Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Vision-Language Foundation Model for Coronary Angiography Supports Heart-Team Treatment Selection

Tae Joon Jun

taejoon@amc.seoul.kr

Asan Medical Center, University of Ulsan College of Medicine https://orcid.org/0000-0002-6808-5149

Byeolhee Kim

Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine

Hanbit Park

Department of Cardiology, Gangneung Asan Hospital, University of Ulsan College of Medicine

Ju Hyeon Kim

Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine

Seong-Bong Wee

Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine

Do-Yoon Kang

Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine

Duk-Woo Park

Asan Medical Center

Seung-Jung Park

Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine

Young-Hak Kim

Asan Medical Center, University of Ulsan College of Medicine https://orcid.org/0000-0002-3610-486X

Jung-Min Ahn

Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine

Article

Keywords: coronary angiography, foundation model, vision–language model, heart-team, revascularisation, external validation

Posted Date: June 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9934333/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

A Vision-Language Foundation Model for Coronary Angiography Supports Heart-Team Treatment Selection

Byeolhee Kim[1], Hanbit Park[2], Ju Hyeon Kim[3], Seong-Bong Wee[3], Do-Yoon Kang[3], Duk-Woo Park[3], Seung-Jung Park[3], Young-Hak Kim[3], Jung-Min Ahn[3*], Tae Joon Jun[1*]

[1]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea.
[2]Department of Cardiology, Gangneung Asan Hospital, University of Ulsan College of Medicine, Gangneung, South Korea.
[3]Department of Cardiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea.

*Corresponding author(s). E-mail(s): d090693@amc.seoul.kr; taejoon@amc.seoul.kr;
Contributing authors: kbh0216@amc.seoul.kr; d210588@gnah.co.kr; jhnkim86@amc.seoul.kr; d240727@amc.seoul.kr; dykang@amc.seoul.kr; dwpark@amc.seoul.kr; sjpark@amc.seoul.kr; mdyhkim@amc.seoul.kr;

Abstract

Coronary artery disease is a leading cause of death worldwide, and the choice between surgical and percutaneous revascularisation shapes patient outcomes. This Heart-Team decision is anchored by the SYNTAX angiographic complexity score, which requires manual 26-segment annotation of the coronary tree and carries substantial inter-observer variability. Here we introduce FMCA, the first vision–language foundation model for coronary angiography at clinically representative scale, trained by cine-level contrastive learning on 113,233 cines from 12,344 patients without segment-level expert annotation. Korean interpretation reports were paraphrased into vessel-structured English for the contrastive objective only, and all downstream predictions take cine video as the sole input. A single frozen encoder recognises coronary anatomy zero-shot across two countries, recovers disease extent and SYNTAX complexity from video on a held-out cohort, and supports the CABG-versus-non-CABG decision at AUROC 0.893,

exceeding the expert clinical-plus-SYNTAX standard by 0.138 absolute AUROC. The advantage is preserved at an operationally independent second institution, and lightweight site-specific tuning further improves external performance. FMCA provides reproducible, video-only decision support across the coronary angiographic pipeline.

Keywords: coronary angiography, foundation model, vision–language model, heart-team, revascularisation, external validation

Introduction

Coronary artery disease is the leading cause of death worldwide[1], and the choice between surgical and percutaneous revascularisation shapes long-term outcomes for affected patients[2]. Invasive coronary angiography (CAG) is the diagnostic reference standard for coronary artery disease. It is also the principal source of evidence for Heart-Team treatment selection among percutaneous coronary intervention (PCI), coronary artery bypass grafting (CABG), and optimal medical therapy[3, 4]. The SYNTAX score[5] is the most widely adopted CAG-derived angiographic complexity index and anchors both European[3] and North American[4] revascularisation guidelines. A high-tertile SYNTAX score is one of the principal indications for surgical revascularisation[2].

SYNTAX scoring is, however, laborious. The calculation requires manual annotation of every coronary segment under a 26-region scheme, with each segment weighted by lesion length, calcification, bifurcation involvement, and other morphological features[5]. Inter-observer variability is substantial even among experienced operators[6], computation is slow, and adoption varies across centres. These constraints limit objective, scalable, and reproducible decision support, even though the underlying CAG video already contains the information that the SYNTAX score summarises.

Foundation models trained on large medical image–text corpora are reshaping clinical AI across multiple domains. Examples include generalist biomedical multimodal models[7], whole-slide pathology models that support cancer subtyping and rare-disease retrieval[8], and retinal models that generalise across ophthalmic diagnoses[9]. For invasive coronary angiography, no equivalent foundation model has been reported at a clinically representative scale. Prior CAG-focused work has targeted narrow tasks such as segment-level lesion detection[10], coronary-dominance classification[11], or treatment-decision support from a single cohort[12], typically using a few thousand images and bespoke architectures.

Building such a model for coronary angiography faces three obstacles that the pathology and retinal domains did not, each of which blocks the standard image–report contrastive pretraining used in those domains. First, interpretation reports are typically written in the native institutional language, often non-English, and are poorly aligned with the predominantly English training corpora of available vision–language backbones. Second, CAG reports follow hospital-specific structured templates that

resist standard contrastive image–report pretraining, which assumes free-form paired captions. Third, each CAG study consists of multiple cine acquisitions and each cine captures only a subset of the coronary tree, requiring cine-level rather than study-level contrastive supervision.

Here we introduce FMCA, the first vision–language foundation model for coronary angiography at clinically representative scale (Figure 1). A single FMCA encoder, trained once on 113,233 routine cine angiograms from 12,344 patients at the Asan Medical Center without any segment-level expert annotation, produces a general-purpose representation that transfers across the entire angiographic decision pipeline. Pretraining aligns each cine with a vessel-organised English paraphrase of its Korean interpretation report, generated by an open-weight large language model[13], through cine-level contrastive learning. All downstream predictions take cine video as the sole input at inference, with the encoder kept frozen. We test FMCA on the three steps of the angiographic clinical decision pipeline. First, we evaluate zero-shot recognition of coronary anatomy on two public external datasets unseen during pretraining[10, 11]. Second, we evaluate recovery of disease extent and SYNTAX complexity from cine video on an internal held-out cohort withheld from pretraining. Third, we evaluate Heart-Team CABG-versus-non-CABG selection from cine video, with external validation at an operationally independent Korean tertiary institution. Finally, we ask whether the foundation-model embedding supports site-tailored deployment from a shared backbone, an architectural advantage that fixed scoring procedures such as SYNTAX cannot provide. Across these tasks, video-only FMCA recognises coronary anatomy zero-shot across two countries, recovers SYNTAX complexity from cine video, and supports the Heart-Team CABG decision at an accuracy exceeding the expert clinical-plus-SYNTAX standard, with the advantage preserved at the external institution.

Results

Study design and cohorts

We evaluated FMCA along the three sequential steps of the angiographic clinical decision pipeline, from coronary anatomy recognition, through disease-extent and complexity stratification, to Heart-Team revascularisation-strategy selection (Figure 1). The foundation model was pretrained by cine-level contrastive learning on the Asan Medical Center (AMC) cohort of 113,233 cines from 12,344 patients. It was then evaluated on a patient-level AMC test set withheld from pretraining, on two public external datasets from two countries for zero-shot anatomy recognition, and at an operationally independent external Korean institution, Gangneung Asan Hospital (GAH), for revascularisation-strategy selection. The flow of each cohort into pretraining, classifier fitting, and downstream evaluation is shown in Figure 2.

Baseline characteristics of the AMC cohort are summarised in Table 1. The patient-level split into training, validation, and test partitions was balanced across every demographic and clinical variable, with all Benjamini–Hochberg-adjusted p-values ≥ 0.121 (full per-split comparison in Extended Data Table 1).

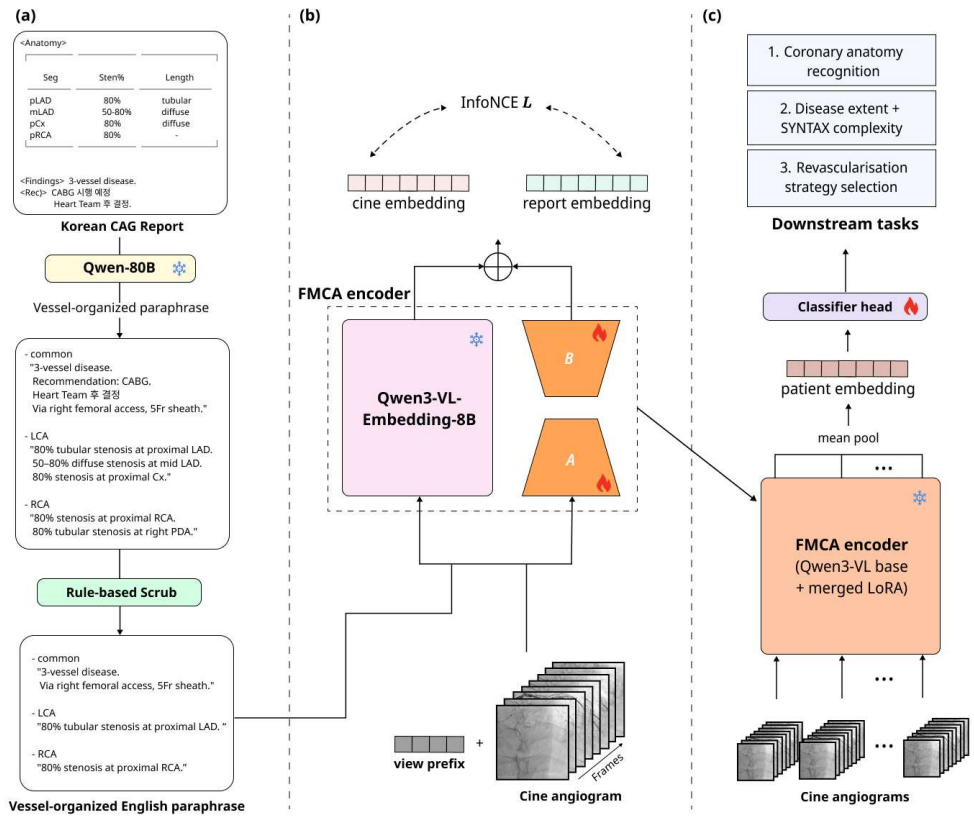


Fig. 1 Overview of FMCA. *(a)* Korean interpretation reports are paraphrased into a vessel-organised English representation with `common`, `LCA`, and `RCA` fields by an open-weight large language model, with treatment-recommendation sentences scrubbed before pretraining. *(b)* Cine-level contrastive pretraining (InfoNCE) aligns cine-video embeddings with paraphrase embeddings, with low-rank adapter (LoRA) parameters trained on a frozen Qwen3-VL-Embedding-8B backbone. Each cine is prefixed with its acquisition-angle token. *(c)* Downstream evaluation takes cine video as the sole input, and no text is consumed at inference. The frozen FMCA encoder maps each cine to a 4,096-dimensional embedding, and a task-specific linear classifier head produces the predicted label. For the disease-extent and treatment tasks, per-patient embeddings are obtained by mean-pooling across a patient's cines, whereas the anatomy task is evaluated at the cine or study level. The same frozen encoder supports the three downstream tasks: coronary anatomy recognition (Table 2), disease extent and SYNTAX complexity (Table 3), and Heart-Team revascularisation strategy selection (Table 4).

FMCA recognises coronary anatomy zero-shot across countries

We first evaluated the entry step of the angiographic decision pipeline, recognition of coronary anatomy, as a zero-shot test of whether the FMCA representation transfers beyond its training institution. A linear vessel-identification classifier (L2-regularised logistic regression) was fitted on 745 manually labelled cines from 199 patients in the AMC cohort and then applied, without any fitting on the target data,

Table 1 Baseline characteristics of the Asan Medical Center cohort.

Variable	Overall (n = 12,343)
Age (years)	64.0 [57.0–70.0]
Male sex, n (%)	9,323 (75.5%)
Body-mass index (kg/m^2)	24.8 [23.0–26.8]
LV ejection fraction (%)	60.0 [56.0–64.0]
Serum creatinine (mg/dL)	0.9 [0.8–1.1]
SYNTAX score	18.0 [12.0–26.5]
Hypertension, n (%)	7,998 (65.4%)
Diabetes mellitus, n (%)	4,362 (35.6%)
Hyperlipidemia, n (%)	7,473 (61.3%)
Current smoker, n (%)	2,986 (24.4%)
Atrial fibrillation, n (%)	262 (2.1%)
COPD, n (%)	218 (1.8%)
Chronic kidney disease, n (%)	436 (3.6%)
Prior myocardial infarction, n (%)	721 (5.8%)
Prior stroke, n (%)	858 (7.0%)
Prior PCI, n (%)	1,845 (15.0%)
Prior CABG, n (%)	33 (0.3%)
Left main disease, n (%)	2,219 (18.0%)
LAD disease, n (%)	10,721 (86.9%)
LCX disease, n (%)	7,556 (61.2%)
RCA disease, n (%)	7,724 (62.6%)
Clinical presentation, n (%)	
Stable angina	6,089 (49.3%)
Unstable angina	2,960 (24.0%)
NSTEMI	639 (5.2%)
STEMI	242 (2.0%)
Treatment received, n (%)	
PCI	8,872 (71.9%)
CABG	2,647 (21.4%)
Optimal medical therapy	667 (5.4%)

Numeric variables are reported as median [interquartile range]. Categorical variables are reported as n (%). Variables with missing data are: age (n = 425 missing), body-mass index (n = 6), left-ventricular (LV) ejection fraction (n = 2,708), serum creatinine (n = 1,144), and institutional SYNTAX score (n = 7,729). All other variables are complete. The full per-split distribution is reported in Extended Data Table 1, and variable-level missingness in Supplementary Section S8.1. Treatment received refers to the definitive revascularisation strategy selected during the index hospitalisation. COPD, chronic obstructive pulmonary disease. NSTEMI, non-ST-elevation myocardial infarction. STEMI, ST-elevation myocardial infarction.

to two independent public datasets unseen during pretraining. The CoronaryDominance dataset[11] contributed 10,560 cines from 1,577 patient studies acquired on two Philips imaging systems (Russia). The ARCADE dataset[10] contributed 1,500 single-frame angiograms from the Medical Image Computing and Computer Assisted Intervention 2023 Challenge (Kazakhstan). Zero-shot discrimination between left coronary artery (LCA) and right coronary artery (RCA) reached AUROC 0.970 (95% CI 0.967 to 0.973) on CoronaryDominance and AUROC 0.995 (95% CI 0.991 to 0.999) on ARCADE (Table 2).

