RETINAPACK: FEW-SHOT CONTINUAL RETINAL PRE-SCREENING WITH OFFLINE NEUROMORPHIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Retinal pre-screening in rural and resource-constrained settings requires portable, offline and privacy-preserving inference that can adapt to local acquisition shifts without large-scale retraining. A key difficulty is that current-stage labels can be obtained immediately, whereas prognostic outcomes mature months later, so local adaptation must improve staging without destabilizing longitudinal risk estimates. We introduce RetinaPack, a participant-disjoint longitudinal benchmark of 13,040 participants, and RetinaLPA-2S, a few-shot continual adaptation method for four-stage age-related macular degeneration (AMD) pre-screening. RetinaPack combines adjudicated change and stable pairs with crossed camera/operator repeats to separate disease-related variation from acquisition nuisance and define a pathology-referenced protected subspace. RetinaLPA-2S updates only its orthogonal complement through a 4,096-parameter two-sided low-rank adapter, while a frozen competing-risk branch preserves fixed-input prognosis during stage-only adaptation. A quantized encoder runs on the Akida AKD1000 and signed adaptation remains on the local host, enabling network-independent inference, local updating and feature replay without retinal-image storage. With 10 labels per stage, RetinaLPA-2S reaches 82.20 domain-balanced candidate F1 with 1.07-point forgetting, 1.68 points above BiLoRA-style FA. On 920 staging queries, it achieves 83.78 macro-F1 and 95.33% referral sensitivity. The complete image-to-report path requires 2.150 J, 650.4 ms P95 latency and 224 MiB peak memory, reducing processing energy by 34.35% relative to CPU INT8.

1 INTRODUCTION

Limited access to ophthalmic specialists remains a major barrier in rural, remote and resource-constrained communities, where distance and workforce shortages can delay eye care (Burton et al., 2021; Kiburg et al., 2022). Retinal pre-screening can help triage patients before specialist review, reserving limited clinical capacity for those who require further assessment (Dong et al., 2022; Kiburg et al., 2022). We study four-stage age-related macular degeneration (AMD) grading and six- and twelve-month late-AMD risk for specialist-supported referral, not autonomous diagnosis or treatment. This setting requires portable local inference under limited compute and power, without continuous network access, while keeping retinal images local and adapting to new cameras, operators and acquisition conditions from very few adjudicated labels. Because stage labels are available immediately whereas prognostic outcomes mature months later, local adaptation must learn the current domain without overwriting longitudinal disease information or changing prognosis without corresponding outcome evidence.

Existing approaches address these requirements only in part. Retinal foundation models provide strong transferable representations (Zhou et al., 2023), but full or repeated adaptation of large backbones is costly on portable hardware. Continual-learning and parameter-efficient methods reduce trainable state (Smith et al., 2023), yet their preservation criteria are usually defined by weights, logits, gradients, prompts or generic parameter geometry rather than clinically anchored longitudinal directions. Parameter count alone also does not determine adaptation time or memory cost (Mercea et al., 2024). Medical adaptation can correct an observed domain shift (Liu et al., 2025), but it optimizes the task supervised at update time, when future prognosis is unavailable. Neuromorphic conversion and quantization reduce inference cost (Hu et al., 2023), but efficient execution

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

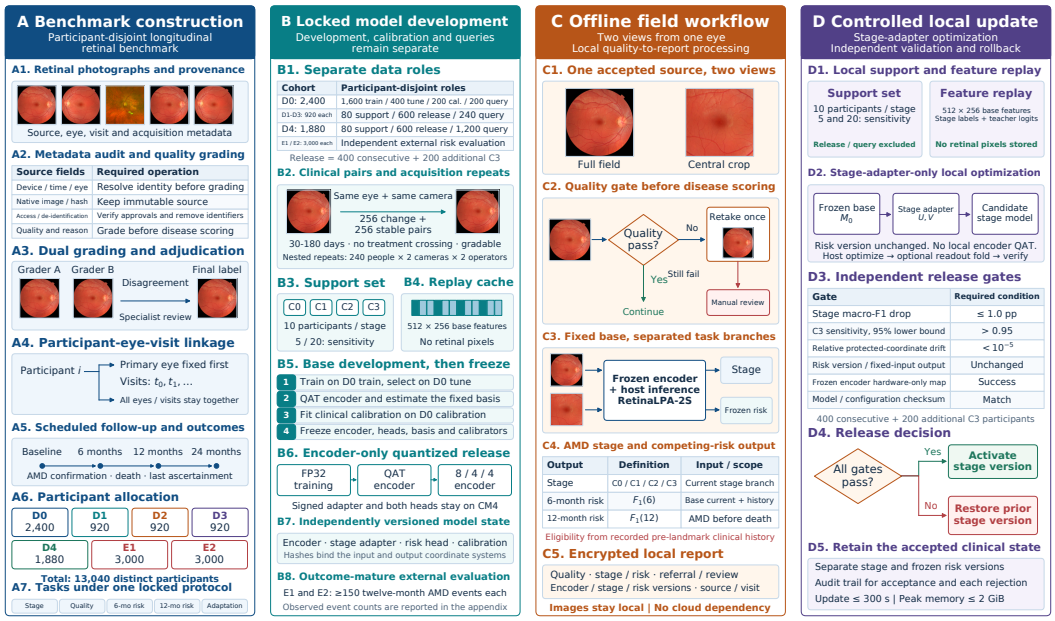

Figure 1: System pipeline across the four-panel study lifecycle. Support, release and query participants remain disjoint. Local updates change only staging. The risk model retains its version. Photographs illustrate input roles rather than study participants or observed longitudinal pairs.

alone does not specify what may be updated safely. The missing interface is a deployment setting that couples few-shot local adaptation, pathology-aware retention, delayed-risk isolation, pixel-free replay and measured device cost.

We address this missing interface with RetinaPack and RetinaLPA-2S. RetinaPack provides the longitudinal supervision and evaluation protocol required for continual retinal pre-screening. Across 13,040 participants, it separates development, support, release, query and external-risk roles, while 256 adjudicated change pairs, 256 stable pairs and a 240-participant crossed camera/operator substudy distinguish pathology change from acquisition variation. These matched sources define a pathology-referenced protected subspace that cross-sectional stage labels cannot identify, so the benchmark supplies both the clinical reference used by the method and the disjoint protocol needed to evaluate it. RetinaLPA-2S applies $A_s = I + QUV^\top Q$ to restrict few-shot updates to the complementary subspace with only 4,096 trainable scalars, while a frozen competing-risk branch isolates delayed prognosis from stage-only adaptation. The deployment pipeline maps the quantized encoder to the Akida AKD1000 and keeps signed adaptation and task heads on the local host, enabling offline inference and local updating with feature replay rather than retinal-image storage.

With 10 labelled participants per stage, RetinaLPA-2S reaches 82.20 domain-balanced candidate F1 with 1.07-point forgetting, exceeding BiLoRA-style FA by 1.68 points under the matched feature-adapter budget. On 920 staging queries, it obtains 83.78 macro-F1 and 95.33% referral sensitivity. The complete Akida-plus-host image-to-report path requires 2.150 J, 650.4 ms P95 latency and 224 MiB peak memory, reducing processing energy by 34.35% relative to CPU INT8. These results characterize the intended operating point of the system: few-shot local adaptation, offline execution and low processing cost while retaining competitive pre-screening performance.

Our main contributions are summarized as follows:

- **RetinaPack benchmark.** A participant-disjoint longitudinal benchmark that separates pathology change from acquisition variation and supports development, adaptation, release and external-risk evaluation under one protocol.
- **Pathology-referenced continual adaptation.** A two-sided low-rank update that adapts only outside a clinically referenced protected subspace, with a frozen prognostic branch isolated from stage-only updates.

- **Offline neuromorphic deployment.** An Akida-plus-host execution design with few-shot local updating, no network dependency and feature-level replay without retinal-image storage.

2 Related Work

RetinaPack sits at the intersection of retinal representation learning, continual and parameter-efficient adaptation, and deployable medical inference under tight resource constraints.

Retinal representation and longitudinal prediction. Retinal foundation models have substantially improved transfer from large-scale fundus and multimodal pretraining. RETFound learns a general retinal representation from unlabeled images (Zhou et al., 2023), RET-CLIP and FLAIR add clinical text supervision (Du et al., 2024; Silva-Rodríguez et al., 2025), and VisionFM, UrFound and RetiZero broaden task or disease coverage (Qiu et al., 2024; Yu et al., 2024b; Wang et al., 2025b). ADAM and related benchmarks provide well-defined AMD detection and lesion tasks (Fang et al., 2022). Longitudinal models use prior fundus images and clinical variables to predict AMD progression (Peng et al., 2020; Lee et al., 2022), while ORBIT-AMD extends the setting with ordinal, bilateral and trajectory supervision (Cui et al., 2026). These works establish strong representations and prognostic baselines, but their primary objective is prediction from a fixed training protocol. They do not define which representation coordinates should remain invariant when a deployed model receives new local stage labels, and cross-sectional accuracy cannot distinguish disease progression from camera or operator variation.

Continual learning and parameter-efficient adaptation. Classical continual learning controls forgetting through weight regularization, episodic replay, gradient projection and output preservation (Kirkpatrick et al., 2017; Li & Hoiem, 2018; Chaudhry et al., 2019; Aljundi et al., 2019; Buzzega et al., 2020; Saha et al., 2021). Pretrained-model approaches instead keep most backbone parameters fixed and adapt prompts or small modules, including L2P, CODA-Prompt, Consistent Prompting, SSIAT and mixture-of-experts adapters (Wang et al., 2022b; Smith et al., 2023; Gao et al., 2024; Tan et al., 2024; Yu et al., 2024a). Recent work studies selective, low-rank or adaptive PEFT in greater detail (Hu et al., 2022; Mercea et al., 2024; Zhang et al., 2024; Wang et al., 2025c; Tian et al., 2025; Zhu et al., 2025; Wang et al., 2025a; Mai et al., 2025). These methods improve the stability-plasticity trade-off and reduce trainable state, yet the constraint is generally defined by weights, logits, gradients, prompts or generic parameter geometry. Such restrictions do not establish that the preserved directions correspond to longitudinal pathology, and equally small adapters can preserve different information. This distinction is central in our setting because acquisition variation and disease change coexist in the same feature space.

Medical adaptation and deployable inference. Medical PEFT and continual adaptation address limited annotations and center-specific domain shift (Dutt et al., 2024; Liu et al., 2025). Test-time adaptation can refine medical representations efficiently, including latent updates that avoid full parametric optimization (Chen et al., 2025a), while recent buffer-based methods further control adaptation state (Kim et al., 2025). These approaches typically optimize the task observed at adaptation time. They do not separate current-task supervision from prognostic outcomes that are not yet observed. Efficient visual adaptation also shows that trainable-parameter count alone is not an adequate deployment metric because memory use and backpropagation cost can remain high (Mercea et al., 2024; Mai et al., 2025). Neuromorphic conversion and quantization reduce inference cost (Rueckauer et al., 2017; Li et al., 2022; Hu et al., 2023; Shen et al., 2025), but low-bit mapping can change signed linear constraints unless operator ownership is specified. The remaining challenge is to combine efficient stage adaptation with explicit longitudinal-coordinate protection and a separate evidence boundary for delayed prognosis.

3 RetinaPack Benchmark Construction

RetinaPack organizes 13,040 participants into participant-disjoint development, local-adaptation, release and external-risk roles (Table 1). All eyes, visits, retakes and derived views from one participant remain in the same partition. The primary staging snapshot contains 920 query participants from D0–D3, while D4 is a separate 1,200-participant held-out workflow cohort. E1 and E2 provide two independent 3,000-participant external-risk cohorts. Support and release-validation participants

Table 1: Participant allocation. Release sets are additional to, and disjoint from, support and query sets. Counts denote independent participants. E1/E2 are independent external-risk cohorts.

Cohort	People	Train	Tune	Cal.	Support	Release	Query
D0	2,400	1600	400	200	0	0	200
D1	920	0	0	0	80	600	240
D2	920	0	0	0	80	600	240
D3	920	0	0	0	80	600	240
D4	1,880	0	0	0	80	600	1200
E1	3,000	0	0	0	0	0	3000
E2	3,000	0	0	0	0	0	3000
Total	13,040	1,600	400	200	320	2,400	8,120

never enter the corresponding query metrics. This separation is required because local labels may train an adapter, release labels may accept or reject it, and neither can serve as test evidence.

Quality and longitudinal labels are defined independently of model output. Two graders record quality, drusen category, pigment abnormality and late-lesion evidence, with specialist adjudication for disagreement. D0 training supplies 256 adjudicated change pairs and 256 stable pairs from distinct participants. A nested 240-participant repeat substudy crosses two cameras and two operators within the same visit so acquisition variation can be estimated without treating it as disease progression. Each local domain reserves 80 support participants and an independent 600-person release stream, composed of 400 consecutive participants plus 200 additional C3 cases for sensitivity precision. D4 collection begins after protocol freeze. The two external systems each exceed the prespecified minimum of 150 observed twelve-month AMD events. Exact object definitions, acquisition metadata, visit timing, annotation rules, cohort flow and missingness accounting are reported in the appendix.

4 Problem Formulation

For participant i, eye e and time t, an accepted image $x_{i,e,t}$ is encoded as $z_t \in \mathbb{R}^{256}$ with reference stage $y_t \in \{0, 1, 2, 3\}$. Quality failure is a separate state. The primary update uses 10 labelled participants per stage, while 5 and 20 per stage are sensitivity settings. Query data are excluded from adaptation, normalization and model selection. Prognosis is the cumulative incidence $F_1(h)$ of first documented late AMD before death by $h \in \{6, 12\}$ months after a non-late-AMD landmark. Because these future labels are unavailable during immediate stage adaptation, encoder, stage adapter, risk head and calibration are versioned separately.

5 Methodology

The method separates acquisition quality, protected longitudinal coordinates and task-specific updates. All settings are fixed before query evaluation.

