when $v^\top v > 0$. For same-view rays sharing source S, exclusion of the source gives

$$D(r_j, r_i) = \ell_{\min} |\sin\theta|, \tag{15}$$

where θ is the angle between the rays and $\ell_{\min} > 0$ is the distance from the source to the retained query segment. Thus, neighboring same-view rays receive a smaller penalty, while across views the distance provides a depth-restricted epipolar prior.

A.3 Training Details

AngioFM uses a DINOv3 ViT-L/16 encoder and a randomly initialized 12-layer cross-attention predictor. We train on 512-pixel inputs. Masked-anchor prediction, contrast translation, synthetic cross-view prediction, and branch supervision are sampled with probabilities 0.125, 0.25, 0.1875, and 0.125, respectively; the remaining 0.3125 of updates use inherited DINO/iBOT self-distillation. We train for 100k optimizer steps with AdamW (Loshchilov & Hutter, 2017), weight decay 0.05, and $\beta = (0.9, 0.95)$. The predictor learning rate is 10^{-4} and the encoder uses a $0.1\times$ multiplier. The EMA momentum is 0.996.

A.4 Baseline Checkpoint Provenance

DINOv3-L denotes the generic image-pretrained checkpoint (Siméoni et al., 2025). DINOv3-L-FT denotes the same architecture after additional self-supervised training on the in-house CAG collection; this checkpoint also initializes AngioFM before its conditional and structural pretraining stage. DINO-world (Baldassarre et al., 2025) and LeVJEPA (Kuhn et al., 2026) are likewise pretrained on the in-house CAG videos before downstream evaluation. All downstream comparisons freeze these pretrained encoders and train only the task-specific readout.

DINO-world pretraining. DINO-world uses the same pool of 31,559 unlabeled CAG videos and the same DINOv3-L-FT encoder initialization as AngioFM; its video predictor is initialized randomly. It is trained for 100k optimizer steps on 512-pixel inputs, with the AdamW settings and learning-rate configuration in Appendix A.3. These shared conditions do not replace its predictive objective with AngioFM's acquisition-conditioned task mixture or CCTA-derived structural supervision.

B Additional Evaluation Protocols

B.1 Downstream Datasets

Downstream evaluation combines public benchmarks with in-house datasets covering lesion localization, coronary anatomy, cardiac motion, and interventional devices. All methods use the same task-specific training, validation, and test partitions. Images are grouped by patient, study, or sequence, and official or released public splits are followed where available.

For lesion localization, the in-house calcification dataset contains 393 training, 67 validation, and 121 test images. The stenosis subset of ARCADE (Popov et al., 2024) contains 1,000 training, 200 validation, and 300 test images. Coronary-dominance classification uses CoronaryDominance (Kruzhilov et al., 2025), with five-fold cross-validation on a pool of 922 training studies. Its 549 held-out test studies comprise 400 real-distribution and 149 domain-shift studies. Cardiac-phase matching uses an in-house collection of 2,849 training, 959 validation, and 132 test cines.

For dense prediction, the coronary-segment subset of ARCADE also contains 1,000 training, 200 validation, and 300 test images. CARDIAG (Lau et al., 2026) supports coronary-segment, vessel, and catheter segmentation, using 437 training, 60 validation, and 148 test images. Vessel segmentation is additionally evaluated on XCAD (Ma et al., 2021), with 80 training, 20 validation, and 26 test images, and XACV (Wu et al., 2025), with 155 training, 32 validation, and 34 test images. XACV is also used for catheter segmentation with the same partitions. Guidewire segmentation uses an in-house dataset containing 5,120 training, 640 validation, and 640 test frames.

For single-view centerline regression, ARCADE-SYNTAX contributes 972 training, 197 validation, and 300 test frames. A separate set of 141 CARDIAG frames is used exclusively for external testing. Multi-view centerline regression uses synthetic projections rendered from public ImageCAS CCTA volumes (Zeng et al., 2023). This dataset contains 1,775 training and 210 test case-sides, with no separate validation split; these counts refer to case-sides rather than individual projection images.

B.2 Statistical Analysis

Unless otherwise specified, comparisons with AngioFM use 20,000 paired bootstrap resamples and raw two-sided $p < 0.05$ without correction for multiple comparisons. Lesion and dense-prediction experiments resample images, whereas coronary-dominance and cardiac-phase experiments resample studies. These tests quantify sampling uncertainty conditional on the fitted models and do not include variability across independently retrained pretrained encoders.