The CoronaryDominance release includes three vendor-specific sub-splits across two Philips imaging systems, Allura Clarity and Azurion. Per-split AUROCs ranged from 0.947 to 0.981 (Supplementary Section S2.3), indicating that anatomy recognition is preserved across scanner generations within the same vendor family. The ARCADE single-frame result is informative for a different reason. AMC pretraining used only

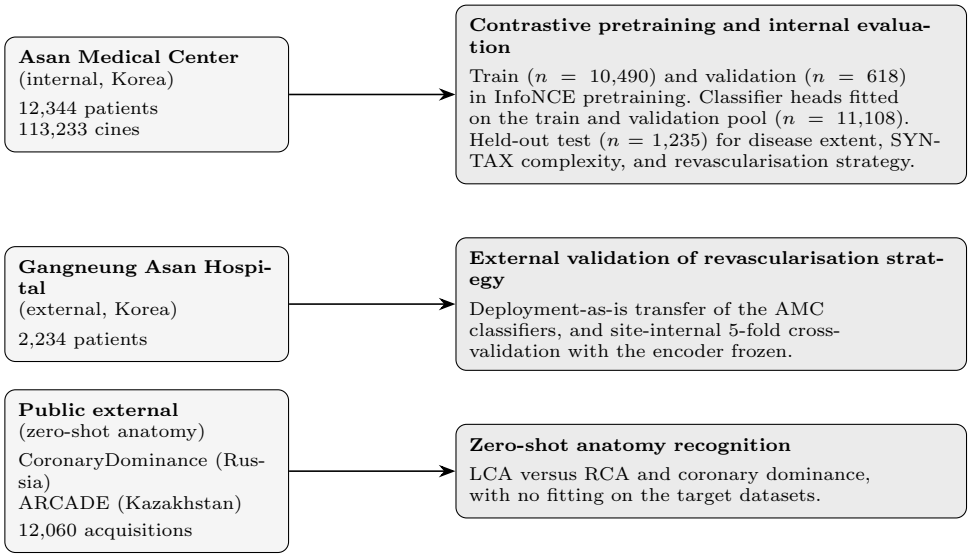


Fig. 2 Data flow across cohorts and downstream evaluation roles. Each cohort (left) maps to its role in pretraining and downstream evaluation (right). Patient-level splits apply only to the AMC internal cohort. The external cohorts have no train, validation, or test partition. For coronary anatomy recognition, the vessel-identification classifier was fitted on 745 cines from 199 AMC patients, a cine-level annotation orthogonal to the patient-level split, and then applied without further fitting to the public external datasets. Cohort details are described in Methods (Study cohorts) and downstream protocols in Methods (Downstream evaluation).

multi-frame cines, yet single-frame AUROC reached 0.995. This suggests that FMCA encodes vessel anatomy largely as a per-frame feature.

We next investigated whether FMCA captures anatomic structure that was not directly supervised by paired text during pretraining. Coronary dominance, defined by the artery supplying the posterior descending artery, is a standard anatomic descriptor that was not encoded as a discrete label in the AMC interpretation reports. We fitted a logistic-regression probe on the FMCA embeddings of the CoronaryDominance studies and evaluated it by 5-fold patient-grouped cross-validation ($n = 1{,}577$). The probe achieved AUROC 0.924 ($\pm$0.016 across folds) and accuracy 0.870. A linear probe on the frozen FMCA embeddings therefore approaches the fully supervised ConvNeXt baseline (89.4% accuracy) reported on the same dataset's domain-shift split[11], despite coronary dominance never appearing as a label in the AMC interpretation reports. This indicates that the FMCA representation captures anatomic structure beyond the descriptors that AMC paired text explicitly supervised.

Taken together, FMCA recognises coronary vessel anatomy from cine video alone and generalises zero-shot across two countries and across both multi-frame cine and single-frame acquisitions. The representation additionally encodes an anatomic concept, coronary dominance, that was not part of the paired-text supervision at AMC.

Table 2 Coronary anatomy recognition on public external datasets (video-only inference).

Dataset (country)	n	Setting	AUROC [95% CI]
CoronaryDominance (Russia)	10,560 cines	Zero-shot transfer	0.970 [0.967, 0.973]
ARCADE (Kazakhstan)	1,500 frames	Zero-shot transfer	0.995 [0.991, 0.999]

A linear L2-regularised logistic-regression vessel-identification classifier was fitted on 745 manually labelled AMC cines from 199 patients and applied without further fitting to the two external datasets. Confidence intervals are 1,000-iteration bootstrap, resampled at the cine level for CoronaryDominance and at the frame level for ARCADE.

FMCA recovers disease extent and SYNTAX complexity from video

We next asked whether FMCA recovers, from cine video alone, the angiographic disease descriptors that SYNTAX scoring computes by manual annotation. On the AMC held-out test cohort ($n = 1{,}235$ patients withheld from contrastive pretraining and all classifier training), per-patient FMCA embeddings were passed to task-specific linear or shallow classifiers. Hyperparameters were selected on the validation split, and each classifier was evaluated once on the held-out test. Multi-vessel disease (MVD), defined as significant disease in at least two of the left anterior descending (LAD), left circumflex (LCx), and right coronary (RCA) arteries, reached AUROC 0.810 (95% CI 0.782 to 0.836). Three-vessel disease (3VD) reached AUROC 0.804 (95% CI 0.777 to 0.828). Both predictions used cine video as the sole input (Table 3).

To quantify the contribution of FMCA's contrastive pretraining, we evaluated the same classifier pipelines on embeddings from a Base configuration: the frozen open-weight backbone used by FMCA, without low-rank adaptation (LoRA)[14] or contrastive training. The Base configuration measures what the off-the-shelf multimodal backbone alone recovers from coronary cine images, before any angiography-specific pretraining. Base AUROC was 0.594 for multi-vessel disease and 0.623 for three-vessel disease, so FMCA contrastive pretraining added between +0.181 and +0.216 absolute AUROC on these tasks. This gain indicates that performance is driven by the angiography-specific representation, not by the off-the-shelf backbone.

Beyond binary vessel-disease detection, we examined whether FMCA recovers the continuous SYNTAX complexity score from cine video alone. SYNTAX is a clinical-standard angiographic complexity index[5], computed by manual annotation of 26 coronary segments, with each segment weighted by lesion length, calcification, and bifurcation involvement. On the 465 held-out patients with a documented institutional SYNTAX score, regression from FMCA embeddings achieved R^2 0.330 (95% CI 0.210 to 0.403) and RMSE 8.97 on the SYNTAX range 0 to 78. Thresholded into tertiles (Low ≤ 22, Intermediate 23 to 32, and High ≥ 33) following the SYNTAX trial[2], the regression achieved a 3-class macro-AUROC of 0.748 (95% CI 0.712 to 0.783). The clinically pivotal High-SYNTAX class (≥ 33) reached a regression-derived AUROC of 0.842 (Supplementary Section S1.1).

These results show that FMCA recovers the angiographic descriptors that SYNTAX scoring manually computes from cine video. The strongest signal appeared on

multi-vessel disease and on the clinically pivotal High-SYNTAX class. The representation therefore captures both basic coronary anatomy and the disease-extent descriptors that anchor SYNTAX scoring.

Table 3 Disease extent and SYNTAX complexity from cine video alone (AMC held-out test).

Task	n_{test}	Metric	Base [95% CI]	FMCA [95% CI]
Multi-vessel disease (MVD)	1,235	AUROC	0.594 [0.556, 0.627]	**0.810** [0.782, 0.836]
Three-vessel disease (3VD)	1,235	AUROC	0.623 [0.588, 0.655]	**0.804** [0.777, 0.828]
SYNTAX, continuous	465*	R^2	—	**0.330** [0.210, 0.403]
SYNTAX, 3-class tertile	465*	macro-AUROC	—	**0.748** [0.712, 0.783]

Held-out single-shot evaluation on the AMC test cohort, withheld from contrastive pretraining and all classifier training. Base reports the result on the frozen Qwen3-VL-Embedding-8B backbone[15] without LoRA or contrastive pretraining, quantifying the contribution of FMCA's contrastive learning. The Base contrast is prespecified for binary AUROC tasks. For SYNTAX tasks, the clinically meaningful comparator is the institutional SYNTAX score itself, analysed in the SYNTAX-known-subset sensitivity (Supplementary Section S6.4). Best classifiers were L1-regularised logistic regression ($C = 0.1$) for the vessel-disease tasks and a multi-layer perceptron (128 hidden units, $\alpha = 0.01$) for the SYNTAX regression. Confidence intervals are 1,000-iteration patient-level bootstrap. Per-class AUROCs and per-vessel disease-detection results are in Supplementary Section S1.1. *SYNTAX tasks are restricted to held-out patients with a documented institutional SYNTAX score.

Video-only FMCA exceeds expert SYNTAX for the Heart-Team decision

Building on these capabilities, we tested whether FMCA supports selection between coronary artery bypass grafting (CABG) and non-CABG revascularisation strategies, among the most consequential clinical decisions that coronary angiography is performed to inform[3]. On the AMC held-out test cohort (n = 1,220 patients with a definitive treatment label, withheld from contrastive pretraining and all classifier training), we compared five feature combinations. These were M1 (16 clinical variables), M2 (clinical plus institutional SYNTAX score), M3 (FMCA video embedding only), M4 (clinical plus FMCA), and M5 (clinical plus SYNTAX plus FMCA). Models M3, M4, and M5 took cine video as the sole imaging input at inference. Video-only M3 reached AUROC 0.893 (95% CI 0.872 to 0.912), exceeding the expert clinical-plus-SYNTAX comparator M2 by Δ AUROC +0.138 (95% CI +0.100 to +0.175, $p < 0.001$). The full M5 stack reached AUROC 0.907 (95% CI 0.887 to 0.925), an absolute improvement of +0.294 over clinical variables alone (Table 4). A pre-specified sensitivity analysis on the 460 held-out patients carrying both a definitive treatment label and a documented institutional SYNTAX score (five of the 465 SYNTAX-documented test patients lacked a definitive treatment label) confirmed M3 exceeded M2 even when the comparator was at its strongest, with M2 AUROC 0.845, M3 AUROC 0.913, and Δ +0.068 (95% CI +0.026 to +0.116). Calibration and reclassification metrics are in Supplementary Section S6.2.

To test whether the internal advantage generalises across institutions, we applied the AMC-trained M1–M5 classifiers without further fitting at Gangneung Asan Hospital (GAH), an operationally independent Korean tertiary institution. The GAH

cohort comprised 2,234 consecutive patients with a definitive revascularisation label, of whom 91 received CABG and 2,143 received percutaneous coronary intervention (PCI). The cohort shared no clinicians, scanner fleet, or patient overlap with AMC. The GAH electronic export lacked institutional SYNTAX and parts of the 16-variable clinical panel. Median imputation from the AMC train pool therefore made M1 and M2 numerically identical at GAH, without affecting the central M3 evaluation. Under this deployment-as-is protocol, no GAH outcome label was observed by any classifier during fitting. Video-only M3 reached AUROC 0.744 (95% CI 0.694 to 0.793), exceeding the comparator M2 by Δ AUROC +0.201 (95% CI +0.121 to +0.281, $p < 0.001$). The Δ AUROC for M5 versus M2 was +0.189 (95% CI +0.109 to +0.265). Both deltas exceed the pre-specified +0.05 threshold with confidence intervals strictly above zero, satisfying STRONG support under the external-validation rule fixed before analysis (Supplementary Section S10). On GAH transfer, M3, M4, and M5 were statistically indistinguishable (Δ M3 minus M5 = +0.012, 95% CI −0.005 to +0.029, $p = 0.14$). Video-only M3 is therefore sufficient on its own for the GAH external task.

Across both the internal held-out cohort and the external institution, video-only FMCA supports CABG-versus-non-CABG selection at performance exceeding expert clinical-plus-SYNTAX scoring (Table 4, Figure 3).

Table 4 Revascularisation strategy selection: video-only FMCA versus expert clinical-plus-SYNTAX, internal and external.