5.1 Quality and same-eye representation

Quality failures can confound pathology differences, so a four-block CPU network gates 128×128 RGB inputs before disease scoring. The blocks use 3×3 stride-2 convolutions with 8, 16, 32 and 64 channels, ReLU6, global average pooling and two logits. The D0 calibration threshold maximizes specificity subject to 98% failure sensitivity. One retake is allowed and a second failure withholds both stage and risk outputs.

Each accepted source produces a full-field view and a central half-diameter crop, both resized to 224×224. Both views share a compact encoder. Its frozen post-QAT features define common coordinates for longitudinal pairs, replay and prognosis. A stride-2 16-channel stem is followed by depthwise-separable blocks with 24, 32, 64, 96 and 128 output channels and strides 2, 2, 2, 2 and 1. Each block uses a linear 3×3 depthwise convolution, a 1×1 pointwise convolution, batch normalization and ReLU6. Global pooling yields 128 features per view and the concatenated base feature is $z \in \mathbb{R}^{256}$. The fused encoder contains 24,512 parameters and 18,555,712 convolutional MACs for the two views. The encoder remains frozen during all local stage updates.

5.2 Nuisance-referenced longitudinal anchoring

The protected information is estimated from three matched sources of within-eye variation rather than from adapter rank alone. Eligible longitudinal pairs span 30–180 days in the same eye and

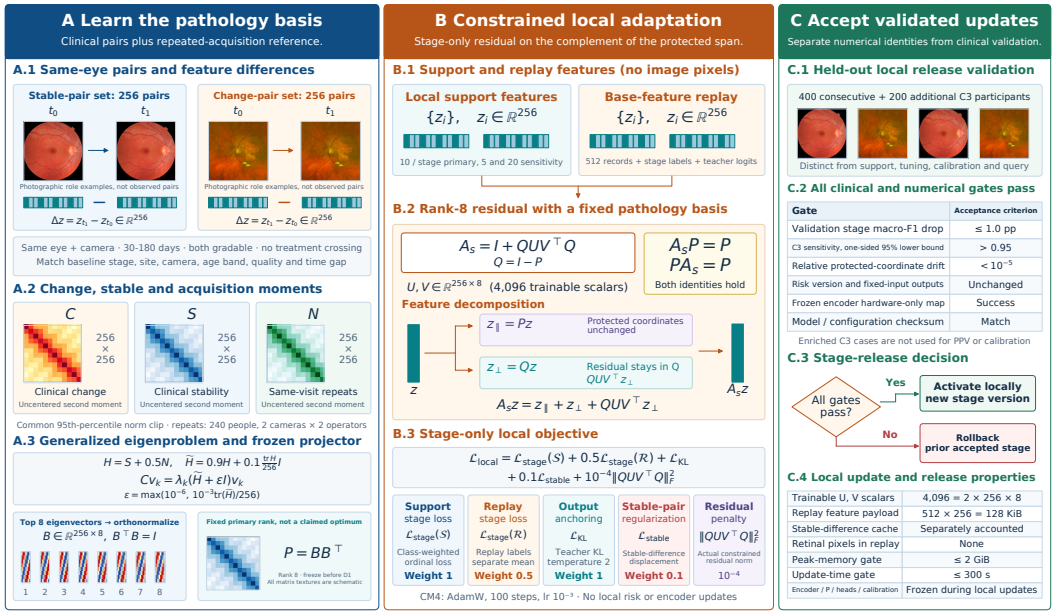

Figure 2: Method framework. Clinical change, stable and repeated-acquisition moments define the frozen projector. Two-sided projection restricts the stage residual at both ends. The local objective excludes risk and quantization-folding losses. Independent clinical and numerical gates control release. Matrix textures and photographic pairs are schematic.

camera, are gradable at both visits, and do not cross treatment. We use 256 adjudicated change pairs and 256 stable pairs. A nested 240-participant same-visit substudy crosses two cameras and two operators, providing a same-visit reference for acquisition variation. Before moment estimation, change and stable pairs are matched on baseline stage, camera, site, age band, quality and time gap. One shared 95th-percentile norm clip limits outlying norms while preserving difference directions.

Let C, S and N be the uncentered second moments of change, stable and repeat differences. Uncentered moments retain directional mean change that groupwise centering would remove. Each repeat participant contributes six contrasts weighted by $1/6$, giving participants equal total weight in N. The basis should emphasize lesion change relative to stable and repeated-acquisition variation. With $d = 256$ we define

$$H = S + 0.5N, \qquad \widetilde{H} = 0.9H + 0.1\frac{\operatorname{tr} H}{d} I, \qquad Cv_k = \lambda_k(\widetilde{H} + \epsilon I)v_k, \tag{1}$$

where $\epsilon = \max(10^{-6}, 10^{-3}\operatorname{tr}(\widetilde{H})/d)$. The top eight generalized eigenvectors are Euclidean-orthonormalized into $B \in \mathbb{R}^{256\times 8}$, giving $P = BB^\top$ and $Q = I - P$. Unlike a generic PCA, gradient or random span, this construction prefers directions with large adjudicated longitudinal change relative to stable and repeated-acquisition variation.

Local staging uses

$$A_s = I + QUV^\top Q, \qquad U, V \in \mathbb{R}^{256\times 8}, \qquad A_s P = P A_s = P, \qquad P(A_s z - z) = 0. \tag{2}$$

The two projections enforce distinct constraints. Right projection prevents the residual from acting on a protected input component, while left projection prevents complementary variation from being written back into protected coordinates. A one-sided map $I + UV^\top Q$ guarantees only the former. We initialize $U = 0$ and $V \sim \mathcal{N}(0, 0.01^2)$, which exposes 4,096 trainable scalars and starts from the identity map. The constraint is deliberately limited to the selected span. A required correction inside that span limits adaptation because the constraint excludes it. For a desired linear correction E, the best admissible rank-r residual has

$$\min_{\substack{R=QRQ \\ \operatorname{rank}(R)\le r}} \|E - R\|_F^2 = \|E - QEQ\|_F^2 + \sum_{j>r} \sigma_j(QEQ)^2. \tag{3}$$

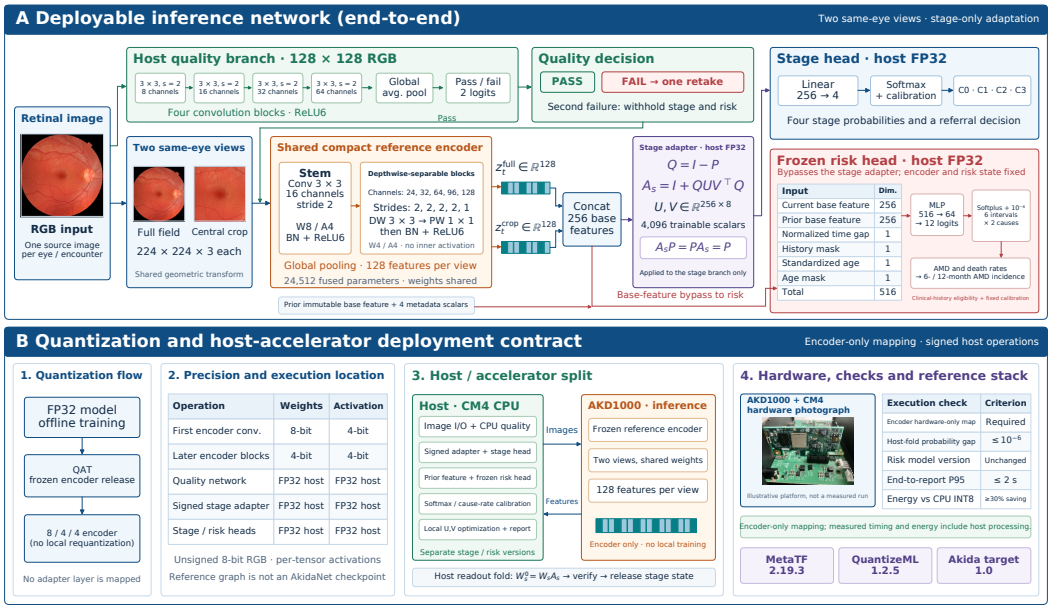

Figure 3: Network and physical execution. The explicit shared encoder is the only accelerator block. The red base-feature path bypasses stage adaptation and feeds the frozen $516 \to 64 \to 12$ competing-risk head. Quality, signed adaptation, readouts and calibration run on CM4. The photograph illustrates the platform, while processing measurements are reported separately.

The first term measures the correction excluded by protection and the second the residual rank cost. Matched basis and rank controls in Appendix P test whether the clinically selected span, rather than low rank alone, explains the observed adaptation behavior.

5.3 Stage Updates and a Frozen Competing-Risk Branch

With P frozen, only U and V change locally, preventing backbone drift. The stage head has four logits. For class counts n_c, effective-number weights use $\beta = 0.999$ and are normalized to mean one. Stage loss combines weighted cross-entropy with expected ordinal distance weighted by 0.15. For local support $\mathcal{S}$ and feature replay $\mathcal{R}$, the update minimizes

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{stage}}(\mathcal{S}) + 0.5\mathcal{L}_{\text{stage}}(\mathcal{R}) + \mathcal{L}_{\text{KL}} + 0.1\mathcal{L}_{\text{stable}} + 10^{-4}\|QUV^{\top}Q\|_F^2. \tag{4}$$

Support loss supplies plasticity, replay preserves earlier stage boundaries, and temperature-2 KL anchoring retains the previous accepted replay-output distribution. Support and replay losses are averaged separately, and distillation includes the squared-temperature multiplier. The stable-pair term constrains non-progressive variation separately from replay. It averages $\|(A_s - I)\delta\|_2^2/256$ over the 32 sampled stable differences δ. The residual penalty limits unnecessary motion inside the allowed complement. AdamW runs for 100 steps at 10^{-3} with weight decay 10^{-4}, gradient clipping at 1.0, batches of 16 support records, 64 replay records and 32 stable differences. The primary update uses 10 labelled participants per stage, with 5 and 20 per stage used only for sensitivity analysis. Replay stores 512 base features with stage labels and teacher logits, not retinal pixels.

Prognostic updates require observed follow-up outcomes. The risk branch bypasses A_s and receives the current and prior immutable base features, normalized time gap, history mask, standardized age and age mask, for 516 inputs. History uses the nearest eligible prior same-eye, same-camera capture within 180 days. Missing history is zero-filled and masked, while the gap is divided by 180 and clipped to $[0, 1]$, distinguishing unavailable history from an observed short interval. A $516 \to 64 \to 12$ MLP uses ReLU6 and training-only dropout 0.1. Softplus plus 10^{-6} yields AMD and death rates for six two-month intervals. For interval k, $\Lambda_{ik} = \lambda_{ik1} + \lambda_{ik2}$ and $d_k(t) = \text{clip}(t - 2(k-1), 0, 2)$, giving

$$F_{ij}(t) = \sum_{k=1}^{6} \exp\left[-\sum_{\ell<k} \Lambda_{i\ell} d_\ell(t)\right] \frac{\lambda_{ikj}}{\Lambda_{ik}} \left(1 - e^{-\Lambda_{ik} d_k(t)}\right). \tag{5}$$

For observed time $Y_i \le 12$ and cause $J_i \in \{0, 1, 2\}$, the cause-specific negative log-likelihood is

$$\mathcal{L}_{\text{risk}} = \frac{1}{n} \sum_i \left[\sum_k \Lambda_{ik} d_k(Y_i) - \mathbf{1}\{J_i > 0\} \log \lambda_{i,k(Y_i),J_i} \right]. \tag{6}$$

Death is treated as a competing event and partial follow-up contributes only observed exposure. The risk head, calibration and clinical-history eligibility remain fixed during stage adaptation, so updating U and V leaves fixed-input risk outputs unchanged.

5.4 Executable precision and release boundaries

The deployment boundary follows the same separation used by the learning rule. Only the frozen encoder is mapped to AKD1000 with unsigned 8-bit RGB input, 8-bit first-layer weights and 4-bit later weights and activations. Quality, the signed adapter, both task heads and calibration remain FP32 on CM4. Inference and host-side updates require no network connection. This avoids independently quantizing A_s, which would not preserve the two projection identities exactly. A stage readout may instead fold the affine map as $W'_s = W_s A_s$ on the host.

A candidate is frozen before release validation. The 400-person consecutive release stream measures stage performance, and an additional 200 C3 cases measure sensitivity precision without contributing to PPV or calibration. Release requires a stage macro-F1 drop no greater than 1.0 point, a one-sided 95% C3 sensitivity lower bound above 0.95, relative protected-coordinate drift below 10^{-5}, unchanged risk version and fixed-input risk output, successful encoder mapping, matching configuration checksum, update time at most 300 s and peak host memory at most 2 GiB. Any failed gate restores the prior accepted stage state. Candidate accuracy and retained-checkpoint performance are therefore distinct endpoints.

6 Experimental Setup

Staging uses the same 920 D0–D3 query participants across methods, while prognosis uses 6,000 E1/E2 participants. Continual adaptation covers all six D1–D3 domain orders and five seeds with nested 5, 10 and 20 labels per stage. Matched feature-adapter methods share the encoder, readout, support labels, replay budget and factor space. Foundation models use native resolutions with separate probe, PEFT and full-tuning regimes. Each tunable regime uses 24 D0-tuning trials and five-seed confirmation. Base training uses 80 AdamW epochs, batch 64, learning rate 3×10^{-4}, five warm-up epochs and cosine decay. QAT uses 20 epochs at 3×10^{-5}. Query labels are excluded from selection. Release labels only accept or reject frozen candidates.

Primary staging performance is macro-F1. Continual F1 averages the four query-domain F1 values, while forgetting averages peak-minus-final F1 over the base domain and the first two target domains. Participant-level uncertainty uses 1,000 paired bootstrap draws with shared resampled identities across methods. External cumulative/dynamic AUC and Brier scores treat death as a competing event and loss censoring with system-specific cross-fitted reverse Kaplan–Meier IPCW (Heagerty et al., 2000; Gerds & Schumacher, 2006). Calibration, within-C2 discrimination and net benefit accompany AUC. Full grids, uncertainty definitions and retained-state analyses are in the appendix.