B.3 Lesion Localization

Lesion localization uses non-interpolated AP. The in-house calcification test set contains 121 images and 132 lesions; a candidate is correct when it lies within 8 pixels of an annotated lesion line. ARCADE (Popov et al., 2024) contains 300 test images and 386 lesions; candidates inside a stenosis polygon or within an 8-pixel exterior tolerance are counted as matches. Candidate extraction uses a 27-pixel suppression radius and at most 20 predictions per image. Each lesion is matched at most once. Checkpoints are selected on validation data and no test-set confidence threshold is selected.

B.4 Baseline Inputs

All downstream tasks use 512×512 inputs. ARCADE contains only isolated images; AngioFM and DINO-world therefore receive the target with black context frames, while single-image encoders receive the target alone. For multi-frame in-house calcification, context frames are spaced 1 s apart; because most annotated targets occur near the beginning of a cine, later frames substitute when earlier context is unavailable.

B.5 Coronary Dominance

Coronary dominance is evaluated using a frozen encoder and linear classifier on the CoronaryDominance dataset (Kruzhilov et al., 2025). Spatial features, and temporal features for video encoders, are pooled without learned attention. Frame probabilities are averaged within each view and view

probabilities within each study. Five fold models are trained after a stratified 100-study hold-out, and study-level accuracy, ROC-AUC, and MCC are reported on the released real-distribution and domain-shift test sets.

Split audit. The released data contain one training study that shares acquisitions with a domain-shift test study and several duplicate test studies, including duplicates with conflicting labels. We retain the released split for comparability with prior work. Removing the identified train/test-overlap study changes accuracy by at most 1.21 percentage points in our check, but does not remove duplicate records within the test data. The domain-shift results should therefore be interpreted as performance on the released benchmark rather than independently verified external generalization.

B.6 Cardiac-Phase Matching

Phase matching uses 16 consecutive frames. Frozen spatial features are mean-pooled and a linear projection produces a 1024-dimensional L_2-normalized embedding. Training uses the cosine triplet objective adapted from Liu et al. (2024). At evaluation, each anchor is represented by five consecutive embeddings and candidate phases are ranked by the mean cosine similarity of the five aligned pairs.

Evaluation contains 132 acquired cines from 47 studies and 6,396 valid anchors. Error is reported in frames and as a percentage of the cardiac cycle. For context, Liu et al. (2024) report 0.89 frames (7.01% of a cycle) for CNN-ConvLSTM and 0.85 frames (6.72%) for CNN-Transformer. These fully trained roadmapping models use a different test list and live-fluoroscopy-to-recorded-angiogram matching, so their scores are not paired comparisons with our results.

B.7 CNN Decoder

The CNN-256 decoder projects frozen encoder features to 256 channels, combines them with a trainable image-convolution stem, and uses multiscale skip-connected upsampling to predict 512×512 masks. Final and auxiliary outputs use binary cross-entropy and Dice losses. The same decoder body and task splits are used for all frozen encoders; only feature adapters differ with encoder dimensionality.

Binary tasks report Dice after pooling true-positive, predicted, and ground-truth pixels over the full test set. Coronary-segment parsing reports the unweighted mean of class-wise pooled Dice over segment classes present in the test set (24 classes for ARCADE and 22 for CARDIAG). Binary tasks select the best validation Dice within 100 epochs and coronary-segment parsing within 200 epochs. Significance testing follows Appendix B.2.

B.8 Centerline Decoder

Centerline regression predicts anatomically labelled coronary-root-to-endpoint routes with 128 learned queries. Each route is parameterized by $K = 48$ Catmull–Rom spline control points and sampled at $N = 192$ ordered positions. Hungarian matching uses vessel-class and ordered-point costs. The decoder contains six transformer layers followed by four local-feature refinement blocks. The encoder remains frozen for all comparisons. Evaluation uses symmetric Chamfer distance between predicted and annotated routes; this metric is distinct from the ordered-point regression loss used for training.

C Supplementary Results and Expert Models

C.1 Complete Coronary-Dominance Results

The released split limitations are described in Appendix B.5. For literature context, Kruzhilov et al. (2025) report 97.3%/0.878 accuracy/MCC on the real-distribution set and 89.4%/0.765 under domain shift using a fully trained task-specific pipeline. These values are not paired comparisons with the frozen-feature readouts.