Model	AMC held-out [95% CI]	GAH transfer [95% CI]	GAH internal CV
M1 — Clinical only	0.613 [0.571, 0.657]	0.545 [0.485, 0.608]*	0.625 ± 0.061
M2 — Clinical + SYNTAX	0.755 [0.718, 0.789]	0.545 [0.480, 0.604]*	0.625 ± 0.061
M3 — FMCA only	**0.893 [0.872, 0.912]**	**0.744 [0.694, 0.793]**	**0.810 ± 0.048**
M4 — Clinical + FMCA	0.898 [0.878, 0.917]	0.735 [0.683, 0.785]	0.815 ± 0.038
M5 — Full stack	0.907 [0.887, 0.925]	0.732 [0.680, 0.783]	0.815 ± 0.038
Δ M3 − M2	**+0.138 [+0.100, +0.175]**	**+0.201 [+0.121, +0.281]**	+0.185
Δ M5 − M1	+0.294 [+0.247, +0.339]	+0.188 [+0.107, +0.268]	+0.190

AMC held-out: 1,220 patients (265 CABG), single-shot evaluation of classifiers fit on the AMC train and validation patients with a definitive treatment label ($n = 10{,}966$). GAH transfer: 2,234 patients (91 CABG), same AMC-trained classifiers applied without target-domain refit (deployment-as-is). GAH site-internal CV: independent 5-fold stratified cross-validation within GAH, with the FMCA encoder kept frozen and only the classifier head refitted per fold. All M3, M4, and M5 results take cine video as the sole imaging input at inference. M5 = full stack (Clinical + SYNTAX + FMCA). *The GAH electronic export lacked institutional SYNTAX and parts of the M1 clinical-16 panel. Deployment-as-is median imputation from the AMC train pool therefore made M1 and M2 numerically identical at GAH. Confidence intervals are 1,000-iteration patient-level bootstrap on AMC and GAH transfer. The GAH site-internal CV column reports mean ± standard deviation across folds. P-values for the Δ rows are $p < 0.001$ for all four comparisons (1,000-iteration bootstrap resolution floor). These are raw two-sided bootstrap p-values. The full panel of paired comparisons, all of which retain significance under Benjamini–Hochberg control of the false-discovery rate, is reported in Supplementary Section S4.2. Best classifier: L2-regularised logistic regression with $C = 0.01$. Calibration and decision-curve analyses are in Supplementary Section S6.2, additional treatment-decision tasks in Section S4, and external-validation analyses in Section S12.

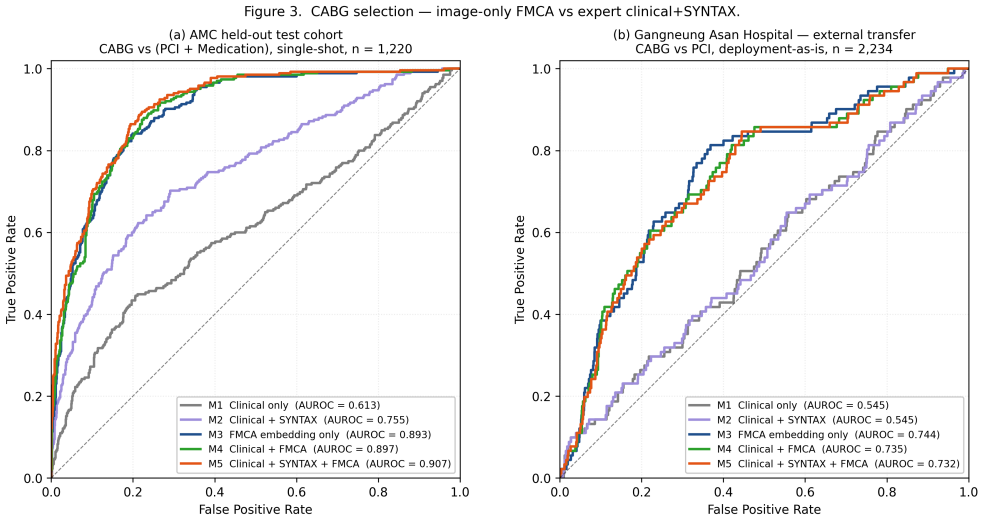


Fig. 3 Video-only FMCA versus expert clinical-plus-SYNTAX for the CABG-versus-non-CABG decision. Receiver-operating-characteristic curves on *(a)* the Asan Medical Center held-out test cohort (n = 1,220) and *(b)* the Gangneung Asan Hospital external transfer cohort (n = 2,234). Five feature combinations are compared: M1 clinical only, M2 clinical with institutional SYNTAX, M3 FMCA video only, M4 clinical with FMCA, and M5 full stack (clinical, SYNTAX, and FMCA). On both cohorts, the M3 video-only curve sits above the M2 clinical-plus-SYNTAX baseline across the threshold range. At GAH (panel *b*), the electronic export did not include institutional SYNTAX or the full M1 clinical-16 panel, so M1 and M2 are numerically identical and the M3, M4, and M5 curves are visually indistinguishable, indicating that video-only deployment suffices when target-site clinical fields are unavailable. AUROC values appear in each legend, and full 95% confidence intervals and paired-Δ p-values are reported in Table 4.

A shared backbone enables site-tailored deployment across sites

The deployment-as-is transfer above keeps the AMC-trained classifier head fixed. We next asked how much a small amount of local labelling recovers, exploiting a property afforded by the foundation-model embedding. We computed a site-internal 5-fold stratified cross-validation within GAH, with the FMCA encoder kept frozen and only the linear classifier head refitted on small GAH labelled folds. Under this protocol, video-only M3 reached AUROC 0.810 ($\pm$0.048 across folds), and the full M5 reached 0.815 ($\pm$0.038). The transfer-to-internal AUROC gap for M3 was 0.066 (from 0.744 to 0.810). This gap reflects cross-site shift in the imaging-derived prediction-probability distribution rather than any deficit in the underlying imaging representation.

Site-tailoring required no encoder retraining, no transfer of patient data outside the institution, and no manual SYNTAX annotation. Unlike fixed scoring rubrics such as SYNTAX, whose definition is fixed once and applied uniformly across sites, the foundation-model embedding therefore supports site-tailored deployment from a shared backbone using only small local labelled cohorts.

Discussion

We developed FMCA, the first vision–language foundation model for coronary angiography at clinically representative scale. A single contrastive recipe, trained on routine cine angiograms paired with non-English interpretation reports, produced a representation that supports three downstream clinical tasks from cine video alone. FMCA recognised coronary anatomy with zero-shot transfer to public external datasets, recovered the disease-extent and complexity descriptors that SYNTAX scoring manually computes, and supported revascularisation-strategy selection between CABG and non-CABG. On the revascularisation-selection task, video-only FMCA exceeded expert clinical-plus-SYNTAX. The advantage was preserved at an operationally independent Korean tertiary institution under deployment-as-is conditions, satisfying the pre-specified STRONG-support external-validation rule.

The clinical relevance of this finding rests on the role that SYNTAX scoring plays in current revascularisation guidelines. The 2018 ESC/EACTS guidelines anchor strategy selection on the SYNTAX score together with patient-level clinical features[3], and a high-tertile SYNTAX score is one of the principal indications for surgical revascularisation. SYNTAX scoring requires manual annotation of every coronary segment under a 26-region scheme, with weighting by lesion length, calcification, and bifurcation involvement. SYNTAX is also limited in practice by inter-observer variability and by the computation burden of segment-by-segment annotation (operator-count distribution in our cohort is reported in Supplementary Section S11). Our results indicate that a foundation-model encoder applied directly to cine video can serve as a reproducible alternative to the manual SYNTAX computation in this decision. The encoder output is reproducible by construction and exceeded the SYNTAX-anchored expert comparator on both the internal held-out cohort and at the external institution. FMCA does not displace clinical judgement. It provides an automated decision-support layer available at the bedside immediately after image acquisition, without segment-level expert annotation.

A methodological contribution of this work is the use of non-English clinical reports for vision–language pretraining. Most multimodal embedding backbones available today are pretrained predominantly on English text, and direct contrastive alignment of Korean angiography reports with cine video would not fully exploit the backbone's representational capacity. We instead converted each Korean report to a vessel-structured English paraphrase using an open-weight language model. Cine-level contrastive learning then aligned the cine embedding with the paraphrase. The text was used only at pretraining and was never consumed at inference, so each downstream prediction depends on the cine video alone. This recipe parallels recent vision–language foundation-model work in other clinical domains[8]. The same approach should transfer to other non-English clinical settings: wherever large image–report datasets exist in a language that English-centric backbones handle poorly, the native-language reports can be paraphrased into English before contrastive alignment, as we did for Korean.

A third contribution concerns clinical deployment. Foundation-model embeddings carry multi-dimensional information that supports lightweight site-specific tailoring through refitting of a small classifier head on a frozen encoder. At GAH, this procedure improved the video-only CABG-selection AUROC from 0.744 under deployment-as-is

to 0.810 with site-internal cross-validation. The procedure required no encoder retraining, no transfer of patient data outside the institution, and no manual SYNTAX annotation. Fixed scoring rubrics such as SYNTAX cannot be refit in this way. Their definition is established once and applied uniformly across sites. Site-tailored deployment from a shared backbone therefore offers a structural alternative for institutions whose patient population, operator practices, or imaging fleet differ from those of the pretraining cohort.

Several limitations remain. External validation is restricted to two operationally independent Korean tertiary institutions, and multi-region Western validation is required before broader deployment. The reference standard for the revascularisation task is the Heart-Team's actual treatment decision rather than a patient outcome, so FMCA reproduces and standardises expert decisions rather than being validated against clinical endpoints. The principal decision is reported as CABG-versus-non-CABG, with finer three-class outcomes and subgroup analyses in Supplementary Sections S4 and S5, and the design is retrospective. The prevalence shift between AMC (21.7% CABG) and GAH (4.1% CABG) left the deployment-as-is probabilities poorly calibrated to the GAH base rate, and isotonic recalibration corrected this without altering discrimination (Supplementary Section S12.5). Routine local recalibration should therefore be expected at sites with substantially different prevalence. The institutional SYNTAX comparator was documented for 37.7% of the held-out test cohort, and a pre-specified sensitivity analysis on this subset confirmed the M3 advantage under the strongest comparator (Supplementary Section S6.4). Future work will pursue multi-region external validation, prospective and outcome-anchored deployment, and integration with catheterisation-laboratory workflows.

More broadly, our results show that a single vision–language foundation model, learned from routine angiographic reports without segment-level annotation, can deliver reproducible video-only decision support across the coronary angiographic pipeline. This offers a scalable and site-adaptable path towards automated support for Heart-Team revascularisation decisions.

Methods

Ethics statement

This study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis with Artificial Intelligence (TRIPOD+AI) reporting guidance for clinical prediction models that use machine-learning methods[16]. TRIPOD+AI adherence is documented in Supplementary Section S13. Retrospective use of de-identified imaging and structured-report data was approved by the Asan Medical Center (AMC) Institutional Review Board (IRB No. 2023-0186), with waiver of informed consent. External validation at Gangneung Asan Hospital (GAH) was approved by the Gangneung Asan Hospital Institutional Review Board (IRB No. 2024-09-010). No patient overlap exists between the AMC pretraining cohort and any external cohort.

Study cohorts

Three patient cohorts and two public external datasets were used, with their roles, splits, and downstream evaluation assignments summarised in Figure 2. Each is described in full below.

Internal cohort at the Asan Medical Center. Asan Medical Center is a high-volume tertiary cardiovascular centre in Seoul, Republic of Korea, with an established interventional cardiology and cardiac surgery programme and a multidisciplinary Heart Team. Between 2002 and 2018, 12,344 consecutive patients undergoing invasive coronary angiography (CAG) at AMC were retrospectively enrolled. The cohort yielded 113,233 cine acquisitions, with a mean of 9.2 cines per patient (range 1 to 22) and 35 to 200 frames per cine. For every patient we obtained the structured Korean interpretation report authored by an interventional cardiologist, 16 baseline clinical variables, and the definitive treatment strategy selected during the index hospitalisation, recorded as percutaneous coronary intervention (PCI), coronary artery bypass grafting (CABG), or optimal medical therapy. The patient-level stratified split fixed under Statistical analysis yielded 10,490 training, 618 validation, and 1,235 test patients. Cohort baseline characteristics are reported in Table 1.

Public external datasets for coronary anatomy recognition. Two publicly available datasets totalling 12,060 angiographic acquisitions served as zero-shot benchmarks. CoronaryDominance[11] (Russia) is a dataset of 1,577 patient studies and 10,560 cines acquired on two Philips imaging systems, Allura Clarity and Azurion, providing cine-level left coronary artery (LCA) versus right coronary artery (RCA) labels and patient-level left-versus-right coronary dominance labels. ARCADE[10] (Kazakhstan) is a publicly released benchmark of 1,500 single-frame angiograms released with the Medical Image Computing and Computer Assisted Intervention 2023 Automatic Region-based Coronary Artery Disease diagnostics Challenge, providing single-frame LCA-versus-RCA labels.

External validation cohort at Gangneung Asan Hospital. GAH is an operationally independent Korean tertiary institution approximately 200 km from AMC. The two institutions share no clinicians, scanner fleet, or patient overlap, and they operate independent Heart Teams. Between 2003 and 2023, 2,234 consecutive patients with a definitive revascularisation label were retrospectively enrolled, of whom 2,143 received PCI and 91 received CABG. The GAH revascularisation registry contains no medical-therapy-only stratum, and the revascularisation strategy selection task at GAH is therefore operationally CABG versus PCI. The GAH electronic export contained age, sex, ejection fraction, serum creatinine, and diabetes mellitus, but did not include the institutional SYNTAX score, vessel-level anatomy labels, or the remaining baseline clinical fields. The deployment-as-is handling of these absent fields is described under Downstream evaluation.

Image and report preprocessing

Cine sampling. Each cine acquisition was uniformly subsampled to 35 RGB frames spanning the full acquisition, and cines with fewer than 5 extracted frames were excluded. At training time, frame indices were drawn with a random jitter of ± 1.5

frames around the uniform grid. Sampled frames were augmented with brightness adjustment (±10%), contrast adjustment (±10%), rotation (±3°), and additive Gaussian noise of standard deviation 2 on the 0–255 intensity scale. At inference time, deterministic uniform sampling without augmentation was used. The DICOM primary and secondary positioner angles were extracted from the acquisition header and prepended to the cine as a text token during pretraining, supplying view-angle context to the multimodal backbone.