7 Results

We first evaluate staging and continual adaptation, then examine external prognosis and deployment cost. Predictive performance, numerical invariance and release decisions are evaluated separately.

7.1 Four-stage staging and deployment comparison

RetinaLPA-2S reaches 83.78 macro-F1 versus 83.57 for the RetiZero probe (Table 2). The unrounded gain is 0.20 points (0.24% relative), but the paired interval does not establish superiority. At the calibration-locked threshold 0.33, referral sensitivity is 95.33%, distinct from 90.00% argmax C3 recall. Appendix H provides paired and threshold diagnostics, and Appendix R reports tuned baselines.

Table 2: A: staging comparisons on 920 shared query participants. F1 includes a participant-bootstrap 95% interval. C3 recall is argmax classification, not thresholded referral. B: measured execution under the same image-to-report processing boundary.

A. Representation	F1 [95% CI]	Acc.	κ_w	C3 recall	C1 F1
MobileNetV3-L (Howard et al., 2019)	78.95 [76.22, 81.60]	80.87	0.921	85.33	63.06
ResNet-50 (He et al., 2016)	77.54 [74.75, 80.22]	79.67	0.917	84.67	59.76
EfficientNet-B0 (Tan & Le, 2019)	79.36 [76.40, 81.90]	81.30	0.923	85.33	63.64
ConvNeXt-T (Liu et al., 2022)	80.18 [77.41, 82.63]	81.96	0.925	86.00	64.67
ConvNeXtV2-T (Woo et al., 2023)	81.38 [78.64, 83.82]	83.26	0.931	86.00	66.87
ViT-B/16 (Dosovitskiy et al., 2021)	80.51 [77.74, 83.01]	82.39	0.929	85.33	65.24
Swin-T (Liu et al., 2021)	81.31 [78.52, 83.82]	83.26	0.931	87.33	66.25
MaxViT-T (Tu et al., 2022)	82.17 [79.52, 84.78]	84.02	0.934	87.33	67.71
MAE ViT-B probe (He et al., 2022)	81.65 [78.92, 84.21]	83.48	0.932	87.33	66.25
DeepDrAMD-style (Xu et al., 2023)	81.49 [78.88, 84.00]	83.48	0.932	87.33	66.88
DINOv2 ViT-B (Oquab et al., 2024)	83.12 [80.53, 85.49]	84.78	0.937	87.33	70.55
MedCLIP probe (Wang et al., 2022a)	79.47 [76.66, 82.17]	81.52	0.924	84.67	63.58
RETFound probe (Zhou et al., 2023)	83.37 [80.67, 85.84]	85.00	0.939	88.67	69.33
RET-CLIP probe (Du et al., 2024)	82.29 [79.54, 84.68]	83.91	0.935	88.00	67.47
VisionFM probe (Qiu et al., 2024)	83.56 [81.00, 85.81]	85.33	0.939	88.00	69.84
FLAIR probe (Silva-Rodríguez et al., 2025)	82.55 [79.91, 84.95]	84.35	0.937	88.67	67.29
UrFound probe (Yu et al., 2024b)	83.35 [80.77, 85.77]	85.00	0.938	90.00	69.14
RetiZero probe (Wang et al., 2025b)	83.57 [80.98, 85.92]	85.33	0.941	88.00	69.81
RetinaLPA-2S	83.78 [81.23, 86.10]	85.43	0.941	90.00	69.16

B. Measured execution under the same processing boundary

Execution path	Median ms	P95 ms	Energy J	Peak MiB	Scope
CM4 FP32	1082.4	1265.0	5.010	768	whole path
CM4 INT8	679.0	787.3	3.275	432	whole path
AKD encoder + host heads	563.5	650.4	2.150	224	whole path

Table 3: Candidate-trajectory results across five seeds and all six domain orders. Dispersion is trajectory SD, not a participant-level confidence interval. FA denotes a matched feature-adapter reduction. Candidate performance is distinct from accepted-state deployment.

Adaptation rule	5/stage	10/stage ± SD	20/stage	Forget. pp	Runs
Frozen base	72.23	72.23 ± 1.77	72.23	0.00	30
Fine-tune	74.93	76.65 ± 1.92	77.43	6.29	30
LwF-FA (Li & Hoiem, 2018)	75.71	77.06 ± 1.76	77.58	2.66	30
EWC-FA (Kirkpatrick et al., 2017)	75.88	77.29 ± 1.68	77.83	2.35	30
A-GEM-FA (Chaudhry et al., 2019)	76.68	78.41 ± 1.63	79.08	2.51	30
ER-FA	76.92	78.76 ± 1.70	79.46	3.03	30
MIR-FA (Aljundi et al., 2019)	77.63	79.53 ± 1.63	80.32	2.17	30
DER++-FA (Buzzega et al., 2020)	78.24	80.33 ± 1.60	81.08	1.39	30
GPM-FA (Saha et al., 2021)	78.39	80.39 ± 1.69	81.15	1.03	30
BiLoRA-style FA (Zhu et al., 2025)	78.45	80.52 ± 1.72	81.22	1.13	30
Unconstrained rank-8 (Hu et al., 2022)	77.90	79.92 ± 1.55	80.68	1.94	30
One-sided LPA	79.05	81.25 ± 1.62	82.09	1.36	30
RetinaLPA-2S	79.93	82.20 ± 1.31	83.09	1.07	30

7.2 Few-shot domain-incremental adaptation

The two-sided adapter reaches 82.20 final domain-balanced F1, 1.68 points and 2.09% relative above BiLoRA-style FA (Table 3). Mean forgetting is 1.07 points. The one-sided control reaches 81.25 F1 with 1.36-point forgetting, while the frozen base avoids forgetting but remains at 72.23 F1. These results distinguish retention from the ability to learn new domains. Paired inference and separate gate-aware trajectory analyses are reported in Appendices U and S.

7.3 Longitudinal risk and uncertainty

The external cohorts contain 295 AMD events by six months and 553 by twelve months. The frozen compact risk branch achieves 0.781 twelve-month AUC versus 0.785 for the fully tuned temporal-foundation comparator (Table 4). The paired AUC-difference interval is [-0.0046, -0.0025], favoring the foundation comparator. Separate-system results and paired-risk interpretation are reported in Appendix V.

Table 4: External-risk comparison, $n = 6,000$. All methods share outcomes, loss censoring and landmarks. Death is a competing event. Risk outputs bypass local stage adaptation.

Risk comparator	AUC_6	AUC_{12}	BS_6	BS_{12}	ECE%	Slope
Age + stage	0.725	0.733	0.0475	0.0862	3.90	0.97
DeepSurv-style (Katzman et al., 2018)	0.732	0.755	0.0495	0.0863	2.59	0.71
Nnet-survival-style (Gensheimer & Narasimhan, 2019)	0.739	0.762	0.0484	0.0841	2.04	0.76
DeepHit-style (Lee et al., 2018)	0.741	0.763	0.0482	0.0839	1.97	0.77
Dynamic-DeepHit-style (Lee et al., 2020)	0.753	0.774	0.0470	0.0813	1.17	0.87
Current-image hazard	0.741	0.763	0.0479	0.0833	1.74	0.78
CNN-LSTM temporal (Lee et al., 2022)	0.757	0.777	0.0467	0.0807	0.94	0.90
CNN-Transformer temporal (Lee et al., 2022)	0.759	0.779	0.0466	0.0804	0.90	0.92
RETFound temporal PEFT (Zhou et al., 2023)	0.763	0.782	0.0463	0.0799	0.75	0.95
VisionFM temporal full (Qiu et al., 2024)	0.766	0.785	0.0461	0.0794	0.78	0.97
RetinaLPA-2S frozen risk	0.761	0.781	0.0464	0.0800	0.77	0.94

Table 5: A: selected configurations of the complete 16-cell factorial ablation. Bits denote projection/stable/replay/anchor, with all other factors fixed. B: measured latency and energy, including the transfer-stress condition.

A. Mechanism ablation	Bits	Final F1	SD	Forget. pp
Full two-sided rule	1111	82.20	1.31	1.07
No projection, other terms fixed	0111	80.21	1.65	1.55
No stable-pair term	1011	81.15	1.64	1.30
No replay CE	1101	79.31	1.65	2.75
No output anchor	1110	80.31	1.57	2.23
Projection only	1000	76.56	1.69	4.37
Replay and anchors only	0011	79.13	1.63	2.36
None of the four factors	0000	74.81	1.84	4.68
B. Measured deployment: latency and energy from execution traces				
Execution path	Median ms	P95 ms	Mean J	Peak MiB
CM4 FP32	1082.4	1265.0	5.010	768
CM4 INT8	679.0	787.3	3.275	432
AKD encoder + host heads	563.5	650.4	2.150	224
Transfer stress	727.9	840.2	2.925	240

7.4 Mechanism, numerical checks and deployment

Double-precision random-matrix tests verify both projection identities and the constrained-rank approximation formula. These tests establish algebraic properties rather than clinical semantics. Table 5 isolates four factors, with its full configuration identical to Table 3. The measured encoder-plus-host path uses 2.150 J versus 3.275 J for CPU INT8, a 34.35% reduction under the shared processing boundary. Its benefit is diluted by idle energy. In independent release validation, 192/200 C3 detections exceed 95% point sensitivity but fail the prespecified lower-confidence-bound gate. Extended numerical and hardware results are reported in Appendix W.

These results show a trade-off, not universal superiority. Two-sided anchoring improves final F1 over the matched one-sided and unconstrained controls, whereas GPM-FA has slightly lower forgetting. The paired staging interval does not establish superiority over RetiZero, and the temporal foundation model achieves higher external-risk AUC. The compact system combines stage-only adaptation, fixed-input risk invariance and lower processing energy. Fixed-input risk invariance alone does not establish calibration after a population shift. Protection is limited to the selected span. The stage head can still respond to complementary features, so coordinate invariance does not imply unchanged stage decisions. Shifts within that span can impede correction. Candidate accuracy and sensitivity-subgate results do not imply that every update passes all release gates.

8 Conclusion

We presented RetinaLPA-2S, a two-sided stage adapter with a frozen prognostic branch and encoder-only accelerator execution. The experiments evaluate local adaptation under a fixed parameter and replay budget, with stage updates separated from delayed prognosis. Numerical invariance, independent release gates and measured processing costs provide distinct evidence for offline adaptation without silently changing prognosis.

AI USE STATEMENT

Generative AI assisted only with manuscript wording, sentence-level expression and organizational polishing. It was not used for literature search, evidence collection, experimental design, result generation or scientific decision making. The authors conducted the experiments, supplied and verified the reported results, reviewed all AI-assisted edits and are responsible for the scientific claims and submitted artifacts.

ETHICS STATEMENT

The study protocol, use of participant-linked clinical records and retinal images, and all associated experimental procedures were reviewed and approved by the relevant institutional ethics committee before study conduct, with ethics oversight maintained throughout the project. All data were handled under the approved protocol and applicable institutional governance and privacy requirements. To preserve double-blind anonymity, the committee name and approval identifier are omitted from the review version and will be reported in the camera-ready manuscript. The intended use is specialist-supported pre-screening, not autonomous diagnosis or treatment. Quality failures require clinical review, and rejected local updates retain the prior accepted stage version. Public release of RetinaPack is subject to a separate institutional data-governance and release review. Any released records will be de-identified and distributed only under the access conditions permitted by the final approval.

REPRODUCIBILITY STATEMENT

The main paper and appendices specify the data partitions, preprocessing, architecture, losses, optimization settings, evaluation metrics and extended results. The core implementation is supplied separately as `source_code.zip`. Reproducing data-dependent results also requires the corresponding model artifacts and authorized clinical records. The Overleaf archive contains manuscript sources and figures rather than participant-linked data or the experimental training archive.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximally interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11849–11860, 2019.

Matthew J. Burton, Jacqueline Ramke, Ana Patricia Marques, et al. The lancet global health commission on global eye health: Vision beyond 2020. *The Lancet Global Health*, 9(4):e489–e551, 2021. doi: 10.1016/S2214-109X(20)30488-5.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930, 2020. URL https://arxiv.org/abs/2004.07211.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Kecheng Chen, Xinyu Luo, Tiexin Qin, Jie Liu, Hui Liu, Victor Ho Fun Lee, Hong Yan, and Haoliang Li. Test-time adaptation for foundation medical segmentation model without parametric updates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20075–20084, 2025a.

Qingyu Chen, Tiarnan D. L. Keenan, Elvira Agron, Alexis Allot, Emily Guan, Bryant Duong, Amr Elsawy, Benjamin Hou, Cancan Xue, Sanjeeb Bhandari, Geoffrey Broadhead, Chantal Cousineau-Krieger, Ellen Davis, William G. Gensheimer, Cyrus A. Golshani, David Grasic, Seema Gupta, Luis Haddock, Eleni Konstantinou, Tania Lamba, Michele Maiberger, Dimosthenis Mantopoulos, Mitul C. Mehta, Ayman G. Elnahry, Mutaz AL-Nawaflh, Arnold Oshinsky,

Brittany E. Powell, Boonkit Purt, Soo Shin, Hillary Stiefel, Alisa T. Thavikulwat, Keith James Wroblewski, Yih Chung Tham, Chui Ming Gemmy Cheung, Ching-Yu Cheng, Emily Y. Chew, Michelle R. Hribar, Michael F. Chiang, and Zhiyong Lu. AI workflow, external validation, and development in eye disease diagnosis. *JAMA Network Open*, 8(7):e2517204, 2025b. doi: 10.1001/jamanetworkopen.2025.17204.

Xuehao Cui, Dejia Wen, Patrick Yu-Wai-Man, and Xiaorong Li. ORBIT-AMD: Ordinal risk, bilateral imaging, and trajectory learning for age-related macular degeneration in multi-cohorts. *Advanced Science*, pp. e76361, 2026. doi: 10.1002/advs.76361.