Table 4: **Coronary dominance on the released splits.** Study-level accuracy/AUC (%) and MCC are averaged over five fold models. ↓ denotes significantly lower performance than AngioFM within the same input setting under study-level bootstrap ($p < 0.05$).

	Real distribution			Domain shift		
Method	Acc.	AUC	MCC	Acc.	AUC	MCC
Single frame						
AngioFM	90.40	91.10	0.583	70.20	65.80	0.322
DINOv3-L	89.65	86.90$^{\downarrow}$	0.449$^{\downarrow}$	70.20	59.16$^{\downarrow}$	0.295
DINO-world	90.40	90.26	0.547	68.86	60.46$^{\downarrow}$	0.269
LeVJEPA	89.30	83.90$^{\downarrow}$	0.433$^{\downarrow}$	61.21$^{\downarrow}$	57.05	0.136$^{\downarrow}$
Video						
AngioFM	91.05	91.85	0.622	70.60	67.40	0.328
DINO-world	89.95	89.87$^{\downarrow}$	0.529$^{\downarrow}$	69.40	60.91$^{\downarrow}$	0.280
LeVJEPA	88.75	84.54$^{\downarrow}$	0.413$^{\downarrow}$	67.65	55.44$^{\downarrow}$	0.217

C.2 Segmentation Expert Models

We additionally train task-specific nnU-Net experts on the same training, validation, and test splits as the CNN-256 readouts. All use the default 2D configuration and are trained for 200 epochs; coronary segments use region-based training to support overlapping labels. The same pooled Dice definition is used for nnU-Net and AngioFM. nnU-Net is single-frame; its predictions are also compared with the video AngioFM decoder on the same annotated target frames.

Table 5: **Task-specific nnU-Net experts versus frozen AngioFM features.** Test Dice (%). ↑ / ↓ marks nnU-Net scores significantly above/below AngioFM under 20,000 paired image-bootstrap resamples with raw two-sided $p < 0.05$.

		Single frame		Video	
Dataset	Target	AngioFM	nnU-Net	AngioFM	nnU-Net
ARCADE	Coronary segments	40.52	42.84	—	—
CARDIAG	Coronary segments	43.30	41.59$^{\downarrow}$	44.01	41.59$^{\downarrow}$
CARDIAG	Vessel	88.11	89.05$^{\uparrow}$	88.09	89.05$^{\uparrow}$
XCAD	Vessel	82.83	83.55	—	—
XACV	Vessel	86.92	87.16	87.04	87.16
CARDIAG	Catheter	84.57	83.20	85.04	83.20$^{\downarrow}$
XACV	Catheter	87.29	78.82$^{\downarrow}$	87.59	78.82$^{\downarrow}$
In-house wire	Guidewire	60.43	60.07	60.67	60.07$^{\downarrow}$

For additional literature context, anatomy-specific YOLOv8 reports 52% on ARCADE coronary segments (Tran et al., 2025); ConvNeXt V2 with DeepLabV3+ reports 45.6% on CARDIAG segments and approximately 79% on catheter segmentation (Lau et al., 2026); SPIRONet reports 81.76% on XCAD vessel segmentation (Huang et al., 2024); and TransUNet and VasoMIM report 82.49% and 86.09%, respectively, on XACV (Huang et al., 2026). These results use different splits and/or metric aggregation and are included only as literature references.

C.3 Task-Specific Lesion Detectors

YOLO26-M (Sapkota et al., 2025) and RF-DETR-M (Robinson et al., 2026) are trained as task-specific detectors and evaluated on our test splits using box centers as point candidates. Their supervision and output geometry differ from the frozen-feature lesion probes.

Table 6: Single-frame lesion AP (%) under the same evaluation rule as the linear probes.

Method	In-house calcification	ARCADE stenosis
AngioFM	30.91	76.38
YOLO26-M	12.40	62.59
RF-DETR-M	8.79	64.01

C.4 Centerline Regression Results

Table 7: **Centerline regression with frozen encoder features.** Single-view Chamfer distance is evaluated on in-domain ARCADE-SYNTAX (300 frames, 984 routes) and out-of-domain CARDIAG (141 frames, 948 routes). Multi-view evaluation uses 210 held-out CCTA-derived case-sides with five identical seeded view draws for each encoder. Lower Chamfer is better.