Korean-to-English vessel-structured paraphrase. Each AMC patient's structured Korean interpretation report was converted to a vessel-organised English paraphrase by the open-weight Qwen3-Next-80B-A3B-Instruct large language model[13], a mixture-of-experts variant of Qwen3 with 80 billion total parameters and 3 billion active parameters per inference step. Paraphrasing used a deterministic system prompt with temperature 0 and top-p 1, producing strict JSON output. Each report was reorganised into a `common` field for procedural context, together with separate `LCA` and `RCA` fields for vessel-specific findings. All reported numeric measurements were preserved verbatim, including stenosis percentage, fractional flow reserve (FFR), lesion length, and flow grade. The paraphrased text was used only during contrastive pretraining and was not consumed by any downstream classifier at inference. Full prompt constraints, fallback rules, and automated full-corpus quality-control statistics are reported in Supplementary Section S3.

Recommendation-text scrub. Treatment-recommendation sentences in the `common` field of each paraphrased report were programmatically removed before pretraining, to prevent leakage of the downstream treatment label into the contrastive objective. The scrub script and the cleaned-corpus SHA-256 checksum are reported in Supplementary Section S9, together with a strong-form-removal sensitivity ablation that retrains the encoder without any `common`-field text.

Model architecture and contrastive pretraining

Backbone and adapter. FMCA uses Qwen3-VL-Embedding-8B[15] as its backbone, an open-weight multilingual video-capable multimodal embedding model, with the backbone parameters frozen during contrastive pretraining. Trainable parameters are restricted to a low-rank adapter (LoRA)[14] of rank 32, $\alpha = 64$, and dropout 0.05, inserted into the query, key, value, and output projections together with the feed-forward projections of every transformer block.

Contrastive pretraining. Each cine and its paraphrased report formed a positive pair for cine-level InfoNCE training[17], with all other pairs in the global batch serving as negatives. Cross-rank gradient gathering pooled negatives across GPUs to expand the effective contrastive batch. The contrastive temperature was a learnable parameter initialised at 0.07.

Optimisation. Optimisation used AdamW[18] with learning rate 2×10^{-4}, weight decay 0.01, cosine learning-rate decay, and a 500-step warm-up. Each optimisation step combined a per-device batch of 4 across two GPUs to yield 8 cines per InfoNCE matrix, with 16 gradient accumulation steps. Training used bfloat16 mixed precision on two NVIDIA RTX PRO 6000 GPUs and was configured for up to 10 epochs with early stopping. The validation contrastive loss reached its minimum at epoch 2 (0.507),

training halted during epoch 6 after a three-epoch plateau, and the epoch-2 checkpoint was adopted, at approximately 4 GPU-hours of compute.

Video-only downstream inference. All downstream FMCA predictions take a cine video as the sole input at inference. The frozen FMCA encoder maps each cine to a 4,096-dimensional embedding, and per-patient embeddings are obtained as the mean across that patient's cines. A task-specific linear classifier head, fitted as described under Downstream evaluation, then maps the patient embedding to the predicted label. The paraphrased English report is not consumed at inference, including for any external-validation prediction.

Downstream evaluation

Three downstream tasks were evaluated: coronary anatomy recognition, angiographic disease extent and complexity stratification, and revascularisation strategy selection. For each task, a hyperparameter search over logistic regression (L1 or L2 penalty), a multi-layer perceptron, and a gradient-boosted decision tree was conducted on the AMC training and validation splits. The validation-selected combination was then evaluated once on the AMC held-out test split, with no test-set tuning. The FMCA encoder was kept frozen throughout. For the disease-extent and revascularisation-strategy tasks, per-patient embeddings were obtained by mean-pooling the 4,096-dimensional embeddings of that patient's cines. The anatomy task was evaluated at the level of individual cines or patient studies, as described below.

Coronary anatomy recognition. Vessel-identification labels were derived from the operator's catheter engagement recorded for 745 manually annotated AMC cines from 199 patients, a cine-level annotation orthogonal to the patient-level training, validation, and test split used for the disease-extent and revascularisation-strategy tasks. Each cine was labelled as left coronary artery (LCA) or right coronary artery (RCA). An L2-regularised logistic-regression classifier was fitted on these 745 cines and applied without further fitting to the external CoronaryDominance and ARCADE acquisitions for zero-shot evaluation. The 199 vessel-identification patients distribute across the AMC training, validation, and test partitions as 176, 13, and 10 patients, respectively. This cross-experiment overlap does not compromise the independence of the disease-extent and revascularisation-strategy evaluations, which use separate classifier heads on disjoint label sets (Supplementary Section S2). Coronary dominance was evaluated as a separate task on the 1,577 CoronaryDominance patient studies, defined by the artery supplying the posterior descending artery, with right-dominant and left-dominant labels provided by the dataset.

Angiographic disease extent and complexity stratification. Per-patient vessel-disease labels DIS_LAD, DIS_LCX, and DIS_RCA were taken as binary operator-recorded indicators of significant disease from the AMC structured report, with missing labels treated as no disease, reflecting a reporting convention adopted by the AMC catheterisation laboratory rather than a fixed stenosis-percentage threshold. Multi-vessel disease (MVD) was defined as two or more diseased vessels among the left anterior descending (LAD), left circumflex (LCX), and right coronary (RCA) arteries, with left main not counted. Three-vessel disease (3VD) was defined as all three of LAD, LCX, and RCA diseased. The institutional SYNTAX score was taken

as a continuous value from 0 to 78, and tertile classification followed the SYNTAX trial cutoffs[2] adopted by the 2018 ESC/EACTS revascularisation guidelines[3]: Low ≤ 22, Intermediate 23 to 32, and High ≥ 33. Among the 12,186 AMC patients with a definitive treatment label, 4,614 (37.9%) had a documented SYNTAX score. The 3-class SYNTAX tertile result is reported as macro-AUROC. Per-class binary AUROCs and per-vessel disease-detection results are reported in Supplementary Section S1.1.

Revascularisation strategy selection. The definitive treatment strategy selected during the index hospitalisation was recorded as percutaneous coronary intervention (PCI), coronary artery bypass grafting (CABG), or optimal medical therapy (OMT). The principal Heart-Team decision-support analysis compared CABG versus non-CABG, with PCI and OMT pooled, on the AMC held-out test cohort. Five feature combinations were evaluated:

- **M1**, 16 clinical variables (age, sex, body-mass index, ejection fraction, serum creatinine, hypertension, diabetes mellitus, smoking, hyperlipidemia, prior myocardial infarction, prior stroke, prior PCI, prior CABG, atrial fibrillation, chronic obstructive pulmonary disease, and chronic kidney disease).
- **M2**, M1 with the institutional SYNTAX score added.
- **M3**, the FMCA video embedding alone, projected to 64 dimensions by principal-component analysis fitted on the AMC training and validation splits to mitigate overfitting from the full 4,096-dimensional embedding.
- **M4**, M1 with the FMCA embedding added.
- **M5**, M1 with the SYNTAX score and the FMCA embedding added.

A pre-specified L2-regularised logistic regression with $C = 0.01$ and the lbfgs solver was fit on the 10,966 AMC training and validation patients with a definitive treatment label, then evaluated once on the held-out test split. No hyperparameter tuning was performed on the held-out test data. Additional task variants (PCI versus CABG, invasive versus medical, and a 3-class macro outcome) are reported in Supplementary Section S4.

External validation at Gangneung Asan Hospital. The revascularisation-strategy task was externally validated under a deployment-as-is protocol, so that the GAH numbers remain directly comparable to the AMC held-out result. Each GAH patient's cines were passed through the same frozen FMCA encoder and mean-pooled to a 4,096-dimensional patient embedding, identical to the AMC internal pipeline. The five M1–M5 classifiers fit on the AMC training and validation pool were then applied without further fitting to the GAH embeddings and clinical features, with no GAH outcome label observed by any classifier during fitting. The GAH electronic export lacked the institutional SYNTAX score and parts of the M1 clinical-16 panel. These absent fields were imputed at the median of the AMC training and validation pool, which makes M1 and M2 numerically identical at GAH without affecting the central M3 evaluation. Independently, a site-internal benchmark was computed by 5-fold stratified cross-validation within GAH, with the FMCA encoder kept frozen and only the linear classifier head refitted on each fold. Calibration, continuous net reclassification improvement, integrated discrimination improvement, and pre-specified age-stratum subgroup analyses are reported in Supplementary Section S12. The pre-specified

external-validation rule (Supplementary Section S10) defines STRONG support as ΔAUROC(M3 versus M2) $\geq +0.05$ and ΔAUROC(M5 versus M2) $\geq +0.05$, with both 95% confidence intervals strictly above zero. This rule was fixed in writing before any GAH analysis was run.

Pre-specified subgroup and sensitivity analyses. Subgroup analyses on the AMC held-out test were pre-specified by clinical presentation (stable angina, unstable angina, non-ST-elevation myocardial infarction, and ST-elevation myocardial infarction) and by age stratum (less than 60 years, 60 to 75 years, and 75 years or older), with results reported in Supplementary Section S5. Sensitivity analyses include per-model regularisation tuning (Supplementary Section S6.3) and a SYNTAX-known-subset analysis restricted to AMC patients with a documented institutional SYNTAX score (Supplementary Section S6.4).

Statistical analysis

AUROC point estimates and pairwise ΔAUROCs were reported with 95% confidence intervals from a 1,000-iteration patient-level bootstrap of the held-out test predictions[19], with predictions fixed and patients resampled with replacement. Paired-Δ p-values used the same paired bootstrap distribution, with two-sided tail mass beyond zero. Within the treatment task family (four outcomes by five non-trivial contrasts), the false-discovery rate across all 20 paired comparisons was controlled by Benjamini–Hochberg[20], with the full panel reported in Supplementary Section S4.2. Calibration was quantified by Brier score and expected calibration error. Decision-curve analysis was computed on the AMC held-out test (Supplementary Section S6.2). Continuous net reclassification improvement and integrated discrimination improvement were computed on both the AMC held-out test and the GAH cohort (Supplementary Sections S6.2 and S12.5).

The AMC cohort was split at the patient level under stratification by treatment and clinical presentation, with random seed 42, yielding 10,490 training (85.0%), 618 validation (5.0%), and 1,235 test (10.0%) patients. The classifier fit and bootstrap seeds were also set to 42.

Computing software and hardware

Contrastive pretraining used two NVIDIA RTX PRO 6000 GPUs with bfloat16 mixed precision. Downstream classifier fitting (logistic regression, multi-layer perceptron, gradient-boosted decision trees, and principal-component analysis) ran on CPU. The software stack was Python 3.12, PyTorch 2.11[21], and the Hugging Face Transformers 4.53, PEFT 0.13, and Accelerate 0.35 libraries for backbone access, LoRA adaptation, and multi-GPU orchestration, respectively. Downstream classifiers used scikit-learn 1.7[22] (logistic regression, multi-layer perceptron, principal-component analysis, and bootstrap implementation) and XGBoost 3.0[23] (gradient-boosted decision trees).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article. TRIPOD+AI[16] adherence is documented in Supplementary Section S13.

Data availability. The CoronaryDominance[11] and ARCADE[10] datasets used for zero-shot coronary anatomy evaluation are publicly available at the URLs cited in the respective references. The patient-level imaging, structured-report, and clinical data from Asan Medical Center and Gangneung Asan Hospital, including the English paraphrase corpus derived from the patient reports, cannot be shared. These data contain sensitive personal health information, and their transfer outside the institutions is prohibited under the patient-privacy regulations of the Republic of Korea and the governing institutional review board approvals (AMC IRB No. 2023-0186 and Gangneung Asan Hospital IRB No. 2024-09-010).

Code availability. The training, evaluation, and statistical-analysis code is provided to the editor and reviewers as a reproducibility package (`code_for_reviewer/`) accompanying this submission. The package contains scripts for the recommendation-text scrub, the cine-level contrastive trainer, the embedding extractor, the downstream classifier and treatment-decision evaluators, and the external-validation evaluators. The Korean-to-English paraphrase pipeline (the large-language-model caller and the verbatim system prompt) is not included in the reviewer code package. The high-level prompt constraints and automated full-corpus quality-control statistics are reported in Supplementary Section S3. The full code is available from the corresponding author on reasonable request post-publication.

Acknowledgements. This work used the open-weight Qwen3-Next-80B-A3B-Instruct large language model for the report paraphrase preprocessing step only. No model distillation was performed, and no API-based external services were used in any step that touched protected patient data. We thank the patients of Asan Medical Center and Gangneung Asan Hospital whose anonymised data made this work possible, and the cardiology services of both institutions for their collaboration in assembling the external-validation cohort.

Declarations

Funding. This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number RS-2023-KH136685). This research was also supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HR21C0198).

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. The retrospective use of de-identified imaging and structured-report data was approved by the Asan Medical Center Institutional Review Board (IRB No. 2023-0186) with waiver of informed consent. External validation at Gangneung Asan Hospital was approved by its Institutional Review Board (IRB No. 2024-09-010). All methods were performed in accordance with the relevant guidelines and regulations.

Author contributions. **Byeolhee Kim**: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft. **Hanbit Park**: Data curation, Validation. **Ju Hyeon Kim**: Data curation, Methodology. **Seong-Bong Wee**: Methodology, Software. **Do-Yoon Kang**: Validation, Resources. **Duk-Woo Park**: Resources, Supervision. **Seung-Jung Park**: Resources, Supervision. **Young-Hak Kim**: Conceptualization, Writing - review & editing, Supervision. **Jung-Min Ahn**: Conceptualization, Methodology, Writing - review & editing, Supervision. **Tae Joon Jun**: Conceptualization, Methodology, Software, Writing - review & editing, Supervision, Project administration.

References

[1] Roth, G. A. *et al.* Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. *Journal of the American College of Cardiology* **76**, 2982–3021 (2020).

[2] Serruys, P. W. *et al.* Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. *New England Journal of Medicine* **360**, 961–972 (2009).

[3] Neumann, F.-J. *et al.* 2018 ESC/EACTS guidelines on myocardial revascularization. *European Heart Journal* **40**, 87–165 (2019).

[4] Lawton, J. S. *et al.* 2021 ACC/AHA/SCAI guideline for coronary artery revascularization: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *Journal of the American College of Cardiology* **79**, e21–e129 (2022).