Li Dong, Wanji He, Ruiheng Zhang, Zongyuan Ge, Ya Xing Wang, Jinqiong Zhou, Jie Xu, Lei Shao, Qian Wang, Yanni Yan, Ying Xie, Lijian Fang, Haiwei Wang, Yenan Wang, Xiaobo Zhu, Jinyuan Wang, Chuan Zhang, Heng Wang, Yining Wang, Rongtian Chen, Qianqian Wan, Jingyan Yang, Wenda Zhou, Heyan Li, Xuan Yao, Zhiwen Yang, Jianhao Xiong, Xin Wang, Yelin Huang, Yuzhong Chen, Zhaohui Wang, Ce Rong, Jianxiong Gao, Huiliang Zhang, Shouling Wu, Jost B. Jonas, and Wen Bin Wei. Artificial intelligence for screening of multiple retinal and optic nerve diseases. *JAMA Network Open*, 5(5):e229960, 2022. doi: 10.1001/jamanetworkopen.2022.9960.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jiawei Du, Jia Guo, Weihang Zhang, Shengzhu Yang, Hanruo Liu, Huiqi Li, and Ningli Wang. Retclip: A retinal image foundation model pre-trained with clinical diagnostic reports. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024*, volume 15012 of *Lecture Notes in Computer Science*, pp. 709–719. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72390-2_66.

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In *Proceedings of the 7th International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 406–425, 2024. URL https://proceedings.mlr.press/v250/dutt24a.html.

Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenole Quellec, Sarah Matta, Sharath M. Shankaranarayana, Yi-Ting Chen, Chuen-heng Wang, Nisarg A. Shah, Chia-Yen Lee, Chih-Chung Hsu, Hai Xie, Baiying Lei, Ujjwal Baid, Shubham Innani, Kang Dang, Wenxiu Shi, Ravi Kamble, Nitin Singhal, Ching-Wei Wang, Shih-Chang Lo, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, Yanwu Xu, and iChallenge-AMD study group. ADAM challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022. doi: 10.1109/TMI.2022.3172773.

Frederick L. Ferris, C. P. Wilkinson, Alan Bird, Usha Chakravarthy, Emily Chew, Karl Csaky, and Srinivas R. Sadda. Clinical classification of age-related macular degeneration. *Ophthalmology*, 120(4):844–851, 2013. doi: 10.1016/j.ophtha.2012.10.036. URL https://pubmed.ncbi.nlm.nih.gov/23332590/.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28463–28473, 2024.

Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. doi: 10.7717/peerj.6257.

Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006. doi: 10.1002/bimj.200610301.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000. doi: 10.1111/j.0006-341X.2000.00337.x.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yangfan Hu, Qian Zheng, Xudong Jiang, and Gang Pan. Fast-snn: Fast spiking neural network by converting quantized ann. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (12):14546–14562, 2023. doi: 10.1109/TPAMI.2023.3275769. URL https://doi.org/10.1109/TPAMI.2023.3275769.

Mrudula Jeeva and Santhosh Miriala. Performance evaluation of AkidaNet converted to spiking domain for the classification of weeds in cotton fields. *Scientific Reports*, 16:27266, 2026. doi: 10.1038/s41598-026-54153-4.

Jared L. Katzman, Uri Shaham, Jonathan Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:24, 2018. doi: 10.1186/s12874-018-0482-1.

Aditya Kelkar, Jai Kelkar, Yash Garg, Harsh H. Jain, and Sabyasachi Sengupta. Smartphone-based offline AI for multi-disease retinal screening: Real-world accuracy. *European Journal of Ophthalmology*, 2026. doi: 10.1177/11206721261462321.

Katerina V. Kiburg, Angus Turner, and Mingguang He. Telemedicine and delivery of ophthalmic care in rural and remote communities: Drawing from australian experience. *Clinical & Experimental Ophthalmology*, 50(7):793–800, 2022. doi: 10.1111/ceo.14147.

Hyeongyu Kim, Geonhui Han, and Dosik Hwang. Buffer layers for test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/ce7dfc19702cc3e2178f2d3cb5c87058-Paper-Conference.pdf.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 698–705, 2018. doi: 10.1609/aaai.v32i1.11842.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020. doi: 10.1109/TBME.2019.2909027.

Junghwan Lee, Tingyi Wanyan, Qingyu Chen, Tiarnan D. L. Keenan, Benjamin S. Glicksberg, Emily Y. Chew, Zhiyong Lu, Fei Wang, and Yifan Peng. Predicting age-related macular degeneration progression with longitudinal fundus images using deep learning. In *Machine Learning*

in Medical Imaging, volume 13583 of *Lecture Notes in Computer Science*, pp. 11–20. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-21014-3_2.

Chen Li, Lei Ma, and Steve Furber. Quantization framework for fast spiking neural networks. *Frontiers in Neuroscience*, 16:918793, 2022. doi: 10.3389/fnins.2022.918793. URL https://doi.org/10.3389/fnins.2022.918793.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Yihang Liu, Ying Wen, Longzhen Yang, Lianghua He, and Heng Tao Shen. CoSMIC: Continual self-supervised learning for multi-domain medical imaging via conditional mutual information maximization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23051–23062, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

Zhuang Liu, Hai Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Quang-Huy Nguyen, Li Zhang, and Wei-Lun Chao. Lessons and insights from a unifying study of parameter-efficient fine-tuning (peft) in visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14845–14857, 2025. doi: 10.1109/CVPR52734.2025.01383.

Otniel-Bogdan Mercea, Alexey Gritsenko, Cordelia Schmid, and Anurag Arnab. Time-, memory- and parameter-efficient visual adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5536–5545, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Yifan Peng, Tiarnan D. Keenan, Qingyu Chen, Elvira Agron, Alexis Allot, Wai T. Wong, Emily Y. Chew, and Zhiyong Lu. Predicting risk of late age-related macular degeneration using deep learning. *npj Digital Medicine*, 3:111, 2020. doi: 10.1038/s41746-020-00317-z. URL https://www.nature.com/articles/s41746-020-00317-z.

Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, Yuyang Zhao, Xuehui Shi, Junfang Xian, Xiaoxia Qu, Sirui Zhu, Lijie Pan, Xiaoniao Chen, Xiaojia Zhang, Shuai Jiang, Kebing Wang, Chenlong Yang, Mingqiang Chen, Sujie Fan, Jianhua Hu, Aiguo Lv, Hui Miao, Li Guo, Shujun Zhang, Cheng Pei, Xiaojuan Fan, Jianqin Lei, Ting Wei, Junguo Duan, Chun Liu, Xiaobo Xia, Siqi Xiong, Junhong Li, Kyle Lam, Benny Lo, Yih Chung Tham, Tien Yin Wong, Ningli Wang, and Wu Yuan. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12), 2024. doi: 10.1056/AIoa2300221.

Bodo Rueckauer, Ioana-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 11:682, 2017. doi: 10.3389/fnins.2017.00682. URL https://www.frontiersin.org/articles/10.3389/fnins.2017.00682/full.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3AOj0RCNC2.

Sicheng Shen, Dongcheng Zhao, Linghao Feng, Zeyang Yue, Jindong Li, Tenglong Li, Guobin Shen, and Yi Zeng. STEP: A unified spiking transformer evaluation platform for fair and reproducible benchmarking. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/e9fece10c0f4ffdcf015ab676f370fbd-Paper-Datasets_and_Benchmarks_Track.pdf.

Julio Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. doi: 10.1016/j.media.2024.103357.

Christos Skevas, Nicolás Pérez de Olaguer, Albert Lleó, David Thiwa, Ulrike Schroeter, Inês Valente Lopes, Luca Mautone, Stephan J. Linke, Martin S. Spitzer, David F. W. Yap, and Di Xiao. Implementing and evaluating a fully functional AI-enabled model for chronic eye disease screening in a real clinical environment. *BMC Ophthalmology*, 24:51, 2024. doi: 10.1186/s12886-024-03306-y.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023. doi: 10.1109/CVPR52729.2023.01146.

Jan-Martin O. Steitz and Stefan Roth. Adapters strike back. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23449–23459, 2024.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. Semantically-shifted incremental adapter-tuning is a continual vitransformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23252–23262, 2024.

Zichen Tian, Yaoyao Liu, and Qianru Sun. Meta-learning hyperparameters for parameter efficient fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23037–23047, 2025.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *Computer Vision – ECCV 2022*, volume 13684 of *Lecture Notes in Computer Science*, pp. 459–479. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20053-3_27.

Huiyi Wang, Haodong Lu, Lina Yao, and Dong Gong. Self-expansion of pre-trained models with mixture of adapters for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10087–10098, 2025a. doi: 10.1109/CVPR52734.2025.00943.

Meng Wang, Tian Lin, Aidi Lin, Kai Yu, Yuanyuan Peng, Lianyu Wang, Cheng Chen, Ke Zou, Huiyu Liang, Man Chen, et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nature Communications*, 16:5528, 2025b. doi: 10.1038/s41467-025-60577-9.

Qiang Wang, Xiang Song, Yuhang He, Jizhou Han, Chenhao Ding, Xinyuan Gao, and Yihong Gong. Boosting domain incremental learning: Selecting the optimal parameters is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4839–4849, 2025c.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022a. doi: 10.18653/v1/2022.emnlp-main.256.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b. doi: 10.1109/CVPR52688.2022.00024.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Ke Xu, Shenghai Huang, Zijian Yang, Yibo Zhang, Ye Fang, Gongwei Zheng, Bin Lin, Meng Zhou, and Jie Sun. Automatic detection and differential diagnosis of age-related macular degeneration from color fundus photographs using deep learning with hierarchical vision transformer. *Computers in Biology and Medicine*, 167:107616, 2023. doi: 10.1016/j.compbiomed.2023.107616.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024a.

Kai Yu, Yang Zhou, Yang Bai, Zhi Da Soh, Xinxing Xu, Rick Siow Mong Goh, Ching-Yu Cheng, and Yong Liu. Urfound: Towards universal retinal foundation models via knowledge-guided masked modeling. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024*, volume 15012, pp. 753–762. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72390-2_70.

Zhi Zhang, Qizhe Zhang, Zijun Gao, Renrui Zhang, Ekaterina Shutova, Shiji Zhou, and Shanghang Zhang. Gradient-based parameter selection for efficient fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28566–28577, 2024.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622:156–163, 2023. doi: 10.1038/s41586-023-06555-x. URL https://www.nature.com/articles/s41586-023-06555-x.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. BiLoRA: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25613–25622, 2025.

SUBMISSION MATERIALS, AI ASSISTANCE AND BENCHMARK RELEASE

AI assistance. Generative AI assisted only with manuscript wording, sentence-level expression, organizational polishing and LaTeX formatting. It was not used for literature search, evidence collection, experimental design, result generation or scientific decision making. The authors conducted the experiments, supplied and verified the reported results, reviewed all AI-assisted edits and are responsible for the scientific claims and submitted artifacts. Numerical verification remains distinct from clinical evaluation and physical-device measurement.

Code and data-dependent reproduction. An anonymized `source_code.zip` is supplied separately with the supplementary submission. It contains the core implementation for protected-basis estimation, two-sided adaptation, local optimization, competing-risk inference, evaluation and numerical checks. Data-dependent reproduction additionally requires authorized clinical records and the corresponding model artifacts. The Overleaf archive contains editable manuscript sources, the bibliography, official style files and figure assets. Compiling this archive does not rerun the experiments.

Benchmark release. The completed study and experimental procedures were conducted under prior institutional ethics approval. Public release of RetinaPack is a separate governance process and remains subject to the required institutional data-release and privacy review. The released benchmark will use de-identified participant records, documented split manifests, task definitions and evaluation protocols consistent with the approved study protocol and final release conditions. Where public redistribution is permitted, the benchmark will be made available without charge for research use. Components requiring controlled access will be released through the corresponding approved access mechanism rather than through an unrestricted download.

A ADDITIONAL RELATED WORK

Prior work establishes dataset documentation, retinal representation learning, low-rank adaptation, competing-risk likelihoods and quality-gated reporting. The methodological question is whether lesion-adjudicated, nuisance-referenced directions improve constrained adaptation beyond a generic span of equal dimension. Equations in the main paper establish a property of a linear map. The clinical value of the selected span requires comparison against the following alternatives.

Table 6: Literature positioning and comparison consequences.

Family	Primary sources	Established contribution	Required distinction
AMD benchmark	ADAM (Fang et al., 2022)	AMD classification and lesion/localization tasks.	Use the supported cross-sectional tasks; longitudinal endpoints require separate labels.
Retinal SSL/VLM	RETFound, RET-CLIP, FLAIR (Zhou et al., 2023; Du et al., 2024; Silva-Rodríguez et al., 2025)	Retinal image or clinical-text pretraining.	Compare probe, PEFT and full tuning with identical outcome access.
Generalist retinal models	VisionFM, UrFound, RetiZero (Qiu et al., 2024; Yu et al., 2024b; Wang et al., 2025b)	Multimodal or knowledge-rich representation.	Match the CFP-only input track and disclose unavailable pretrained data.
Longitudinal prediction	Peng, Lee, ORBIT-AMD (Peng et al., 2020; Lee et al., 2022; Cui et al., 2026)	Prior images, ordinal/bilateral/trajectory supervision.	Separate unilateral primary from bilateral secondary, and 12-month from other horizons.
Clinical workflow	RAIDS, Skevas, DeepSeeNet+ workflow (Dong et al., 2022; Skevas et al., 2024; Chen et al., 2025b)	Referral and external clinical evaluation already exist.	Workflow integration is distinct from the adaptation contribution.