Evaluation	DINOv3-L	DINO-world	AngioFM
ARCADE-SYNTAX, single view	4.36	3.94	**3.25**
CARDIAG, single view	9.72	8.07	**7.27**
Multi-view	—	5.39	**3.87**

The multi-view Chamfer point estimates are 5.39 pixels for DINO-world and 3.87 pixels for AngioFM.

D Initialization and Targeted Ablations

D.1 Initialization Control

DINOv3-L-FT is the CAG-adapted checkpoint used to initialize AngioFM before conditional and structural pretraining.

Table 8: **Selected initialization comparisons.** AP and Dice are percentages. $\downarrow$ denotes significantly lower performance than AngioFM under paired bootstrap ($p < 0.05$).

Task	DINOv3-L-FT	AngioFM
ARCADE segments (Dice)	$37.00^{\downarrow}$	40.52
CARDIAG segments (Dice)	$41.29^{\downarrow}$	43.30
In-house calcification (AP)	28.95	30.91
ARCADE stenosis (AP)	75.55	76.38

The segment-level differences are significant, whereas the lesion-localization differences are not.

D.2 Branch Supervision

The branch-supervision ablation uses the same 100k-step training schedule as full AngioFM but removes the branch contrastive loss while retaining the remaining objectives. Frozen representations are evaluated with the same CNN-256 decoder. Removing branch supervision lowers coronary-segment Dice from 40.52 to 37.86 on ARCADE, from 43.30 to 41.47 on single-frame CARDIAG, and from 44.01 to 42.31 on video CARDIAG; all three differences are significant under paired image bootstrap ($p < 0.05$). The ablation isolates the branch loss rather than all CCTA-derived supervision, since cross-view and tangent objectives remain active.

D.3 Input Context

This experiment uses the same frozen AngioFM checkpoint and separately fitted single-frame and video readouts. It therefore measures the utility of additional observations at inference rather than removing a temporal pretraining objective.

Table 9: **Input-context comparison for AngioFM.** Scores are percentages except MCC. Δ denotes video minus single-frame performance. Differences are descriptive and are not assigned within-model significance tests.

Task (metric)	Single frame	Video	Δ
In-house calcification (AP)	30.91	31.63	+0.72
CARDIAG segments (Dice)	43.30	44.01	+0.71
CARDIAG vessel (Dice)	88.11	88.09	-0.02
XACV vessel (Dice)	86.92	87.04	+0.12
CARDIAG catheter (Dice)	84.57	85.04	+0.47
XACV catheter (Dice)	87.29	87.59	+0.30
In-house guidewire (Dice)	60.43	60.67	+0.24
Dominance, real (accuracy)	90.40	91.05	+0.65
Dominance, real (AUC)	91.10	91.85	+0.75
Dominance, real (MCC)	0.583	0.622	+0.039
Dominance, shift (accuracy)	70.20	70.60	+0.40
Dominance, shift (AUC)	65.80	67.40	+1.60
Dominance, shift (MCC)	0.322	0.328	+0.006

For comparison, DINO-world calcification AP changes from 27.38 to 30.55 and its CARDIAG segment Dice from 41.12 to 41.91. Thus, higher absolute video performance does not imply that AngioFM benefits more from additional frames. Context can provide changes in contrast visibility as well as motion cues, so these comparisons do not isolate temporal reasoning.

D.4 Physical Conditioning

We evaluate physical ray and acquisition-state conditioning against a positional-only variant that retains distinct time and spatial position embeddings but removes physical ray coordinates, contrast/acquisition-state conditioning, and the geometric attention bias. Multi-view evaluation uses identical held-out case-sides and view draws for both encoders.

Table 10: **Physical-conditioning comparison for multi-view centerline regression.** Chamfer distance is averaged over annotated route–view pairs. Only the Chamfer difference is assigned an inferential test.

Variant	Chamfer (px) ↓	Identity (%)	Recall@10 px (%)
Full AngioFM	3.87	96.07	89.04
Positional-only conditioning	5.43	96.27	85.99

The positional-only variant has a higher mean Chamfer distance (5.43 versus 3.87 pixels; paired difference $+1.56$ pixels, 95% CI $[1.36, 1.77]$, $p < 10^{-4}$). Branch identity is 96.27% versus 96.07%, and Recall@10 is 85.99% versus 89.04%; no significance tests are reported for these two metrics.