[5] Sianos, G. *et al.* The SYNTAX score: an angiographic tool grading the complexity of coronary artery disease. *EuroIntervention* **1**, 219–227 (2005).

[6] Généreux, P. *et al.* SYNTAX score reproducibility and variability between interventional cardiologists, core laboratory technicians, and quantitative coronary measurements. *Circulation: Cardiovascular Interventions* **4**, 553–561 (2011).

[7] Zhang, S. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* (2024).

[8] Ding, T. *et al.* A multimodal whole-slide foundation model for pathology. *Nature Medicine* **31**, 3749–3761 (2025).

[9] Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).

[10] Popov, M. *et al.* Dataset for Automatic Region-based Coronary Artery Disease Diagnostics using X-ray angiography images. *Scientific Data* **11**, 20 (2024).

[11] Kruzhilov, I. *et al.* CoronaryDominance: Angiogram dataset for coronary dominance classification. *Scientific Data* **12**, 341 (2025).

[12] Kim, B., Kim, J., Kim, Y.-H., Jun, T. J. & Ahn, J.-M. Deep learning-based treatment decision support framework for multi-vessel coronary artery disease using integrated coronary angiography and clinical data. *BMC Medical Informatics and Decision Making* (2026). In press.

[13] Yang, A. *et al.* Qwen3 technical report. *arXiv preprint* (2025).

[14] Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. *arXiv preprint* (2021).

[15] Li, M. *et al.* Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint* (2026).

[16] Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).

[17] van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint* (2018).

[18] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019). ArXiv:1711.05101.

[19] Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26 (1979).

[20] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

[21] Paszke, A. *et al. PyTorch: An imperative style, high-performance deep learning library*, 8024–8035 (2019).

[22] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[23] Chen, T. & Guestrin, C. *XGBoost: A scalable tree boosting system*, 785–794 (2016).

Extended Data

Extended Data Table 1 Baseline characteristics of the Asan Medical Center cohort by patient-level data split.

Variable	Train (n = 10,490)	Val (n = 618)	Test (n = 1,235)	p (BH)
Age (years)	64.0 [57.0–70.0]	63.0 [57.0–71.0]	63.0 [56.0–70.0]	0.893
Male sex, n (%)	7,941 (75.7%)	448 (72.5%)	934 (75.6%)	0.893
Body-mass index (kg/m^2)	24.8 [23.0–26.8]	24.9 [23.0–26.9]	24.9 [23.1–26.8]	0.893
LV ejection fraction (%)	60.0 [56.0–64.0]	60.0 [55.0–64.0]	61.0 [56.0–64.0]	1.000
Serum creatinine (mg/dL)	0.9 [0.8–1.1]	0.9 [0.8–1.1]	0.9 [0.8–1.1]	1.000
SYNTAX score	18.0 [12.0–26.0]	17.0 [10.0–28.0]	19.0 [11.0–27.0]	1.000
Hypertension, n (%)	6,786 (64.7%)	401 (64.9%)	811 (65.7%)	1.000
Diabetes mellitus, n (%)	3,709 (35.4%)	222 (35.9%)	431 (34.9%)	1.000
Hyperlipidemia, n (%)	6,373 (60.8%)	370 (59.9%)	730 (59.1%)	1.000
Current smoker, n (%)	2,536 (24.2%)	148 (23.9%)	302 (24.5%)	1.000
Atrial fibrillation, n (%)	235 (2.2%)	12 (1.9%)	15 (1.2%)	0.893
COPD, n (%)	178 (1.7%)	10 (1.6%)	30 (2.4%)	0.893
Chronic kidney disease, n (%)	356 (3.4%)	25 (4.0%)	55 (4.5%)	0.893
Prior myocardial infarction, n (%)	617 (5.9%)	38 (6.1%)	66 (5.3%)	1.000
Prior stroke, n (%)	713 (6.8%)	48 (7.8%)	97 (7.9%)	0.893
Prior PCI, n (%)	1,558 (14.9%)	95 (15.4%)	192 (15.5%)	1.000
Prior CABG, n (%)	31 (0.3%)	1 (0.2%)	1 (0.1%)	0.893
Left main disease, n (%)	1,862 (17.8%)	127 (20.6%)	230 (18.6%)	0.893
LAD disease, n (%)	9,115 (86.9%)	533 (86.2%)	1,073 (86.9%)	1.000
LCX disease, n (%)	6,442 (61.4%)	379 (61.3%)	735 (59.5%)	1.000
RCA disease, n (%)	6,595 (62.9%)	393 (63.6%)	736 (59.6%)	1.000
Clinical presentation, n (%)				1.000
Stable angina	5,174 (49.3%)	305 (49.4%)	610 (49.4%)	
Unstable angina	2,515 (24.0%)	149 (24.1%)	296 (24.0%)	
NSTEMI	543 (5.2%)	32 (5.2%)	64 (5.2%)	
STEMI	206 (2.0%)	12 (1.9%)	24 (1.9%)	
Treatment received, n (%)				1.000
PCI	7,540 (71.9%)	444 (71.8%)	888 (71.9%)	
CABG	2,249 (21.4%)	133 (21.5%)	265 (21.5%)	
Optimal medical therapy	567 (5.4%)	33 (5.3%)	67 (5.4%)	

Numeric variables are reported as median [interquartile range] and categorical variables as n (%). Overall-cohort values are reported in Table 1 of the main manuscript. The patient-level stratified split (treatment by clinical presentation, random seed 42, Methods) is balanced across every clinical and demographic variable, with all Benjamini–Hochberg-adjusted Kruskal–Wallis (continuous) or χ^2 (categorical) p-values ≥ 0.121. Variable-level missingness is reported in Supplementary Section S8.1. Treatment received refers to the definitive revascularisation strategy selected during the index hospitalisation. COPD, chronic obstructive pulmonary disease. NSTEMI, non-ST-elevation myocardial infarction. STEMI, ST-elevation myocardial infarction.

Supplementary Information: A Vision-Language Foundation Model for Coronary Angiography Supports Heart-Team Treatment Selection

S1. Internal downstream panel (AMC held-out test cohort)

We report the principal internal tasks evaluated on the AMC held-out test cohort ($n = 1{,}235$ patients withheld from contrastive pretraining).

S1.1 Per-task results

Supplementary Table 1 reports the per-task internal results. For each task, hyperparameter search was conducted on the train and validation split ($n = 11{,}108$). The FMCA test metric uses the val-selected classifier-and-hyperparameter combination evaluated once on the held-out test split. The Base column reports the frozen Qwen3-VL-Embedding-8B backbone (no LoRA, no contrastive training) on the same patient-level split, using a default L2-regularised logistic regression ($C = 1.0$, no hyperparameter search), included as a reference to isolate the contribution of contrastive pretraining.

S1.2 Notes on individual tasks

- **DIS_LAD** has 87% positive prevalence in the test set (1,073 of 1,235). The binary AUROC ceiling is therefore tight, and F_1 and accuracy are more informative metrics for this task.
- All vessel-disease binary labels follow the original metadata convention of treating "missing" as "not-positive".
- The continuous SYNTAX-score regression sample is restricted to the 4,614 AMC patients with a documented SYNTAX score. The patient-level split (85 / 5 / 10) gives 465 SYNTAX-known patients in the held-out test cohort.
- The 3-class tertile macro-AUROC is the equal-weighted mean of the three per-class binary AUROCs at the Serruys 2009 and ESC 2018 cutoffs.
- The macro-AUROC (0.748) is driven by separation of the decision-relevant Low ($\leq$ 22) and High ($\geq$ 33) classes (per-class AUROC 0.770 and 0.842). The intermediate band (23 to 32) is the clinically ambiguous zone in which guideline recommendations themselves overlap, and its per-class AUROC (0.632) is correspondingly lower.

Supplementary Table 1 Internal per-task results on the AMC held-out test cohort.

Task	Metric	n_{test}	n_{pos}	Base [95% CI]	FMCA [95% CI]
MVD (≥2-vessel)	AUROC	1,235	901	0.594 [0.556, 0.627]	**0.810 [0.782, 0.836]**
3VD (3-vessel)	AUROC	1,235	426	0.623 [0.588, 0.655]	**0.804 [0.777, 0.828]**
SYNTAX score (continuous)	R^2	465	—	—	**0.330 [0.210, 0.403]**
SYNTAX 3-class tertile	macro-AUROC	465	—	—	**0.748 [0.712, 0.783]**
Low (≤ 22) vs rest	AUROC	465	288	—	0.770
Intermediate (23–32) vs rest	AUROC	465	116	—	0.632
High (≥ 33) vs rest	AUROC	465	61	—	**0.842**
DIS_LAD	AUROC	1,235	1,073	—	0.652 [0.608, 0.692]
DIS_LCX	AUROC	1,235	735	—	0.751 [0.721, 0.774]
DIS_RCA	AUROC	1,235	736	0.598 [0.566, 0.630]	**0.787 [0.762, 0.810]**
n_{diseased} (0–3)	macro-AUROC	1,235	—	—	0.734 [0.695, 0.772]
Treatment (3-class)	macro-AUROC	1,220	—	0.595 [0.563, 0.628]	0.798 [0.776, 0.829]

Both Base and FMCA 95% CIs are 1,000-iteration patient-level bootstraps. FMCA CIs are computed by refitting at the val-selected hyperparameters. The bootstrap point centre reproduces the hyperparameter-search test metric to ≤ 0.005 across the binary and multiclass tasks. The SYNTAX-score regression uses an MLP with early stopping whose initialisation noise produces a slightly wider reproduction band (≤ 0.025 R^2), so the original 0.330 point estimate is retained alongside the bootstrap-derived CI of [0.210, 0.403]. The 3-class SYNTAX tertile macro-AUROC and per-class binary AUROCs are derived from the regression model's per-sample test predictions thresholded at the clinical-standard Serruys 2009 and ESC 2018 cutoffs (Low ≤ 22, Intermediate 23 to 32, High ≥ 33). Validation-selected FMCA classifiers: L1-regularised logistic regression ($C = 0.1$) for MVD, 3VD, and DIS_LCX. L2-regularised logistic regression ($C = 0.01$) for DIS_LAD, DIS_RCA, and n_{diseased}. A multi-layer perceptron ($h = 128$, $\alpha = 0.01$) for the SYNTAX-score regression. Gradient-boosted trees ($n = 400$, depth 4) for the 3-class treatment task. Base reports a default L2 logistic regression ($C = 1.0$).

S2. Anatomy supplementary analyses

S2.1 AMC internal vessel-identification classifier: 5-fold GroupKFold sanity check

The vessel-identification classifier used in the main manuscript was fit on 745 manually labelled AMC cines from 199 patients (cine-level annotation orthogonal to the patient-level train, validation, and test split). Independent of the external zero-shot transfer reported in the main text, we evaluated the same classifier on the AMC labelled cines by 5-fold GroupKFold cross-validation (group = patient identifier, preventing within-patient leakage across folds), with results in Supplementary Table 2.

S2.2 Patient overlap between the 745-cine annotation and the patient-level split

The 199 patients contributing the 745 labelled cines distribute across the patient-level split as shown in Supplementary Table 3.

The 10 of 1,235 patients shared between the 745-cine vessel-identification training set and the held-out test cohort constitute a cross-experiment overlap, not a within-experiment leak. The vessel-identification head and the disease-extent and revascularisation-strategy classifier heads are independent classifiers trained on disjoint label sets (cine-level vessel labels versus patient-level disease, SYNTAX, and treatment

Supplementary Table 2 AMC internal vessel-identification classifier: 5-fold GroupKFold sanity check on 745 cines from 199 patients.

Classifier	Accuracy ± SD	F_1 (RCA-positive)	AUROC
LogReg L2 ($C = 1.0$)	0.938 ± 0.021	0.887	0.973
LogReg L2 ($C = 0.1$)	**0.937 ± 0.021**	**0.884**	**0.974**
Linear SVM	0.940 ± 0.020	0.890	0.970
MLP (256 hidden units)	0.894 ± 0.017	0.800	0.925

The LogReg L2 ($C = 0.1$) head was selected and is the linear probe applied to the external acquisitions in the main manuscript. The internal cross-validation AUROC of 0.974 ± 0.020 is comparable to the external zero-shot AUROC of 0.970 reported in the main manuscript. The parity indicates that anatomic vessel-identity features encoded in the FMCA embedding transfer cleanly across institutions and imaging platforms.

Supplementary Table 3 Distribution of the 199 vessel-identification patients across the AMC patient-level split.

Pool	Vessel-ID patients	% of vessel-ID set	% of pool
Train (10,490)	176	88.4%	1.68%
Val (618)	13	6.5%	2.10%
Test (1,235)	10	5.0%	0.81%

labels). The FMCA encoder is frozen across both analyses, so the encoder representation of any held-out test patient never received a downstream supervisory signal from that patient. The independence of the disease-extent and revascularisation-strategy evaluations is therefore preserved.

S2.3 Vessel identification (LCA versus RCA): per-split detail on CoronaryDominance and ARCADE

Supplementary Table 4 gives the per-split breakdown of the zero-shot transfer reported in the main manuscript (combined AUROC 0.970 [0.967, 0.973]).

S2.4 Coronary dominance: per-split and RCA-only sub-analyses

The dominance probe is a logistic-regression head fit and evaluated by 5-fold cross-validation on the external CoronaryDominance studies (dominance labels do not exist in the AMC pretraining cohort). Per-split and RCA-only results are reported in Supplementary Table 5.

S3. Vessel-structured English paraphrase pipeline

S3.1 Goal and target schema

The pipeline reorganises Korean structured CAG reports into a vessel-aligned English representation that pairs each cine with anatomically focused text. The target JSON schema is shown below.

Supplementary Table 4 Per-split zero-shot vessel-identification AUROC on the external public datasets.