Table 6 (continued)

Family	Primary sources	Established contribution	Required distinction
Offline screening	Smartphone offline screening (Kelkar et al., 2026)	Offline retinal inference is not unique.	Compare whole-path cost on specified physical platforms.
Generic CL	LwF, EWC, A-GEM, MIR, DER++, GPM (Li & Hoiem, 2018; Kirkpatrick et al., 2017; Chaudhry et al., 2019; Aljundi et al., 2019; Buzzega et al., 2020; Saha et al., 2021)	Retention through logits, weights, examples or gradients.	Hold label and feature information constant for attribution.
Pretrained-model CL	L2P, CODA-Prompt, SSIAT, MoE adapters (Wang et al., 2022b; Smith et al., 2023; Tan et al., 2024; Yu et al., 2024a)	Prompt/adapter isolation and routing.	Native-backbone approaches are distinct from the matched feature-adapter comparisons.
Recent PEFT	Adapter+, SOYO, MetaPEFT, BiLoRA (Steitz & Roth, 2024; Wang et al., 2025c; Tian et al., 2025; Zhu et al., 2025)	Improved adapter selection or subspace separation.	Equal rank alone does not mean equal information or compute.
Medical adaptation	Medical PEFT, CoSMIC, medical TTA (Dutt et al., 2024; Liu et al., 2025; Chen et al., 2025a)	Medical center shifts and adaptive representations.	Distinguish label-free TTA from supervised ten-per-stage updates.
SNN and quantization	Conversion, QFFS, Fast-SNN, STEP (Rueckauer et al., 2017; Li et al., 2022; Hu et al., 2023; Shen et al., 2025)	Quantization and event-driven execution have existing controls.	Hardware mapping is distinct from a new spike-neuron model.

B ACQUISITION UNITS, ENROLLMENT AND SOURCE PROVENANCE

The dataset contains 13,040 participants assigned to mutually exclusive development, support, release-validation and query roles. E1 and E2 identify independent external health-system cohorts. Participant, eye, visit and capture keys remain distinct throughout the analysis. Derived views are never counted as additional participants. Table 8 reproduces the allocation in the main paper.

Eligibility is restricted to adults aged at least 50 years. The staging track retains late AMD and prior treatment as recorded factors. The 512 pair participants receive one additional index-eye capture at 90 days, within a 14-day scheduling window, so the primary 30–180-day pairing rule does not depend on a six-calendar-month visit occurring before day 180. The incident-risk track excludes a known late-AMD history or prior late-AMD treatment in the index eye. Death ascertainment and loss-to-follow-up procedures are established before enrollment. Clinical assessments determine referrals independently of the model throughout research collection. No automated delay of clinical care is part of the design.

Table 7: Denominator and object definitions.

Object	Unique key	Counting rule
Participant	Study pseudonym	One person, all eyes and visits. Never count image crops as people.
Eye	Participant + OD/OS	At most two eyes per participant. Primary eye assigned from hash parity before quality inspection.
Visit	Participant + encounter timestamp	One clinical encounter. Multiple cameras or attempts do not create extra visits.
Capture	Source SHA-256 + attempt ID	One native image file. Raw duplicates collapse but remain in the audit log.
Derived view	Parent capture + transform hash	Full and central crops count as views, never independent captures.
Longitudinal pair	Two same-eye visit keys	Pair eligibility uses actual time and treatment metadata.
Risk landmark	Participant + index-eye baseline	One primary eligible landmark per participant. Repeated-landmark analysis is secondary and clustered.
Outcome	Landmark + cause + date	First specialist confirmation or death. Neither stage-prediction date nor last negative visit is an exact onset.

Table 8: Complete mutually exclusive allocation, identical to the main cohort table.

Cohort	n	Train	Tune	Cal.	Support	Release	Query
D0	2400	1600	400	200	0	0	200
D1	920	0	0	0	80	600	240
D2	920	0	0	0	80	600	240
D3	920	0	0	0	80	600	240
D4	1880	0	0	0	80	600	1200
E1	3000	0	0	0	0	0	3000
E2	3000	0	0	0	0	0	3000
Total	13040	1600	400	200	320	2400	8120

Table 9: Capture-capacity arithmetic, not observed acquisition volume.

Capacity item	Formula	Count
Participants	fixed allocation	13,040
Routine scheduled visits	$13{,}040 \times 4$	52,160
Additional pair visits	512×1	512
All visit slots	52,160 + 512	52,672
Routine bilateral captures	$52{,}160 \times 2$	104,320
Additional repeat captures	240×3	720
Additional pair captures	512×1	512
Pre-retake capture capacity	104,320 + 720 + 512	105,552
Retake allowance	ceiling of 15% of 105,552	15,833
Storage/annotation capture ceiling	105,552 + 15,833	121,385

Table 10: Acquisition specification and response to failure.

Item	Fixed primary specification	Failure rule
Source	Native RGB fundus image, retained byte-for-byte. New images have short side at least 1,024 pixels. Camera field of view is stored.	No recompression or reconstructed image substituted for a native source.
Position	Macula-centered capture with recorded fixation and dilation status. Capture both eyes when feasible.	A central crop is not asserted to be foveal when centering is uncertain.
Timing	Baseline, 6, 12 and 24 month target visits. Store exact dates and actual intervals.	Late visits are not rounded backward into a risk horizon.
Quality	One source capture plus at most one immediate quality-driven retake.	Repeat failure goes to specialist workflow and remains in the screening denominator.
Operators	Pseudonymized operator ID, training level and workflow version.	Unresolved operator identity is a missing-data flag, not a fabricated category.
Metadata	Site, camera make/model, serial hash, field of view, resolution, exposure metadata where available and source timestamp.	Resolve required identity/laterality fields before model eligibility.
Clinical linkage	Specialist-confirmation, death, last contact, treatment type and treatment date.	No risk landmark without resolvable late-AMD history. Loss censoring at last verified event-free contact.

B.1 Crossed acquisition substudy and matching

The crossed-acquisition substudy includes 240 D0-training participants with a gradable index eye. Each participant is photographed by two operators with two cameras during the same visit, with randomized acquisition order and the first routine capture counted only once. Six pairwise feature differences per participant receive weight $1/6$ when estimating the repeat moment N. This produces 1,440 contrasts but only 240 independent participant clusters. No statement about lesion change is inferred from same-day appearance differences.

Longitudinal C/S matching requires the same camera, site, baseline-stage category and quality category, with time-gap differences of at most 30 days and age differences of at most five years. A deterministic nearest-neighbor match without replacement is selected by acquisition metadata, with participant ID as the tie-breaker. The 256 eligible pairs per group are formed within D0 training. No query participant enters moment estimation. The primary algorithm uses the fixed $S + 0.5N$ nuisance reference. The matched-set formation is an experimental-design restriction, not a causal guarantee.

Table 11: Nested repeated-acquisition accounting.

Factor	Levels	Participants	Captures	Contrasts
Camera $\times$ operator	2×2	240	960	1440
Additional to scheduled source	3 extra / index eye	0	720	0
Longitudinal change	1 pair / participant	256	512	256
Longitudinal stable	1 pair / participant	256	512	256

C Annotation and data-record definitions

Stage definitions follow the adopted Beckman rubric (Ferris et al., 2013). Stage annotations and first-event confirmation are different records. A late lesion requires specialist confirmation for outcome labeling, while a model's suspected late-AMD output is a referral state. The selected index eye is not replaced because its photograph is difficult to grade. The primary referral endpoint uses the index eye. Bilateral captures do not create independent participant observations.

Table 12: Clinical target definitions.

Target	Operational rule	Recorded uncertainty
C0	No AMD under the adopted rubric, including only small drusen below 63 micrometers without AMD-related pigment abnormality.	Ambiguous diameter or pigment origin receives an adjudication flag.
C1	Medium drusen at least 63 and below 125 micrometers, without AMD-related pigment abnormality or late lesion.	No model-derived lesion size is accepted as the reference.
C2	Large drusen at least 125 micrometers, or AMD-related pigment abnormality with at least medium drusen, without a late lesion.	Store drusen and pigment separately as well as final stage.
C3	Specialist-confirmed geographic atrophy or neovascular AMD under the clinical assessment protocol.	Store subtype, confirmation evidence and treatment history.
Quality failure	Insufficient image evidence for the relevant grading field after the defined capture attempt.	Blur, exposure, field loss, opacity, obstruction or unsupported file.
Stable pair	No adjudicated change in the recorded lesion attributes, both images gradable.	Same coarse stage alone does not establish stability.
Change pair	Adjudicated lesion worsening without treatment crossing.	Record the changed attribute and grader confidence.

Table 13: Annotation workflow, calibration and escalation.

Phase	Allocation and criterion	Response
Grader calibration	100 images spanning stages and quality reasons. Stage weighted kappa at least 0.80, quality kappa at least 0.85.	Retrain graders on disagreements. Use a fresh 100-image batch for reassessment, not repeated memorization of the same test.
Production	Two independent grades for every primary benchmark image, with masked model outputs.	Senior specialist adjudicates all stage, late-lesion and quality disagreements.
Re-read	Random 10% of resolved records at least two weeks later, with shuffled order.	Report within-grader drift and disagreement categories.
Late event	Source specialist record and date, checked independently by two abstractors.	Resolve all event-date/cause disagreements before unblinding queries.
Corrections	Immutable source, label and split versions.	Retain versioned corrections; do not silently overwrite labels.

Table 14: Data-record schema and integrity assertions.

Record type	Required fields	Assertions
Participant record	participant_id, primary_eye, age, age_missing, consent_scope	Globally unique ID and an index eye assigned before image QC.
Capture record	participant_id, eye, visit_id, capture_id, attempt, SHA-256, site, camera, operator, time	Parent participant exists, laterality valid, derived files retain source identity.
Label record	capture_id, quality, reason, stage, lesion attributes, grader IDs, adjudication version	Ungradable sources do not receive model-derived reference labels.
Landmark record	participant_id, origin, history_capture, event_time, event_cause, last_contact	Features precede origin. All outcomes occur after origin. Death and loss are distinct.
Partition record	participant_id, domain, role, release_version	Exactly one role per participant. No support/release/query overlap.
Pair record	earlier_capture, later_capture, group, treatment_crossing, matching_block	Same eye and camera, eligible gap, no unresolved lesion label.
Model-state record	encoder_hash, adapter_hash, risk_hash, calibration_hash, parent_version	Only stage-adapter hash may change in an immediate local update.

D CALENDAR, CENSORING, MISSINGNESS AND EXTERNAL STUDIES

The acquisition protocol fixes an origin T_0 before D4 query collection. Historical observations support the longitudinal analyses only when feature and outcome timestamps are recoverable. D4 provides a held-out staging and quality domain. The six- and twelve-month risk comparisons use the outcome-mature E1/E2 cohorts. The duration of model development is distinct from the clinical observation intervals.

The event of interest is documented first specialist confirmation. This timing may reflect access to care and is not claimed to be biological onset. Death before confirmation is a competing event. A last-negative/first-positive interval belongs only to the secondary onset sensitivity. Intervals crossing a horizon do not become known binary labels. Interval-based onset and documented-confirmation analyses have different estimands. Event dates are not imputed from prediction scores.

Table 15: Fixed time and information restrictions.

Analysis	Allowed information	Excluded information
D0 selection	Training and tune participants before calibration freeze.	Calibration/release/query labels as optimization inputs.
Local stage update	Available current-stage support labels and permitted past replay.	Future six/twelve-month outcomes or query labels.
Risk prediction	Base features at or before origin, age, masks, gap, known prior treatment.	Post-origin treatment or specialist findings as predictors.
D4 zero-shot	Frozen encoder, stage/risk heads and calibrators from prior development.	Unlabeled D4 query statistics for QAT, normalization or TTA.
D4 adaptation	Dedicated 80-person support reserve and independent release cohort.	1,200 consecutive D4 query participants in any update.
E1/E2 primary	Fixed risk model on two independent query cohorts.	External recalibration on the same observations used to claim external validity.

Table 16: External outcome ledger. Administrative survivors, losses, AMD events and deaths are distinct. Counts are shared across all risk methods.

System	People	AMD by 12	Death by 12	Lost <12	Alive/event-free at 12
1	3000	249	77	433	2241
2	3000	304	60	431	2205

E REPRODUCIBLE NETWORK, PREPROCESSING AND TRAINING

The primary encoder is the explicit compact depthwise-separable network specified in Table 17. It is not claimed to be a new convolutional primitive. Its purpose is a reproducible low-cost test bed for the constrained adapter. Each separable block has a linear depthwise/pointwise pair followed by batch normalization and ReLU6. There is no intervening nonlinear activation between depthwise and pointwise operations. Batch-normalization parameters fold into the convolution for the deployment parameter count. The reported MAC total counts only convolutional multiply-accumulates, excluding pooling, quality control, adaptation, task heads, transfer and other host operations.

E.1 FIGURE-TO-METHOD CORRESPONDENCE

Figure 1 describes data roles and the study lifecycle, Figure 2 describes basis estimation and stage-only optimization, and Figure 3 describes tensor dimensions and execution ownership. Shared module names identify the same components across the three figures. In particular, C, S, N are uncentered moments under one norm clip, the primary pair sets contain 256 change and 256 stable pairs, and the 240-person acquisition-repeat study is nested within training. Matrix textures are illustrations, not measured covariances.

The network figure uses the explicit reference encoder in Table 17, not an AkidaNet-0.5 checkpoint. The risk MLP has ReLU6 and dropout 0.1 during base training. Evaluation mode disables dropout and freezes batch-normalization state. The red risk path receives unadapted base features, so stage-parameter changes do not alter a fixed-input risk output. The optional $W'_s = W_s A_s$ fold is a signed FP32 host readout, not a new quantized adapter layer on the accelerator. Image examples and hardware photographs illustrate roles only. They do not document recruitment, temporal progression, device ownership or a measured run.

Table 17: Exact encoder layout. Dimensions and parameter/MAC counts are computed from the declared graph.

Layer	Input	Output	Weights	Fused bias	MACs/view
Conv 3	224 x 224 x 3	112 x 112 x 16	432	16	5,419,008
DS 1	112 x 112 x 16	56 x 56 x 24	528	24	1,655,808
DS 2	56 x 56 x 24	28 x 28 x 32	984	32	771,456
DS 3	28 x 28 x 32	14 x 14 x 64	2336	64	457,856
DS 4	14 x 14 x 64	7 x 7 x 96	6720	96	329,280
DS 5	7 x 7 x 96	7 x 7 x 128	13152	128	644,448

Table 18: Arithmetic counts from the reference architecture, not measured memory or runtime.

Object	Computed quantity
encoder_fused_parameters	24,512
macs_per_view	9,277,856
macs_two_views	18,555,712
stage_parameters	1,028
risk_parameters	33,868
adapter_trainable	4,096
adapter_bytes_fp 32	16,384
adam_params_grad_m_v_bytes	65,536
replay_feature_bytes	131,072
basis_B_bytes	8,192

Table 19: Deterministic image preprocessing and conservative training augmentation.