Split	Imaging system	n cines	n RCA-pos.	Zero-shot AUROC [95% CI]
CoronaryDominance, main split	Philips Allura Clarity	6,764	1,783	0.968 [0.964, 0.972]
CoronaryDominance, real-distribution split	Philips Allura Clarity	2,890	857	0.981 [0.977, 0.985]
CoronaryDominance, domain-shift split	Philips Azurion (distinct scanner)	906	234	0.947 [0.930, 0.961]
CoronaryDominance, combined	mixed	10,560	2,874	0.970 [0.967, 0.973]
ARCADE single-frame	Kazakhstan, single-frame	1,500	—	**0.995**

The "main" split is the dataset authors' standard benchmark split. The "real-distribution" split is held out and preserves the epidemiologically realistic class prevalence. The "domain-shift" split is acquired on a different Philips imaging system (Azurion versus Allura Clarity used for the main and real splits) and serves as a cross-scanner out-of-distribution test. AUROC $\geq$ 0.94 on every split is direct evidence that FMCA vessel-identification representations transfer without target-domain supervision across class-prevalence and scanner-platform shifts.

Supplementary Table 5 Coronary-dominance probe: per-split and RCA-only AUROC.

Split	n studies	n right-dom.	AUROC (all cines) ± SD	AUROC (RCA only) ± SD
CoronaryDominance, main split	1,028	709 (69.0%)	0.922 ± 0.007	0.914 ± 0.007
CoronaryDominance, real-distribution	400	350 (87.5%)	0.807 ± 0.062	0.880 ± 0.031
CoronaryDominance, domain-shift	149	97 (65.1%)	0.742 ± 0.118	0.857 ± 0.119
Combined	**1,577**	**1,156 (73.3%)**	**0.924 ± 0.016**	**0.922 ± 0.012**

The published supervised state of the art on this dataset (ConvNeXt with majority vote)[11] reports accuracy 89.4% on the domain-shift split. The FMCA dominance probe reaches 0.870 accuracy and AUROC 0.924 across the larger combined cohort. The RCA-only sub-analyses confirm that dominance information is concentrated in RCA-injected cines, consistent with the standard anatomic observation that the posterior-descending-artery-origin segment is most visible in RCA acquisitions.

```
{
  "common":   "Procedure context, overall finding, recommendation",
  "LCA":      "Left coronary artery findings
               (segments, stenosis, morphology)",
  "RCA":      "Right coronary artery findings",
  "segments": [
    {"name": "pLAD", "vessel": "LCA",
     "stenosis": "80%", "lesion_length": "tubular 10-20mm"}
  ]
}
```

S3.2 Model and configuration

- **Model.** Qwen3-Next-80B-A3B-Instruct[13] (open-weight large language model, mixture-of-experts, 80 billion total parameters with 3 billion active per inference step, instruction-tuned).
- **Inference.** 16 concurrent workers on a single GPU node, approximately 40 minutes for the full corpus of 12,344 reports.

- **Sampling.** Temperature 0, top-p 1 (deterministic JSON output).
- **Output validation.** Strict JSON parsing with a fallback single retry on parse failure.

S3.3 System prompt: key constraints

The system prompt is deterministic and enforces the following constraints. The verbatim prompt is available from the corresponding author on reasonable request.

1. **Translate** all Korean clinical text to concise English. Preserve numeric values verbatim (stenosis %, FFR, minimal lumen area, plaque burden, TIMI flow grade).
2. **Reorganise** content into the three labelled fields (`common`, `LCA`, `RCA`).
3. **Vessel routing rule (critical).** Every coronary segment is assigned to the `LCA` or `RCA` family using a deterministic AHA mapping. Left main, LAD, LCX, diagonal, obtuse marginal, and ramus segments route to `LCA`. RCA, posterior descending, posterolateral, and conus segments route to `RCA`.
4. **Two-way fallback rule.** If a vessel family has zero findings, the field receives exactly the canonical sentence *"No significant stenosis in the [left, right] coronary artery."* If any finding is present, the fallback sentence must not appear.
5. **JSON output only.** No prose or markdown fences.

Across all 12,344 paraphrased reports, automated full-corpus quality control confirmed zero residual Korean characters, zero violations of the two-way fallback rule, and no unrecovered API errors.

S4. AMC additional treatment tasks

Beyond the headline CABG-versus-non-CABG comparison reported in the main manuscript, the AMC held-out test cohort supports three additional treatment-decision tasks. These are PCI versus CABG (medical therapy excluded), invasive versus medical (any revascularisation versus medical therapy alone), and the 3-class macro-AUROC across PCI, CABG, and medical therapy. All evaluations follow the same protocol as the main text (held-out single-shot; classifiers fit on the train and validation pool with $C = 0.01$ fixed; per-patient 4,096-dimensional embedding projected to PCA-64; video-only inference for M3, M4, and M5).

S4.1 Per-task M1–M5 AUROCs on the AMC held-out test cohort

Supplementary Table 6 reports the M1–M5 AUROCs for each of the four treatment-decision tasks on the AMC held-out test cohort.

S4.2 Paired Δ AUROCs and unadjusted *p*-values (Benjamini–Hochberg controlled)

Held-out single-shot evaluation with paired bootstrap (1,000 iterations, two-sided tail mass beyond zero), Benjamini–Hochberg adjustment[20] applied across all 20 paired comparisons (4 tasks × 5 non-trivial contrasts).

Supplementary Table 6 Per-task M1–M5 AUROCs on the AMC held-out test cohort.

Task	n_{test}	n_{pos}	M1	M2	M3	M4	M5
CABG vs non-CABG	1,220	265	0.613	0.755	**0.893**	0.898	**0.907**
PCI vs CABG	1,153	265	0.624	0.763	**0.901**	0.906	0.915
Invasive vs medical	1,220	1,153	0.715	0.711	0.632	0.721	0.723
3-class macro	1,220	—	0.657	0.744	0.797	0.831	0.840

AUROC point estimates on the held-out test (1,000-iteration patient-level bootstrap). The 95% confidence intervals for the CABG-versus-non-CABG task are reported in Table 4 of the main manuscript, and the paired ΔAUROC confidence intervals for all four tasks are in Supplementary Table 7.

Supplementary Table 7 Paired Δ AUROCs and unadjusted p-values across the four treatment-decision tasks.

Task	Comparison	Δ AUROC	95% CI	p (unadjusted)
CABG vs non-CABG	M2 − M1	+0.141	[+0.086, +0.198]	<0.001
CABG vs non-CABG	M3 − M1	+0.280	[+0.230, +0.328]	<0.001
CABG vs non-CABG	**M3 − M2**	**+0.138**	**[+0.100, +0.175]**	**< 0.001**
CABG vs non-CABG	M4 − M1	+0.284	[+0.239, +0.328]	<0.001
CABG vs non-CABG	**M5 − M1**	**+0.294**	**[+0.247, +0.339]**	**0.001**
CABG vs non-CABG	M5 − M2	+0.153	[+0.115, +0.195]	<0.001
PCI vs CABG	M2 − M1	+0.139	[+0.085, +0.193]	<0.001
PCI vs CABG	M3 − M1	+0.278	[+0.233, +0.324]	<0.001
PCI vs CABG	M4 − M1	+0.282	[+0.237, +0.327]	<0.001
PCI vs CABG	**M5 − M1**	**+0.291**	**[+0.246, +0.336]**	**0.001**
PCI vs CABG	M5 − M2	+0.152	[+0.115, +0.190]	<0.001
Invasive vs medical	M2 − M1	−0.004	[−0.082, +0.083]	0.890
Invasive vs medical	M3 − M1	−0.084	[−0.176, +0.006]	0.060
Invasive vs medical	M4 − M1	+0.006	[−0.079, +0.090]	0.882
Invasive vs medical	M5 − M1	+0.007	[−0.069, +0.089]	0.866
Invasive vs medical	M5 − M2	+0.011	[−0.065, +0.094]	0.778
3-class macro	M2 − M1	+0.087	[+0.043, +0.128]	<0.001
3-class macro	M3 − M1	+0.141	[+0.099, +0.182]	<0.001
3-class macro	M4 − M1	+0.175	[+0.134, +0.214]	<0.001
3-class macro	**M5 − M1**	**+0.184**	**[+0.142, +0.225]**	**0.001**
3-class macro	M5 − M2	+0.096	[+0.056, +0.135]	<0.001

All four-task by non-trivial-comparison treatment-panel deltas with $|\Delta \text{ AUROC}| \geq 0.05$ remain significant after Benjamini–Hochberg adjustment.

Note on the invasive-versus-medical boundary task. None of the four imaging-augmented model contrasts reaches significance at FDR 0.05. The M3-versus-M1 comparison is borderline at uncorrected $p = 0.060$. This is consistent with the clinical interpretation in the main Discussion. The choice between any invasive treatment (PCI or CABG) and conservative medical therapy is dominated by patient frailty and comorbidities, which are captured by the clinical-only feature stack but not by the angiographic embedding. Imaging features neither confound nor break the signal and simply do not add information for this particular boundary task. The CABG-selection task, the principal main-text claim, shows the opposite, expected direction with very large effect sizes (Δ M5 − M1 = +0.294, $q < 0.001$).

S5. Pre-specified subgroup analyses

Subgroup analyses were pre-specified by clinical presentation (stable angina, unstable angina, NSTEMI, STEMI) and age stratum (< 60, 60 to 75, ≥ 75 years) before the main-text evaluation. We report Δ AUROC(M5 $-$ M1) for the CABG-versus-non-CABG task, with 95% confidence intervals from a 1,000-iteration paired bootstrap stratified by subgroup. The STEMI presentation stratum on the held-out test cohort has too few CABG events ($n = 24$, CABG $= 6$) to support a stable subgroup AUROC and is reported but not included in the headline forest.

S5.1 AMC CABG-versus-non-CABG: Δ AUROC(M5 $-$ M1) by subgroup

The imaging contribution to the CABG-versus-non-CABG decision was positive in every pre-specified subgroup (Supplementary Table 8). The clinical-only baseline M1 stayed near chance across subgroups (AUROC 0.55 to 0.63), whereas the full M5 stack reached 0.83 to 0.93. The resulting imaging gain Δ AUROC(M5 $-$ M1) ranged from +0.196 in the NSTEMI stratum to +0.369 in unstable angina, with every reported 95% confidence interval excluding zero. The gain is broadly distributed across clinical presentations and age strata rather than driven by any single subgroup.

Supplementary Table 8 Subgroup analysis of the imaging contribution to the AMC CABG-versus-non-CABG task.

Subgroup	n	n_{CABG}	Δ **AUROC (M5 $-$ M1)**	95% CI
Stable angina	601	105	**+0.350**	[+0.286, +0.416]
Unstable angina	293	88	**+0.369**	[+0.290, +0.449]
NSTEMI	64	19	+0.196	[+0.039, +0.352]
STEMI	24	6	n/a (too few events)	—
Age < 60	425	76	+0.291	[+0.218, +0.366]
Age 60 to 75	657	160	**+0.309**	[+0.250, +0.367]
Age ≥ 75	103	24	+0.280	[+0.136, +0.436]

The imaging gain (M5 over clinical-only M1) is positive with a 95% confidence interval excluding zero in every reported subgroup. Per-subgroup M1 and M5 AUROC ranges are summarised in the text. The STEMI stratum has too few CABG events ($n_{\text{CABG}} = 6$) for a stable estimate and is reported but excluded from the headline analysis. GAH subgroup detail (age strata only, presentation missing from the export) is reported in §S12.6.

S6. Treatment-decision supplementary analyses (AMC held-out test)

In addition to discrimination (AUROC), we report calibration (Brier score and expected calibration error), net benefit at clinically relevant thresholds, and reclassification metrics (continuous net reclassification improvement and integrated discrimination improvement) for the central CABG-versus-non-CABG task on the AMC held-out test cohort ($n_{\text{test}} = 1{,}220$, CABG $= 265$).

S6.1 Calibration

Supplementary Table 9 reports calibration metrics for M1–M5. The FMCA-augmented models achieve the lowest Brier scores, and the expected calibration error remains small (≤ 0.035) for every model.

Supplementary Table 9 Calibration metrics (Brier, expected calibration error, maximum calibration error) for the M1–M5 stack on the AMC held-out test cohort.

Model	Brier (↓)	ECE (↓)	MCE (↓)
M1 (clinical)	0.161	0.032	0.086
M2 (clinical + SYNTAX)	0.139	0.028	0.064
M3 (FMCA only, video, PCA-64)	0.104	0.028	0.053
M4 (clinical + FMCA)	0.103	0.027	0.075
M5 (clinical + SYNTAX + FMCA)	**0.098**	0.035	0.079

Adding FMCA features both improves discrimination and reduces Brier loss. Expected calibration error is small for every model on this 1,220-patient test cohort.

S6.2 Decision-curve analysis and reclassification

Decision-curve analysis. Net benefit was computed across decision thresholds $\theta \in [0.01, 0.99]$ for treating-as-CABG-candidate at each predicted-probability threshold. M5 dominates M2 and M1 across the clinically meaningful threshold range (10 to 60% predicted CABG probability). Net-benefit curves are rendered in Supplementary Figure 1.

Continuous NRI and IDI. Continuous net reclassification improvement and integrated discrimination improvement (Supplementary Table 10) follow the Pencina convention. Component decomposition is reported so that the unsigned sum is interpretable rather than a misleading aggregate. Confidence intervals are a 1,000-iteration patient-level bootstrap[19] on the fixed held-out test predictions.

Supplementary Table 10 Continuous NRI and IDI on the AMC held-out test cohort.