Operation	Primary setting	Reason/check
Field mask	Largest connected component with channel maximum above 8/255, followed by a 5-pixel closing.	Reject empty or fragmented fields instead of silently inventing a crop.
Views	Square retinal-field bounding box, black padding to square, resize bilinear to 224. Central crop side 0.5 of estimated diameter.	One source, two views, shared weights. Store exact crop coordinates.
Input	Unsigned RGB, numerical scale x/255 for the reference network.	No unreported channel-mean subtraction on an unsigned device input.
Geometry	Horizontal flip 0.5, rotation uniform[-7,7] degrees, isotropic scale[0.94,1.00].	Same transform for both views from a source.
Photometry	Brightness and contrast multipliers[0.90,1.10].	No hue jitter, erasing or mixed-image augmentation in the primary configuration.
History	Nearest eligible prior same-eye/same-camera capture within 180 days.	Deterministic tie-breaker by capture ID. No future or fellow-eye replacement.
Age	Training mean/SD, clipped z-score[-4,4], missing value 0 plus mask.	Statistics fitted on D0 training only.
Gap	Days divided by 180 and clipped[0,1], zero if history absent.	Masks distinguish missing history from same-day timing.

Table 20: Base fitting, selection and calibration. Values are fixed experimental settings.

Phase	Exact configuration	Selection boundary
Base model	AdamW 80 epochs, batch 64, lr 3 e-4, betas 0.9/0.999, weight decay 1 e-4, five-epoch warm-up, cosine schedule.	Checkpoints selected by mean of standardized stage/risk validation losses on D0 tune, with standardization constants fitted from epoch 1.
Stage loss	Effective-number weights beta 0.999, normalized to mean 1, plus 0.15 expected ordinal distance.	Class counts from training participants, not query prevalence.
Risk loss	Cause-specific piecewise-exponential NLL, averaged over eligible participant landmarks.	Observed partial follow-up contributes exposure. Unobserved future intervals are not negatives.
Joint base objective	Stage loss plus risk NLL, each averaged over its eligible batch, coefficient 1.	A missing entire task batch contributes zero only for that task.
QAT	20 epochs, lr 3 e-5, first weights 8 bits, later weights/activations 4 bits, batchnorm frozen.	The encoder is the quantized block. Signed head and adapter operations remain FP32.
Stage calibration	One positive temperature T in[0.25,4] minimizing calibration cross-entropy, bounded scalar optimization tolerance 1 e-6.	D0 calibration only, frozen before local updates.
Risk calibration	One shared log hazard multiplier constrained [-2,2], minimizing calibration NLL plus its squared value by bounded scalar optimization.	No independent horizon scaling that violates cumulative-risk coherence.
Quality calibration	Highest failure-score threshold with at least 98% observed failure recall, ties favor fewer false passes.	Calibration labels remain separate from query labels.

E.2 Local objective and all reduction constants

Let class weights be $w_c \propto (1-\beta)/(1-\beta^{n_c})$ with their mean normalized to one and $\beta = 0.999$. For a stage observation, use $-w_y \log p_y + 0.15 \sum_c p_c |c-y|$. Support and replay stage losses are separately batch-averaged before applying the replay coefficient 0.5. The anchor is $4\,\mathrm{KL}(\mathrm{softmax}(a_{old}/2) \| \mathrm{softmax}(a_{new}/2))$, averaged over its replay batch. Stable loss is $|\mathcal{P}_s|^{-1} \sum_\delta \|(A_s - I)\delta\|_2^2/256$ on a 32-pair minibatch. The norm penalty is on the actual residual $QUV^\top Q$, not an unrelated norm of the factors. The small AdamW factor decay is listed separately from that objective penalty. No risk loss, risk distillation or fold-consistency term trains the local stage factors in the primary method.

The previous accepted stage model is the anchor teacher. Rejected candidates never replace the teacher. Adam moments reset for each local domain, while the accepted U and V initialize the next update. The encoder and risk head remain in evaluation mode, fixing batch-normalization statistics and disabling dropout. The stage and risk branches therefore share immutable base features without sharing a locally changing transform.

Table 21: Immediate local update state and budget.

Item	Primary value	Auditable behavior
Parameters	U/V only, 2 times 256 times 8 scalars.	Zero-initialize U for first deployment. V Gaussian SD 0.01.
Optimizer	AdamW 100 steps, lr 1 e-3, betas 0.9/0.999, decay 1 e-4, norm clip 1.	The step count is fixed independently of query performance.
Batches	16 support,64 replay,32 stable-pair differences, with replacement when required.	Same sampled IDs for matched methods.
Support	10 per stage primary,5/20 nested sensitivity.	Per-domain reserve 80 people. The primary balanced support uses 40 participants.
Replay	Stage-stratified reservoir, up to 128 records per stage, fixed seed.	Reservoir contains only permitted prior support or training records.
Teacher	Previous accepted stage logits recomputed before update on replay.	Risk outputs remain those of the unchanged risk branch.
Precision	FP32 factor optimization and signed host inference.	Encoder quantization remains fixed. No local encoder recalibration.
Resource gates	Update time at most 300 s, peak host memory at most 2 GiB.	These are acceptance thresholds, not observed update-time or memory values.

F MATHEMATICAL GUARANTEE, CAPACITY COST AND IMPLEMENTATION CHECKS

For $Q = I - P$ with an orthogonal projector P, the residual $R = QUV^\top Q$ satisfies $PR = RP = 0$. Thus $A_s P = PA_s = P$ and $P(A_s z - z) = 0$ for every vector z. The one-sided residual $UV^\top Q$ need not satisfy $PUV^\top Q = 0$. For example, with $P = \mathrm{diag}(1, 0)$, $U = (1, 0)^\top$, $V = (0, 1)^\top$ and $z = (0, 1)^\top$, the one-sided adapter adds one to the protected coordinate. This is an explicit counterexample, not merely a difference in terminology.

These equalities do not establish preservation of all pathological information. The selected basis can omit a clinically important direction, and an unchanged projected coordinate does not imply an unchanged classifier logit. A rank-limited complementary residual can also be insufficient to correct a real shift.

F.1 APPROXIMATION TRADE-OFF

Let E be a desired linear correction, and require a residual R with $R = QRQ$ and rank at most r. Orthogonal decomposition and best rank-r approximation give

$$\min_{R=QRQ,\ \mathrm{rank}(R)\le r} \|E - R\|_F^2 = \|E - QEQ\|_F^2 + \sum_{j>r} \sigma_j(QEQ)^2. \tag{7}$$

The first term is the correction forbidden by the protected coordinates. The second is the approximation cost of the low-rank budget inside the complement. Both are nonnegative. No optimizer can eliminate the first term without violating the constraint. The numerical program evaluates both sides under five controlled allowed/forbidden correction mixtures. These checks are not a theorem of clinical transfer or a proof of novelty over all prior projected adaptations.

Table 22: Numerical verification of Equation 7. Angle indexes a controlled allowed/forbidden correction mixture in a numerical test, not a measured retinal shift.

Mixture angle	Optimal residual^2	Forbidden^2	Rank-tail^2
0	0.6661931	0.0000000	0.6661931
15	0.6869196	0.0276221	0.6592974
30	0.6540158	0.0937352	0.5602807
60	0.4471521	0.2608177	0.1863344
90	0.4324588	0.4324588	0.0000000

Table 23: Actual random-matrix calculations. Right/left columns are Frobenius norms of $AP - P$ and $PA - P$. Coordinate drift is maximum over 512 random vectors. INT4 means naive rounding of A, not the deployed device path.

Basis rank	Operator	Right residual	Left residual	Max coordinate drift
4	one-sided	1.05 e-15	2.14 e-02	5.55 e-02
4	two-sided	1.04 e-15	1.06 e-15	2.53 e-15
4	two-sided FP32	5.50 e-08	5.49 e-08	1.09 e-07
4	INT4 after two-sided	5.30 e-03	5.30 e-03	1.05 e-02
8	one-sided	1.53 e-15	4.35 e-02	9.28 e-02
8	two-sided	1.52 e-15	1.52 e-15	2.91 e-15
8	two-sided FP32	8.42 e-08	8.42 e-08	1.65 e-07
8	INT4 after two-sided	1.09 e-02	1.09 e-02	1.95 e-02
16	one-sided	2.17 e-15	6.31 e-02	1.31 e-01
16	two-sided	2.17 e-15	2.18 e-15	3.45 e-15
16	two-sided FP32	1.07 e-07	1.07 e-07	1.62 e-07
16	INT4 after two-sided	1.26 e-02	1.26 e-02	2.08 e-02
32	one-sided	3.02 e-15	9.25 e-02	1.70 e-01
32	two-sided	3.04 e-15	3.03 e-15	4.91 e-15
32	two-sided FP32	1.54 e-07	1.54 e-07	2.21 e-07
32	INT4 after two-sided	1.71 e-02	1.71 e-02	2.38 e-02

F.2 Which basis is being compared

The primary moment is computed from immutable post-QAT base features, not from adapted coordinates. The clipping radius is shared across change, stable and repeat groups. Removing their group means can erase a consistent progression direction, and normalizing every difference to unit length can overweight small nuisance changes. Centered and unit-normalized versions therefore remain explicit alternative constructions rather than silently changing the method.

A four-class cross-sectional LDA has at most three discriminant directions. The rank 3 comparison uses rank 3 clinical, random, PCA and LDA bases with the same residual rank. A separate LDA 3+PCA 5 hybrid is named as such if a rank 8 reference is required. Raw principal angles between independent encoder seeds are not meaningful without aligning feature coordinate systems. A cross-seed comparison would require Procrustes alignment fitted on a shared D0-training anchor set. Within-seed pair-resampling angles need no such alignment.

Table 24: Basis controls and matching requirements.

Control	Construction	Fixed factors
One-sided clinical	Same B but residual UV-transpose Q.	All losses, sampling, gates, seeds and support labels identical.
Two-sided clinical	Primary change/stable/repeat ratio.	Rank 8 basis, rank 8 residual,4096 factors.
Output-only	Residual QUV-transpose.	Tests output preservation without identity action on protected inputs.

Table 24 (continued)

Control	Construction	Fixed factors
Random	Seeded orthogonal rank 8 projector.	Matched complement dimension and residual capacity.
PCA	Top 8 eigenvectors of D0-training base features.	No query observations. Same factor budget.
Stable PCA	Top 8 stable-difference directions.	Stable labels available to every matched method.
Stage LDA	Rank 3 discriminants, plus named rank 3 comparison family.	Rank 8 LDA is not defined by four stage classes.
Gradient	A fixed D0-training activation/gradient subspace.	Exporter records exact layer, objective and dimensionality.
Shuffled clinical	Shuffle pair-group labels within matching blocks.	Acquisition distribution unchanged, clinical semantics disrupted.
No repeat moment	Set N weight to 0.	Tests whether crossed acquisition information adds value.

G MATCHED COMPARATORS, NATIVE BASELINES AND MODEL-SELECTION BUDGETS

A comparison called FA is a feature-adapter reduction of an established idea, not a claim to reproduce its native neural architecture. The common feature space and stage readout are frozen. The same support labels, stable-pair labels, replay candidates and previously accepted outputs are available to all matched methods. Methods that do not use particular information still have their accessible information documented. Matching rank alone does not control readout training, stored feature values or access to additional images.

The main comparisons include 18 alternative representations and 12 alternative continual-learning rules alongside RetinaLPA-2S. Numerical results use the stated evaluation protocol. Citations identify the underlying methods rather than the source of their reported performance values. Feature-adapter reductions retain the FA or style qualifier and are not presented as native-architecture reproductions.

Table 25: Feature-adapter control definitions and native-family distinction.

Rule	Matched FA operation	Native experiment
Fine-tune	Optimize unconstrained rank 8 factors on support CE.	Support-only factor update; encoder and head frozen.
LwF	Same factors with temperature 2 old-logit distillation.	Feature-adapter reduction, not the native published architecture.
EWC	Same factors with diagonal Fisher penalty from permitted old records.	Fisher memory and construction time counted.
A-GEM	Project factor gradient against gradient on shared replay batch.	Same reference batch size 64.
ER	Same factors and readout, support and replay CE.	Feature-only replay under the shared buffer limit.
MIR	Retrieve highest predicted loss-increase records from a 128-record candidate subset.	The candidate selection cost is measured.
DER++	Replay CE plus MSE to stored old logits, common feature/factor state.	Original image-level DER++ is not conflated with this reduction.

Table 25 (continued)

Rule	Matched FA operation	Native experiment
GPM	Project common factor updates using D0-derived activation subspace.	Subspace storage added to total bytes.
BiLoRA-style FA	Bilinear feature update $z' = [I + F_{\text{out}}(M_t \odot \Theta_t)F_{\text{in}}^{\top}]z$ with fixed orthogonal bases and a prespecified sparse support mask; only the masked coefficients are trained under the matched 4,096-scalar budget.	Same frozen encoder/readout and information budget as the matched FA track; feature-space reduction, not the native BiLoRA architecture.
One-sided/two-sided	Change only left projection under matched full losses.	Matched comparison of the two projection constraints.

Table 26: Foundation-model regimes and matching rules.

Regime	Configuration	Budget/reporting
Frozen probe	Freeze encoder, train task heads on D0 training. Use supported native image resolution.	18 representation families. Task heads follow each backbone's native feature dimension.
Rank 8 PEFT	Query/value low-rank factors where supported, scale 16, dropout 0.05.	RETFound, VisionFM and RetiZero primary. Adapters use supported query/value projections.
Full tuning	All permitted backbone and task-head parameters.	Same participant labels and outcome eligibility; backbone and head parameters are trainable.
Twenty-four trials	4 learning rates times 3 weight decays times 2 augmentation strengths.	lr in 1 e-5,3 e-5,1 e-4,3 e-4. Decay 1 e-4,1 e-3,1 e-2. Reduced augmentation uses half-width ranges.
Confirmation	Five seeds for the selected setting, at most 80 epochs, patience 10 on D0 tune.	Calibrator fitted afterward on D0 calibration. Trial failures and selection scores retained.
Temporal FM	Same current/prior information, gap, age, masks and cause-specific likelihood.	History and age are available to the strongest risk comparator.