Contrast	$\text{NRI}_{\text{event}}$	$\text{NRI}_{\text{non-event}}$	Continuous NRI [95% CI]	IDI [95% CI]
M5 vs M2 (FMCA-incremental)	+0.562	+0.552	**+1.114 [+0.999, +1.227]**	**+0.224 [+0.197, +0.252]**
M5 vs M1 (combined)	+0.611	+0.631	+1.242 [+1.139, +1.346]	+0.339 [+0.307, +0.371]

Both NRI components are positive and of similar magnitude on the M5-versus-M2 contrast, so the continuous-NRI sum is interpretable. The categorical NRI for M5 versus M2 at the clinical CABG-probability cutoff of 0.40 is +0.36, consistent with the continuous result.

S6.3 Per-model regularisation-tuned sensitivity

The main manuscript reports a fixed L2-regularised logistic regression with $C = 0.01$ across M1–M5. To verify that the fixed-C choice does not unfairly under-fit the lower-dimensional baselines, we re-ran the analysis under a per-model C-tuning protocol (cross-validation-selected $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$) for CABG versus non-CABG (Supplementary Table 11).

Supplementary Table 11 Per-model regularisation-tuned sensitivity for the CABG-versus-non-CABG task.

Model	Selected C	AUROC (per-model C)	AUROC (fixed $C = 0.01$)	Δ to main
M1 (clinical)	10^{-3}	0.615	0.613	+0.002
M2 (clinical + SYNTAX)	10^{-1}	0.755	0.755	∼0
M3 (FMCA only)	10^{-1}	0.894	0.893	+0.001
M4 (clinical + FMCA)	10^{-1}	0.898	0.898	∼0
M5 (full)	10^{-1}	**0.908**	**0.907**	+0.001

AUROCs are shown as point estimates, and the 95% confidence intervals coincide closely between the two protocols (the fixed-C intervals are in Table 4 of the main manuscript). Δ AUROC(M5 − M1) under per-model C is +0.293 (vs main fixed $C = 0.01$: +0.294), so the two protocols agree to within 0.002 AUROC. Notably, the cross-validation-selected C for M1 is stronger than the fixed value (10^{-3} vs 10^{-2}), indicating that the M1 baseline is not over-regularised by the main protocol but, if anything, slightly under-regularised. The 16-dimensional clinical feature stack prefers stronger L2 shrinkage than the more informative PCA-64 imaging features. The headline Δ AUROC(M5 − M1) is therefore not a mechanical artefact of the C choice.

S6.4 SYNTAX-imputation transparency and SYNTAX-known-subset sensitivity

The institutional SYNTAX score is documented for 4,614 of 12,186 patients with a definitive treatment label (37.9%). Missing values were imputed by the median of the train and validation pool (median = 18.0) at the start of the M2, M4, and M5 classifier pipelines. The distribution of documented SYNTAX by treatment received is shown in Supplementary Table 12.

Supplementary Table 12 Among the 1,220 held-out test patients, distribution of documented SYNTAX score by treatment received.

Population	n	CABG	Documented SYNTAX	Median SYNTAX (documented)
Full held-out test (treatment-known)	1,220	265	460 (37.7%)	19.0
Treatment = CABG	265	265	135 (50.9%)	26.5
Treatment ≠ CABG	955	0	325 (34.0%)	16.0

The 50.9% versus 34.0% gap reflects the operator's higher likelihood of recording a SYNTAX score when CABG is a candidate treatment. SYNTAX missingness is therefore not missing-at-random. It is correlated with the treatment-decision label, and median imputation systematically attenuates the standalone discriminative value of the M2 baseline.

We re-ran the CABG-versus-non-CABG evaluation restricted to the 460 held-out test patients with a documented SYNTAX score. Classifiers were fit on the same train and validation pool. Only the test population is restricted. The results are reported in Supplementary Tables 13 and 14.

Supplementary Table 13 SYNTAX-known-subset sensitivity: AUROC on the 460 held-out test patients with a documented institutional SYNTAX score.

Model	AUROC, full test (n = 1,220)	AUROC, SYNTAX-documented test (n = 460)	Δ
M1 (clinical)	0.613	0.600	−0.013
M2 (clinical + SYNTAX)	**0.755**	**0.845**	**+0.090**
M3 (FMCA only)	0.893	0.913	+0.020
M4 (clinical + FMCA)	0.898	0.913	+0.015
M5 (clinical + SYNTAX + FMCA)	**0.907**	**0.931**	**+0.024**

Supplementary Table 14 Paired Δ AUROCs on the SYNTAX-documented subset (1,000-iteration paired bootstrap).

Comparison	Δ AUROC	95% CI	p
M3 vs M2 (FMCA-only vs clinical+SYNTAX)	**+0.068**	[+0.026, +0.116]	<0.001
M5 vs M2	**+0.085**	[+0.041, +0.131]	< 0.001
M5 vs M1	+0.331	[+0.263, +0.397]	< 0.001

Even when the SYNTAX comparator is at its strongest (real, documented score for every test patient), video-only FMCA (M3 = 0.913) exceeds the clinical-plus-SYNTAX baseline (M2 = 0.845) by +0.068 AUROC. The qualitative main-text claim that imaging embeddings exceed expert clinical-plus-SYNTAX holds on the SYNTAX-documented subset and is therefore not driven by imputation.

S7. Hyperparameters and software

The bootstrap, BH-FDR, and TRIPOD+AI conventions are described in the main Methods. This section reports only the hyperparameter grids and software versions used.

Pretraining. AdamW[18] (learning rate 2×10^{-4}, cosine schedule with 500-step warm-up, weight decay 0.01), LoRA $r = 32$, $\alpha = 64$, dropout 0.05, batch size 8 (4 per GPU × 2 GPUs), bfloat16 mixed precision, on two NVIDIA RTX PRO 6000 GPUs. Validation contrastive loss reached its minimum at epoch 2 (eval loss = 0.507). Training was stopped during epoch 6 after a three-epoch plateau, and the epoch-2 checkpoint was adopted.

Downstream classifier hyperparameter search. Logistic regression ($C \in \{0.01, 0.1, 1.0\}$, penalty $\in$ {L1, L2}), multi-layer perceptron (hidden units $\in \{(64), (128), (256), (256,128), (512,256)\}$, $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}\}$), gradient-boosted

decision trees (`n_estimators` $\in \{200, 400\}$, `max_depth` $\in \{3, 4\}$, learning rate $\in \{0.05, 0.10\}$).

Treatment-decision classifier. Fixed L2 logistic regression, $C = 0.01$, solver `lbfgs`. Pre-specified before evaluation, with no tuning on the held-out test.

Software. Python 3.12, scikit-learn 1.7[22], XGBoost 3.0[23], PyTorch 2.11[21]. The random seed was fixed at 42 throughout.

S8. Baseline characteristics and patient flow

S8.1 Baseline characteristics

The full split-stratified baseline characteristics table (train, validation, and test, with Benjamini–Hochberg-adjusted p-values) is presented as Extended Data Table 1 of the main manuscript, and the main-text Table 1 reports the overall-cohort summary. Variable-level missingness across the cohort is summarised in Supplementary Table 15.

Supplementary Table 15 Variable-level missingness in the AMC cohort.

Variable	n missing	Missing (%)
Age	425	3.4
Body-mass index	6	0.05
LV ejection fraction	2,708	21.9
Serum creatinine	1,144	9.3
Institutional SYNTAX score	7,729	62.6

Denominator is the full AMC cohort (n = 12,343). All variables not listed are complete. The patient-level split was stratified and balanced across variables (Extended Data Table 1).

S8.2 Patient-flow diagram

The patient-flow diagram for the AMC cohort is rendered as Supplementary Figure 2 (in the Supplementary Figures section at the end of this document). From 12,400 patient records with at least one CAG study and a structured Korean report, 12,344 had cines successfully extracted, 12,343 met the requirement of at least 5 usable frames per cine, and the patient-level stratified split (seed 42, treatment by clinical presentation strata) yielded 10,490 training, 618 validation, and 1,235 held-out test patients.

Per-task analytic cohorts further restrict to patients with a non-missing label. For example, the SYNTAX-score regression uses the 4,614 patients with a documented score. The treatment-decision benchmark uses the 12,186 patients with a definitive PCI, CABG, or medical-therapy label. The 1,235 held-out test patients never appeared in any pretraining step.

S9. Recommendation-text scrub corpus disclosure

The vessel-structured paraphrase used for contrastive pretraining contained, in its `common` field, both the overall angiographic finding (such as "N-vessel disease") and the Heart-Team's treatment recommendation (occasional explicit "Rec) ... CABG / PCI" strings). This is a label-shortcut concern for the treatment-decision results.

We addressed the concern with a targeted recommendation-text scrub before contrastive pretraining. A strong-form ablation that removes the entire `common` field is reported as a sensitivity bound.

S9.1 Targeted recommendation-text scrub (used for the main-text encoder)

Rules. (i) Cut everything from the first `Recommendation:` header onwards (case-insensitive). This single rule removes 73% of treatment-recommendation leakage. (ii) Drop any remaining sentence containing explicit treatment-decision tokens (`Heart Team`, `Rec)`, `OMT`, `to be decided`, `scheduled for PCI/CABG`, `planned for PCI/CABG`, `medical treatment`, `medical therapy`). (iii) Preserve every sentence about catheter equipment, access route, overall angiographic finding ("N-vessel disease"), FFR, iFR, IVUS, OCT, TIMI flow, collateral notes, patent stents, and prior segment-level interventions. The `LCA` and `RCA` per-vessel anatomic fields are not modified.

After the scrub, on the 12,344-record corpus (78.5% of records had at least one sentence removed), the residual occurrences of treatment-decision tokens in the `common` field are: `Recommendation` 16, `Heart Team` 0, `Rec)` 0, `OMT` 0, `medical treatment` 0, `medical therapy` 0. The mean `common` length dropped from 316 to 281 characters (88.8% retained). The 16 residual `Recommendation` mentions are negative phrasings (14 of 16: "Recommendation not stated"), TAVI work-up context (1 of 16), or prior-PCI-at-another-hospital (1 of 16). Zero of 16 maps to a CABG, PCI, or medical-therapy recommendation on the index hospitalisation.

The main-text encoder used throughout the main manuscript was trained on this targeted-scrub corpus. The validation contrastive loss reached its minimum at epoch 2 (eval loss = 0.507). Every main-text result is computed from this checkpoint.

S9.2 Strong-form ablation (sensitivity)

A strong-form ablation that removed the entire `common` field failed to converge. Validation contrastive loss saturated at 1.10 to 1.20 versus 0.507 for the targeted scrub, because the strong scrub also removed legitimate procedural and overall-finding text required to align video and report. The targeted scrub adopted in §S9.1 is therefore the correct compromise.

S10. Pre-specified external-validation interpretation rule

To protect the GAH external-validation claim from post-hoc rationalisation, an interpretation rule was specified before the GAH analysis was run. The rule defines how the GAH-transferred Δ AUROCs are translated into a verdict about external support.

S10.1 The rule (agreed before analysis)

- **STRONG support.** GAH-transferred Δ AUROC(M3 vs M2) ≥ +0.05 and Δ AUROC(M5 vs M2) ≥ +0.05, with both 95% CIs strictly above zero.
- **MODERATE support.** At least one of the above is significant. The other has a CI lower bound above zero but a smaller point estimate.
- **LIMITED support.** The magnitude attenuates substantially or one CI crosses zero. The manuscript's headline claim is qualified accordingly.
- **FAILURE.** Both deltas are negative with CIs strictly below zero. The classifier and encoder are frozen, so this would imply a substantial AMC-versus-GAH site shift, and the manuscript would be re-titled to a single-institution claim.

S10.2 Applied result

From §S12.2: Δ AUROC(M3 − M2) = +0.201 [+0.121, +0.281] and Δ AUROC(M5 − M2) = +0.189 [+0.109, +0.265]. Both deltas exceed +0.05 and both 95% CIs strictly exclude zero. **Outcome: STRONG support** for the main treatment-decision claim. Both GAH deltas are larger than the AMC held-out values. This is consistent with the GAH M1 and M2 baselines being weaker on the imputed-feature cohort (AUROC 0.55 versus 0.62 on AMC), which expands the gap that FMCA imaging closes. The internal AMC claim remains the more conservative reference point.

S11. SYNTAX comparator provenance

The institutional SYNTAX score used as the comparator throughout the main manuscript is not core-laboratory adjudicated. It is the SYNTAX score recorded prospectively in the institutional catheterisation-laboratory database, calculated by the primary operator at the time of the case. Because the database does not retain the operator identifier as a structured column, we recovered operator information directly from the report text by parsing the standardised header line ("*Operator 1: <name>*") present in every report.

S11.1 Operator-count statistics

Supplementary Table 16 summarises the distribution of primary operators recovered from the report headers.

The institutional SYNTAX comparator is therefore multi-reader with a substantial number of independent scorers (≥ 12 cardiologists each contributing ≥ 100

Supplementary Table 16 AMC primary-operator distribution recovered from the structured report header.

Statistic	Value
Reports parsed	12,400
Reports with retrievable Operator-1 field	9,874
Distinct primary operators	**20**
Operators contributing ≥ 100 cases	12
Cumulative share, top 5 operators	75.7%
Cumulative share, top 10 operators	96.3%
Cumulative share, top 12 operators	99.4%

cases). It is not core-laboratory adjudicated. This composition matches typical real-world institutional SYNTAX scoring and likely under-estimates the predictive value of a research-grade SYNTAX comparator obtained under blinded core-laboratory review[6].

S11.2 Implications for the comparator claim

The principal claim that FMCA imaging embeddings exceed clinical-plus-SYNTAX prediction is therefore evaluated against a real-world institutional comparator rather than a research-grade core-laboratory score. The SYNTAX-documented-subset sensitivity analysis (§S6.4), in which M3 (AUROC 0.913) still exceeds M2 (0.845) on patients with a complete documented SYNTAX score, indicates that the advantage persists even when the comparator is evaluated at its strongest.

S12. Gangneung Asan Hospital external-validation analysis

Methods are described in the main manuscript.

S12.1 Cohort summary

Supplementary Table 17 summarises the Gangneung Asan Hospital external-validation cohort and the clinical fields available in its electronic export.

S12.2 CABG-versus-non-CABG: AMC-trained classifier transferred to GAH (zero target-domain supervision)

Supplementary Table 18 reports the AMC-trained M1–M5 classifiers applied to the GAH cohort without any target-domain supervision (deployment-as-is), and Supplementary Table 19 the corresponding paired ΔAUROCs.

Paired Δ AUROC (patient-level paired bootstrap, 1,000 iterations):

The M3, M4, and M5 differences are within bootstrap noise on GAH transfer. Video-only M3 is statistically indistinguishable from the full M5 stack ($\Delta = +0.012$, $p = 0.14$). On the AMC held-out test, by contrast, M5 does slightly exceed M3 (Δ M3 $-$ M5 $= -0.014$ $[-0.021, -0.007]$, $p < 0.001$).

Supplementary Table 17 Gangneung Asan Hospital cohort summary.

Variable	Value
n total (definitive revascularisation label)	2,234
Age, median (years)	67
Male sex, n (%)	1,633 (73.1%)
Treatment: PCI / CABG, n (%)	2,143 (95.9%) / 91 (4.1%)
Medication-only stratum	not represented in the GAH revascularisation registry
Documented institutional SYNTAX score, n (%)	0 (0%)
Ejection fraction, median (%)	59
Creatinine, median (mg/dL)	1.03
Diabetes mellitus, n (%)	751 (33.6%)
CAG date range	2003 to 2023

The GAH electronic export contained age, sex, ejection fraction, creatinine, and diabetes mellitus but did not include institutional SYNTAX, presentation, body-mass index, smoking, hyperlipidemia, prior MI / stroke / PCI / CABG, atrial fibrillation, COPD, or chronic kidney disease. The remaining M1 features and the SYNTAX score are filled by deployment-as-is median imputation from the AMC train and validation pool (§S6.4). M1 (clinical) and M2 (clinical + SYNTAX) therefore degenerate to the same vector on this cohort and produce numerically identical AUROC.

Supplementary Table 18 GAH transfer M1–M5 AUROCs (zero target-domain supervision).

Model	AUROC	95% CI
M1 (clinical)	0.545	[0.485, 0.608]
M2 (clinical + SYNTAX)	0.545	[0.480, 0.604]
M3 (FMCA, video only)	0.744	[0.694, 0.793]
M4 (clinical + FMCA)	0.735	[0.683, 0.785]
M5 (clinical + SYNTAX + FMCA)	0.732	[0.680, 0.783]

Supplementary Table 19 Paired Δ AUROCs and two-sided p-values on the GAH transfer cohort.

Comparison	Δ AUROC	95% CI	p (two-sided)
M2 − M1	−0.001	[−0.087, +0.093]	0.998
M3 − M1	+0.200	[+0.119, +0.277]	<0.001
M3 − M2	**+0.201**	**[+0.121, +0.281]**	**0.001**
M4 − M1	+0.190	[+0.109, +0.269]	<0.001
M5 − M1	+0.188	[+0.107, +0.268]	<0.001
M5 − M2	**+0.189**	**[+0.109, +0.265]**	**0.001**
M3 − M4	+0.009	[−0.007, +0.025]	0.310
M3 − M5	**+0.012**	**[−0.005, +0.029]**	**0.140**
M4 − M5	+0.003	[−0.0004, +0.007]	0.076

Under deployment-as-is, the embedding-bearing models inherit the AMC prior and over-predict CABG probability on the GAH cohort, owing to a prevalence shift between cohorts (CABG 4.1% in GAH versus 21.7% in the AMC train and validation pool). This calibration gap is recoverable by simple recalibration to the GAH base rate (§S12.5). Because AUROC and reclassification metrics are invariant under monotonic recalibration, the AUROC gains above reflect the model's intrinsic discriminative contribution rather than a calibration artefact.

S12.3 CABG-versus-non-CABG: site-internal 5-fold cross-validation on GAH

Five-fold stratified k-fold within GAH. We report mean ± SD across folds and a bootstrap 95% CI on the out-of-fold predictions (Supplementary Table 20).

Supplementary Table 20 GAH site-internal 5-fold cross-validation AUROCs.

Model	AUROC mean ± SD	95% CI
M1	0.625 ± 0.061	[0.551, 0.664]
M2	0.625 ± 0.061	[0.549, 0.670]
M3 (video only)	0.810 ± 0.048	[0.764, 0.853]
M4	0.815 ± 0.038	[0.769, 0.855]
M5	0.815 ± 0.038	[0.767, 0.856]

The transfer-versus-internal gap (M3: 0.744 vs 0.810; M5: 0.732 vs 0.815) is small, a moderate and expected loss from cross-site deployment without any target-domain refitting. Both numbers are well above the corresponding M1 and M2 baselines.

S12.4 PCI-versus-CABG and other treatment-decision tasks on GAH

The GAH revascularisation registry contains only PCI and CABG patients. There is no medical-therapy-only stratum. Consequently, the main-text revascularisation-strategy task at GAH is operationally CABG versus PCI rather than CABG versus PCI-plus-medical-therapy. The three-class macro and invasive-versus-medical comparisons defined in the main manuscript are not reportable on this cohort. The CABG-versus-non-CABG results in §S12.2 and §S12.3 are equivalent to PCI-versus-CABG on this cohort up to label inversion. Their AUROCs and 95% CIs are therefore identical to the M1–M5 numbers above.

S12.5 Reclassification on the GAH cohort and isotonic recalibration

Supplementary Table 21 reports continuous net reclassification improvement (NRI) and integrated discrimination improvement (IDI) for the FMCA-bearing models on the GAH transfer cohort.

A simple isotonic recalibration of M5 to GAH out-of-fold predictions (5-fold) reduces M5 ECE from 0.44 to 0.07 and Brier from 0.275 to 0.072 without altering AUROC (isotonic regression is strictly monotonic). This is reported as a sensitivity. The deployment-as-is metric is the more stringent test of generalisation.

S12.6 Subgroup analysis: Δ AUROC(M5 − M1) on GAH transfer

GAH clinical presentation was not included in the data export. Only age strata are reportable (Supplementary Table 22).

Supplementary Table 21 Continuous NRI and IDI on the GAH transfer cohort.

Comparison	Continuous NRI [95% CI]	IDI [95% CI]
M5 vs M2	+0.197 [+0.111, +0.266]	+0.162 [+0.122, +0.200]
M5 vs M1	+0.266 [+0.175, +0.335]	+0.161 [+0.120, +0.198]

Component decomposition (M5 vs M2): event-NRI = +0.934, non-event-NRI = −0.737. The strongly positive event component reflects M5 correctly raising the predicted CABG probability on actual CABG cases. The negative non-event component is the expected signature of the prevalence shift discussed in §S12.2. The continuous NRI and IDI remain strongly positive because the event-side gain dominates. IDI in particular is invariant to base-rate shift in a way that confirms a genuine improvement.

Supplementary Table 22 GAH age-stratum subgroup analysis of the imaging contribution.

Subgroup	n	CABG	Δ AUROC (M5 − M1)	95% CI
Age $<$ 60	644	16	+0.083	[−0.096, +0.249]
Age 60 to 75	976	52	+0.175	[+0.102, +0.244]
Age $\geq$ 75	614	23	+0.284	[+0.164, +0.405]

The imaging contribution is positive across all three age strata and is statistically distinguishable from zero in the two strata that contain the bulk of CABG cases. The under-60 stratum has the smallest CABG count (16) and the widest CI.

S13. TRIPOD+AI checklist

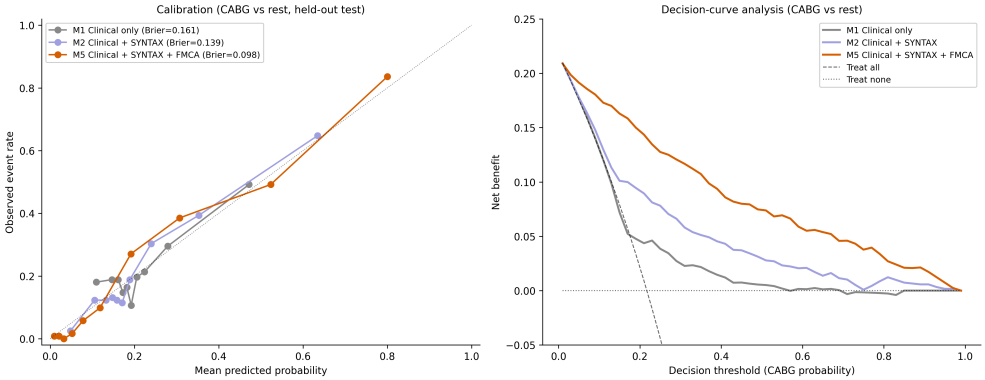
The manuscript is reported in accordance with the TRIPOD+AI guidance for clinical-prediction models that use regression or machine-learning methods[16]. The checklist below maps each TRIPOD+AI reporting domain to the manuscript or supplementary section in which the corresponding information is reported.

The checklist confirms reporting of:

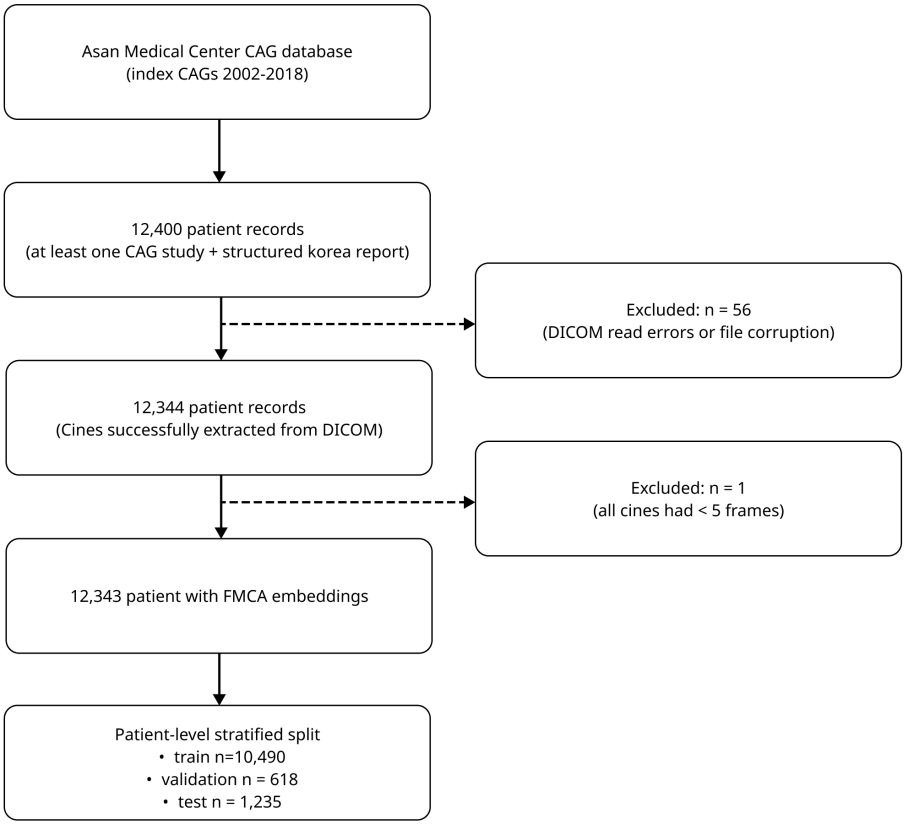
- **Title, abstract, and introduction.** Study objective, model intended use, structured abstract, and medical context.
- **Methods.** Source and eligibility of data (AMC 2002 to 2018; GAH cohort), outcome measurement, predictor construction (M1–M5), sample size and patient flow (§S8), missing-data handling (§S6.4 SYNTAX imputation), predictor selection (§S6.3 per-model C), performance measures, bias and calibration assessment (§S5 subgroups, §S6 calibration / DCA / NRI), internal versus external validation framing, software and hardware versions, and open-science availability (Data and Code availability statements in the main manuscript).
- **Results.** Participant flow (§S8.2 diagram), baseline characteristics (main Table 1, §S8.1), per-task performance, confidence intervals, and methodological refinements from preliminary work to the final reported pipeline (recommendation scrub in §S9, per-model regularisation in §S6.3, SYNTAX-known sensitivity in §S6.4).
- **Discussion.** Limitations enumerated in the main Discussion, overall interpretation, clinical implications, and generalisability (§S10 pre-specified external rule and §S12 GAH analysis).

Items 22 (clinical-implementation pathway) and 23 (regulatory pathway) are flagged as not applicable to the present research-stage publication. They would be relevant only for any future deployment.

Supplementary Figures



Supplementary Figure 1 Calibration and decision-curve analysis for the CABG-versus-non-CABG task on the AMC held-out test cohort ($n = 1{,}220$). Left: reliability curves for M1 (clinical only, Brier 0.161), M2 (clinical plus SYNTAX, Brier 0.139), and M5 (clinical plus SYNTAX plus FMCA, Brier 0.098), with perfect calibration on the diagonal; the FMCA-augmented M5 is the best calibrated. Right: decision-curve net benefit for M1, M2, and M5 across the CABG-probability threshold range, with "treat all" and "treat none" reference strategies. M5 provides the highest net benefit across the clinically relevant range.



Supplementary Figure 2 AMC patient-flow diagram. From 12,400 patient records with at least one CAG study and a structured Korean report, 12,344 had cines successfully extracted, 12,343 met the ≥ 5 usable frames-per-cine criterion, and the patient-level stratified split (seed 42, treatment $\times$ presentation strata) yielded 10,490 train / 618 validation / 1,235 held-out test patients. The 1,235 test patients were withheld from contrastive pretraining and all classifier training.