H COMPLETE STAGING RESULTS AND THRESHOLD ANALYSIS

All 19 methods are evaluated on the same 920 query participants with the same labels and domain allocation. Predictions are evaluated without fitting on query data. The separate D0 calibration partition selects the referral threshold from 0.05 to 0.95 in 0.01 increments. The selected threshold is 0.33 and remains fixed across query domains.

The staging snapshot computes macro-F1 from the pooled participant predictions. Continual F1 is the equal mean of domain-specific F1 after updates and is not expected to equal the snapshot metric. The two metrics therefore answer different evaluation questions.

Table 27: Complete staging results. Bootstrap intervals use paired resampling of shared query participants.

Method	F1	95% interval	Acc.	κ_w	C1 F1
MobileNetV3-L	78.95	[76.22, 81.60]	80.87	0.921	63.06
ResNet-50	77.54	[74.75, 80.22]	79.67	0.917	59.76
EfficientNet-B0	79.36	[76.40, 81.90]	81.30	0.923	63.64
ConvNeXt-T	80.18	[77.41, 82.63]	81.96	0.925	64.67
ConvNeXtV2-T	81.38	[78.64, 83.82]	83.26	0.931	66.87
ViT-B/16	80.51	[77.74, 83.01]	82.39	0.929	65.24

Table 27 (continued)

Method	F1	95% interval	Acc.	κ_w	C1 F1
Swin-T	81.31	[78.52, 83.82]	83.26	0.931	66.25
MaxViT-T	82.17	[79.52, 84.78]	84.02	0.934	67.71
MAE ViT-B probe	81.65	[78.92, 84.21]	83.48	0.932	66.25
DeepDrAMD-style	81.49	[78.88, 84.00]	83.48	0.932	66.88
DINOv2 ViT-B	83.12	[80.53, 85.49]	84.78	0.937	70.55
MedCLIP probe	79.47	[76.66, 82.17]	81.52	0.924	63.58
RETFound probe	83.37	[80.67, 85.84]	85.00	0.939	69.33
RET-CLIP probe	82.29	[79.54, 84.68]	83.91	0.935	67.47
VisionFM probe	83.56	[81.00, 85.81]	85.33	0.939	69.84
FLAIR probe	82.55	[79.91, 84.95]	84.35	0.937	67.29
UrFound probe	83.35	[80.77, 85.77]	85.00	0.938	69.14
RetiZero probe	83.57	[80.98, 85.92]	85.33	0.941	69.81
RetinaLPA-2S	83.78	[81.23, 86.10]	85.43	0.941	69.16

Table 28: Exact confusion counts for the compact staging model.

Reference / predicted	C0	C1	C2	C3	Row total
C0	299	31	0	0	330
C1	15	111	24	0	150
C2	0	29	241	20	290
C3	0	0	15	135	150
Column total	314	171	280	155	920

Table 29: Classwise metrics recomputed from Table 28.

Stage	n	Precision	Recall	F1
C0	330	95.22	90.61	92.86
C1	150	64.91	74.00	69.16
C2	290	86.07	83.10	84.56
C3	150	87.10	90.00	88.52

Table 30: Query-domain snapshots from the same participant records as the pooled staging table. No averaging of domain accuracies is substituted for pooled counts.

Domain	n	F1	Acc.	κ_w	C3 argmax recall
D0	200	83.81	85.50	0.941	96.67
D1	240	80.89	83.75	0.935	85.71
D2	240	83.46	84.58	0.939	93.33
D3	240	86.68	87.92	0.950	85.00

Table 31: The same query predictions at all thresholds. The calibration-locked threshold is 0.33. TP and FP cannot increase with threshold.

Threshold	TP	FN	TN	FP	Sensitivity	Specificity
0.10	149	1	652	118	99.33	84.68
0.20	148	2	704	66	98.67	91.43
0.30	145	5	726	44	96.67	94.29
0.33	143	7	730	40	95.33	94.81
0.40	140	10	741	29	93.33	96.23
0.50	133	17	751	19	88.67	97.53
0.60	123	27	760	10	82.00	98.70
0.70	117	33	766	4	78.00	99.48
0.80	102	48	769	1	68.00	99.87
0.90	73	77	770	0	48.67	100.00

Table 32: The same fixed threshold transported to each query domain. PPV/NPV are derived from these integer counts, not separately assigned.

Domain	TP	FN	TN	FP	Sens.	Spec.	PPV	NPV
D0	29	1	165	5	96.67	97.06	85.29	99.40
D1	33	2	189	16	94.29	92.20	67.35	98.95
D2	44	1	185	10	97.78	94.87	81.48	99.46
D3	37	3	191	9	92.50	95.50	80.43	98.45

Table 33: Confidence-retention subsets. These do not establish improved whole-population screening.

Cutoff	Retained n	Coverage%	F1	C3 argmax recall
0.00	920	100.00	83.78	90.00
0.40	920	100.00	83.78	90.00
0.50	888	96.52	85.70	90.48
0.60	690	75.00	89.87	93.18
0.70	468	50.87	72.61	96.69
0.80	258	28.04	49.88	100.00
0.90	145	15.76	50.00	100.00

Table 34: Paired participant-bootstrap intervals for compact F1 minus each comparator.

Comparator	Difference	Paired 95% interval
MobileNetV3-L	4.83	[3.44, 6.41]
ResNet-50	6.24	[4.60, 8.08]
EfficientNet-B0	4.41	[2.99, 5.96]
ConvNeXt-T	3.59	[2.30, 4.98]
ConvNeXtV2-T	2.39	[1.18, 3.74]
ViT-B/16	3.26	[2.00, 4.60]
Swin-T	2.46	[1.07, 3.97]
MaxViT-T	1.61	[0.24, 2.99]
MAE ViT-B probe	2.12	[0.56, 3.67]
DeepDrAMD-style	2.28	[0.87, 3.74]
DINOv2 ViT-B	0.65	[-0.72, 1.98]
MedCLIP probe	4.30	[2.77, 5.81]
RETFound probe	0.40	[-1.07, 1.83]
RET-CLIP probe	1.48	[0.38, 2.67]
VisionFM probe	0.22	[-0.97, 1.43]

Table 34 (continued)

Comparator	Difference	Paired 95% interval
FLAIR probe	1.22	[0.14, 2.41]
UrFound probe	0.43	[-0.87, 1.83]
RetiZero probe	0.20	[-1.05, 1.47]

I CONTINUAL TRAJECTORIES, FACTORIAL ATTRIBUTION AND ACCEPTED STATES

All six domain orders, 123, 132, 213, 231, 312 and 321, are evaluated with seeds 11, 23, 37, 53 and 71. The three support budgets yield 1,170 candidate trajectories over 13 methods. A trajectory contains the initial state and three domain updates. The state matrix stores a score for each of four query domains at every time. Main-table forgetting uses the base domain and the first two target domains, with the maximum taken from that domain's arrival through the final state. The last target has no subsequent update and is excluded from the forgetting mean.

Candidate trajectories quantify the optimized stage models before release decisions. A rejected update retains the previous accepted checkpoint. Accepted-state evaluation uses that checkpoint rather than the rejected candidate. The independent C3 sensitivity-subgate results in Table 49 are separate from these candidate matrices and do not imply that all 30 trajectories pass every release gate. Multiple method comparisons on a release set are research evaluations, not certification of the selected model for clinical use.

Table 35: All methods and support sizes. The 30 trajectories in each cell share the same query participants.

Method	Shot	Final F1	Trajectory SD	Forget. pp	Runs
Frozen base	5	72.23	1.77	0.00	30
Frozen base	10	72.23	1.77	0.00	30
Frozen base	20	72.23	1.77	0.00	30
Fine-tune	5	74.93	1.98	5.95	30
Fine-tune	10	76.65	1.92	6.29	30
Fine-tune	20	77.43	1.76	6.31	30
LwF-FA	5	75.71	1.97	2.57	30
LwF-FA	10	77.06	1.76	2.66	30
LwF-FA	20	77.58	1.79	2.77	30
EWC-FA	5	75.88	1.93	2.33	30
EWC-FA	10	77.29	1.68	2.35	30
EWC-FA	20	77.83	1.74	2.43	30
A-GEM-FA	5	76.68	1.85	2.28	30
A-GEM-FA	10	78.41	1.63	2.51	30
A-GEM-FA	20	79.08	1.70	2.65	30
ER-FA	5	76.92	1.77	2.57	30
ER-FA	10	78.76	1.70	3.03	30
ER-FA	20	79.46	1.64	2.72	30
MIR-FA	5	77.63	1.74	2.11	30
MIR-FA	10	79.53	1.63	2.17	30
MIR-FA	20	80.32	1.55	2.06	30
DER++-FA	5	78.24	1.64	1.53	30
DER++-FA	10	80.33	1.60	1.39	30
DER++-FA	20	81.08	1.52	1.44	30
GPM-FA	5	78.39	1.71	1.09	30
GPM-FA	10	80.39	1.69	1.03	30
GPM-FA	20	81.15	1.63	0.95	30
BiLoRA-style FA	5	78.45	1.72	1.26	30
BiLoRA-style FA	10	80.52	1.72	1.13	30
BiLoRA-style FA	20	81.22	1.63	1.26	30
Unconstrained rank-8	5	77.90	1.61	1.98	30
Unconstrained rank-8	10	79.92	1.55	1.94	30
Unconstrained rank-8	20	80.68	1.47	1.97	30

Table 35 (continued)

Method	Shot	Final F1	Trajectory SD	Forget. pp	Runs
One-sided LPA	5	79.05	1.69	1.44	30
One-sided LPA	10	81.25	1.62	1.36	30
One-sided LPA	20	82.09	1.36	1.46	30
RetinaLPA-2S	5	79.93	1.67	1.09	30
RetinaLPA-2S	10	82.20	1.31	1.07	30
RetinaLPA-2S	20	83.09	1.01	0.85	30

Table 36: Ten-per-stage order sensitivity. Each row summarizes five seeds, not independent clinical replications.

Method	Order	F1	SD	Forget. pp
DER++-FA	123	80.24	1.69	1.50
DER++-FA	132	80.31	1.49	1.42
DER++-FA	213	80.18	1.98	1.59
DER++-FA	231	80.35	2.00	1.36
DER++-FA	312	80.39	1.57	1.31
DER++-FA	321	80.49	1.77	1.17
GPM-FA	123	80.29	1.79	1.15
GPM-FA	132	80.38	1.63	1.04
GPM-FA	213	80.28	1.97	1.17
GPM-FA	231	80.43	2.06	0.97
GPM-FA	312	80.43	1.69	0.97
GPM-FA	321	80.50	1.93	0.87
BiLoRA-style FA	123	80.45	1.76	1.23
BiLoRA-style FA	132	80.48	1.60	1.18
BiLoRA-style FA	213	80.38	2.15	1.32
BiLoRA-style FA	231	80.50	2.08	1.16
BiLoRA-style FA	312	80.61	1.80	1.01
BiLoRA-style FA	321	80.69	1.87	0.90
One-sided LPA	123	81.17	1.75	1.48
One-sided LPA	132	81.12	1.69	1.54
One-sided LPA	213	81.33	1.82	1.26
One-sided LPA	231	81.25	1.88	1.37
One-sided LPA	312	81.34	1.70	1.25
One-sided LPA	321	81.30	1.82	1.29
RetinaLPA-2S	123	82.29	1.39	0.96
RetinaLPA-2S	132	82.22	1.36	1.05
RetinaLPA-2S	213	82.25	1.55	1.02
RetinaLPA-2S	231	82.20	1.47	1.08
RetinaLPA-2S	312	82.12	1.44	1.19
RetinaLPA-2S	321	82.14	1.39	1.16

Table 37: One complete candidate state matrix, seed 11 and order 123.

Update	D0 F1	D1 F1	D2 F1	D3 F1	Domain mean
0	81.95	70.79	61.44	61.71	68.97
1	81.47	84.73	61.44	61.71	72.34
2	81.05	83.65	81.06	61.71	76.87
3	81.05	83.15	80.15	77.01	80.34

Table 38: Complete four-factor ablation. Cell 1111 is identical to the main continual-learning configuration and result.

Projection	Stable	Replay	Anchor	F1	SD	Forget. pp
0	0	0	0	74.81	1.84	4.68
0	0	0	1	76.28	1.83	3.95
0	0	1	0	77.16	1.76	3.57
0	0	1	1	79.13	1.63	2.36
0	1	0	0	75.70	1.69	4.48
0	1	0	1	77.28	1.76	3.59
0	1	1	0	78.26	1.66	3.12
0	1	1	1	80.21	1.65	1.55
1	0	0	0	76.56	1.69	4.37
1	0	0	1	78.26	1.63	3.52
1	0	1	0	79.15	1.65	2.79
1	0	1	1	81.15	1.64	1.30
1	1	0	0	77.43	1.68	4.23
1	1	0	1	79.31	1.65	2.75
1	1	1	0	80.31	1.57	2.23
1	1	1	1	82.20	1.31	1.07

Table 39: Marginal factorial effects averaged over the other three factors.

Factor	Mean on	Mean off	Difference
projection	79.30	77.35	1.94
stable	78.84	77.81	1.02
replay	79.70	76.95	2.74
anchor	79.23	77.42	1.80

Table 40: Candidate and accepted-state reporting rules.

Quantity	Computation	Invalid alternative
Candidate F1	Evaluate every optimized candidate under the fixed protocol.	Replace poor candidates with the previous checkpoint without labeling the change.
Accepted F1	Evaluate the actual checkpoint retained after release checks.	Report candidate F1 while claiming rollback safety.
Forgetting	Use recorded domain-state matrices and arrival times.	Assign forgetting independently of the trajectory.
Rejected update	Count reason, candidate version and retained parent.	Discard failed trajectories from means.
Information budget	Report support, replay, stable labels and release cohort.	Describe ten-shot adaptation as needing only 40 total clinical labels.
Winner certification	One prespecified candidate per domain receives final validation.	Choose a winner using release labels and quote an unadjusted confidence bound.

J COMPETING-RISK LIKELIHOOD, CENSORING AND CALIBRATION DIAGNOSTICS

The risk branch produces 12 positive cause-specific rates measured per month, rather than twelve unrelated probabilities. Each two-month interval has AMD and death hazards. Their sum determines survival, while the cause fraction allocates event probability. The main cumulative-incidence expression sums the probability of surviving to each interval and experiencing the specified event within its observed duration. AMD cumulative incidence is nondecreasing in time. AMD incidence, death incidence and event-free survival sum to one.

For a documented event at time t, the NLL is the integrated total hazard through t minus the log hazard of the observed cause. A right-censored record contributes integrated hazard through its last ascertainment. Administrative censoring at exactly 12 months contributes an observed event-free control at the 12-month evaluation horizon. It must not be dropped by an erroneous strict greater-than comparison. The numerical checks include this boundary.

The loss-censoring survival function G is estimated using five-fold cross-fitted reverse Kaplan–Meier curves within each external system, treating AMD and death as censoring events for the loss process. These estimates are evaluation nuisances rather than predictive-model tuning. Event weights use the appropriate left limit. IPCW relies on the conditional independent-loss assumption and does not automatically correct informative follow-up. External AUC and Brier intervals are interpreted under that assumption.

Table 41: Complete external-risk metrics under one outcome/censoring ledger. ECE is in percent, Brier is a probability-scale mean square.

Method	AUC 6	AUC 12	BS 6	BS 12	ECE 12	Intercept	Slope
Age + stage	0.725	0.733	0.0475	0.0862	3.90	-0.46	0.97
DeepSurv-style	0.732	0.755	0.0495	0.0863	2.59	-0.73	0.71
Nnet-survival-style	0.739	0.762	0.0484	0.0841	2.04	-0.57	0.76
DeepHit-style	0.741	0.763	0.0482	0.0839	1.97	-0.55	0.77
Dynamic-DeepHit-style	0.753	0.774	0.0470	0.0813	1.17	-0.31	0.87
Current-image hazard	0.741	0.763	0.0479	0.0833	1.74	-0.49	0.78
CNN-LSTM temporal	0.757	0.777	0.0467	0.0807	0.94	-0.24	0.90
CNN-Transformer temporal	0.759	0.779	0.0466	0.0804	0.90	-0.21	0.92
RETFound temporal PEFT	0.763	0.782	0.0463	0.0799	0.75	-0.15	0.95
VisionFM temporal full	0.766	0.785	0.0461	0.0794	0.78	-0.10	0.97
RetinaLPA-2S frozen risk	0.761	0.781	0.0464	0.0800	0.77	-0.16	0.94
No-history frozen risk	0.739	0.762	0.0481	0.0837	1.84	-0.52	0.77

Table 42: Paired stratified participant-bootstrap uncertainty on external records, 1,000 draws.

Method	AUC 12	AUC 95% interval	Brier 12	Brier 95% interval
VisionFM temporal full	0.7849	[0.7678, 0.8042]	0.07944	[0.07413, 0.08470]
RetinaLPA-2S frozen risk	0.7814	[0.7639, 0.8009]	0.08002	[0.07469, 0.08534]
No-history frozen risk	0.7616	[0.7440, 0.7817]	0.08369	[0.07822, 0.08924]

Table 43: Compact frozen risk minus the fully tuned temporal-foundation comparator. Lower AUC and higher Brier favor the comparator.

Metric	Difference	Paired 95% interval
AUC 12	-0.00351	[-0.00459, -0.00245]
Brier 12	0.000580	[0.000442, 0.000727]

Table 44: Calibration intercept and slope with 1,000 participant-bootstrap fits. Diagnostic fitting does not change model predictions.

Method	Intercept [95% CI]	Slope [95% CI]
VisionFM temporal full	-0.10 [-0.28, 0.09]	0.97 [0.89, 1.07]
RetinaLPA-2S frozen risk	-0.16 [-0.35, 0.01]	0.94 [0.85, 1.03]
No-history frozen risk	-0.52 [-0.68, -0.36]	0.77 [0.70, 0.85]

Table 45: Within-C2 discrimination avoids crediting a risk predictor merely for separating baseline stages.

Method	C2 events	C2 deaths	AUC 12	Brier 12	ECE 12%
VisionFM temporal full	410	50	0.672	0.1391	1.93
RetinaLPA-2S frozen risk	410	50	0.666	0.1404	2.22
No-history frozen risk	410	50	0.635	0.1483	4.65

Table 46: IPCW net benefit per participant. Thresholds are fixed before evaluation, not selected to favor a model.

Threshold	Foundation	Compact	No history	Refer all	Refer none
0.05	0.06447	0.06372	0.06187	0.05146	0.00000
0.10	0.03957	0.03987	0.03750	-0.00155	0.00000
0.15	0.02640	0.02677	0.02371	-0.06081	0.00000
0.20	0.01493	0.01456	0.01135	-0.12747	0.00000

Table 47: Compact frozen-risk calibration bins from the same external records. Binning changes the summary, not the underlying predictions.

Binning	Bin	Observed n	Mean predicted	IPCW observed
equal width	1	3262	0.0397	0.0420
equal width	2	1045	0.1423	0.1406
equal width	3	491	0.2429	0.2215
equal width	4	205	0.3426	0.2827
equal width	5	77	0.4453	0.3695
equal width	6	30	0.5316	0.5131
equal width	7	20	0.6491	0.5739
equal width	8	6	0.7455	0.8158
equal mass	1	494	0.0085	0.0020
equal mass	2	514	0.0163	0.0159
equal mass	3	520	0.0253	0.0260
equal mass	4	503	0.0373	0.0404
equal mass	5	518	0.0541	0.0407
equal mass	6	516	0.0753	0.0927
equal mass	7	527	0.1048	0.1166
equal mass	8	510	0.1453	0.1200
equal mass	9	512	0.2086	0.2419
equal mass	10	522	0.3598	0.2923

J.1 Analytical event-precision sensitivity

A fixed minimum number of observed events does not establish power for an arbitrarily small effect. For a planning approximation, let $a = 0.77$, m be event cases and $n = 9m$ controls. Define $q_1 = a/(2-a)$, $q_2 = 2a^2/(1+a)$ and $v = [a(1-a)+(m-1)(q_1-a^2)+(n-1)(q_2-a^2)]/(mn)$. With assumed paired-score covariance represented by correlation ρ, use $\mathrm{SE}_\Delta = \sqrt{2(1-\rho)v}$, interval width $3.92\mathrm{SE}_\Delta$, and two-sided normal power $\Phi(-1.96 - \Delta/\mathrm{SE}_\Delta) + 1 - \Phi(1.96 - \Delta/\mathrm{SE}_\Delta)$. This is an explicitly approximate uncensored working design calculation, not a sample-size guarantee for time-dependent competing risks. The table below is an analytical sensitivity calculation, separate from the participant-bootstrap intervals of the completed external evaluation.

Table 48: Working paired-AUC precision scenarios. Cases, control ratio, score correlation and effect are assumptions. These are analytical precision scenarios, not the observed clinical confidence intervals.

Events	Assumed correlation	AUC effect	95% width	Power%
54	0.80	0.003	0.0959	5.2
54	0.80	0.010	0.0959	6.9
54	0.80	0.020	0.0959	12.9
150	0.80	0.003	0.0575	5.5
150	0.80	0.010	0.0575	10.5
150	0.80	0.020	0.0575	27.6
300	0.80	0.003	0.0407	6.0
300	0.80	0.010	0.0407	16.1
300	0.80	0.020	0.0407	48.7
600	0.80	0.003	0.0288	6.9
600	0.80	0.010	0.0288	27.6
600	0.80	0.020	0.0288	77.8
1200	0.80	0.003	0.0203	8.9
1200	0.80	0.010	0.0203	48.7
1200	0.80	0.020	0.0203	97.1
3000	0.80	0.003	0.0129	15.0
3000	0.80	0.010	0.0129	86.2
3000	0.80	0.020	0.0129	100.0

K RELEASE VALIDATION AND FAILURE HANDLING

The 1,200 consecutive D4 query participants and the independent release cohorts have distinct roles. The quality gate allows one retake and withholds both disease outputs after repeated failure. Candidate updates are assessed on a separate release population. Additional C3 cases improve sensitivity precision but do not estimate population prevalence or PPV.

The clinical release test requires a validation macro-F1 drop no greater than 1.0 percentage point relative to the prior model, a one-sided 95% exact binomial lower bound exceeding 0.95 for C3 sensitivity, and predeclared numerical and operational checks. The 200 additional C3 records improve sensitivity precision but cannot estimate PPV or prevalence. A 200-person enriched sensitivity result is never pooled into the 400-person consecutive release stream. At most one prospectively prespecified candidate per domain uses this cohort for final release testing. A failed candidate remains a failure. Later retesting requires a new untouched cohort or is labeled development only. This gate is a research release criterion, not regulatory clearance.

Table 49: Exact sensitivity-subgate results from independent 200-case release cohorts. Point sensitivity above 95% does not ensure its lower bound passes. Other clinical and hardware gates are not declared passed by this table.

Domain	TP	FN	n	Point%	One-sided 95% lower	Sensitivity subgate
D1	197	3	200	98.50	96.17	pass
D2	196	4	200	98.00	95.48	pass
D3	192	8	200	96.00	92.90	FAIL
D4	198	2	200	99.00	96.89	pass

L ENCODER QUANTIZATION, SIGNED HOST EXECUTION AND MEMORY

Only the shared frozen encoder is assigned to the AKD1000 accelerator. The deployment follows an AKD1000-compatible convolutional pipeline with 8-bit input-layer weights and 4-bit subsequent weights and activations, consistent with prior AkidaNet quantization and AKD1000-compatibility work (Jeeva & Miriala, 2026). Signed adaptation, task heads, calibration and report handling execute

on the CM4 host. The operator boundary is the same for the main and supplementary execution comparisons.

The primary encoder is quantized once. Its dequantized output is the immutable base coordinate system used for the basis, history and replay. The signed adapter and heads execute in host FP32. Updating stage factors U and V neither reruns encoder QAT nor changes history-feature interpretation. The formal identities hold in exact arithmetic. The FP32 residual and logit deviation are tested numerically, while encoder quantization error relative to the floating encoder is a separate comparison. Software emulation alone does not establish physical encoder energy or hardware-only execution.

Table 50: Operator placement and numerical release checks.

Operator	Executor/precision	Verification
Source preprocessing and quality	CM4 CPU, deterministic preprocessing and FP32 quality.	Source/crop hash and identical quality decision for a fixed version.
Shared two-view encoder	AKD1000,8-bit input/first weights,4-bit internal weights/activations.	Encoder-only hardware map, no simulator fallback.
Stage residual	CM4 FP32 low-rank evaluation or folded signed readout.	Relative protected-coordinate drift below 1 e-5 on fixed calibration vectors.
Stage readout	CM4 FP32, W-prime equals W times A.	Maximum folded/unfolded probability deviation at most 1 e-6 in the reference test.
Risk head	CM4 FP32, fixed 516-to 64-to 12 parameters and fixed calibrator.	Fixed-input outputs unchanged after stage-only updates.
Encoder conversion	Post-QAT host reference versus hardware feature output.	Predictions differ at most 0.01 absolute probability on calibration, and held-out F1 conversion loss at most 1 point.
Clinical reports	CM4 host, unchanged risk version and updated stage version when accepted.	Schema, model hashes and rollback parent recorded.

Table 51: Persistent adaptation-memory arithmetic. Metadata is a declared allowance. Framework/allocator peak memory is measured separately.

Item	Bytes	KiB
Replay feature payload	131,072	128.00
Replay logits FP16	4,096	4.00
Replay labels	512	0.50
Replay metadata allowance	32,768	32.00
Stable differences FP32	262,144	256.00
U and V FP32	16,384	16.00
Basis B FP32	8,192	8.00
Total	455,168	444.50

Table 52: Software and execution configuration.

Component	Configuration
Reference runtime	Python 3.11-compatible NumPy, SciPy and PyTorch.
Accelerator target	Akida 1.0 / AKD1000; frozen encoder only.
Vendor stack	MetaTF 2.19.3 and QuantizeML 1.2.5.
Host execution	CM4 FP32 quality, signed stage adapter, task heads and calibration.
Update boundary	Only U, V change; encoder, projector, risk and calibration remain fixed.

M HARDWARE MEASUREMENTS AND PROCESSING-ENERGY ANALYSIS

The hardware tables summarize 1,000 timed processing records per execution path. Each record includes preprocessing, quality, transfer, encoder, heads and report durations, together with total latency and energy. Complete-path P95 is computed from row-wise totals, not by adding component P95 values. The main and supplementary tables report the same measurements at different numerical precision.

The processing boundary starts when the source image is available and ends after report persistence. The encoder-plus-host path includes host and accelerator energy. Camera operation, operator waiting time and image acquisition are outside this boundary. The component table reports the encoder-plus-host breakdown. The idle-time table is a conditional calculation using the declared 3 W common host-idle assumption, not an additional set of measured clinical visits.

Table 53: Complete hardware processing measurements under the shared image-to-report boundary.

Path	Median ms	P95 ms	Mean J	Peak MiB	Trace rows
CM4 FP32	1082.40	1265.00	5.0104	768	1000
CM4 INT8	679.03	787.34	3.2746	432	1000
AKD encoder + host heads	563.51	650.44	2.1498	224	1000
Transfer stress	727.94	840.17	2.9246	240	1000

Table 54: Component and total latency quantiles for the measured encoder-plus-host path. Component quantiles are not additive.

Stage	Median ms	P95 ms
Preprocessing	114.64	134.47
Quality	32.73	38.31
Transfer	44.67	52.87
Encoder	279.35	329.56
Heads	46.91	55.37
Report	44.79	52.69
Total from full traces	563.51	650.44

Table 55: Conditional cycle-energy calculation with 3 W common host idle power. Savings use the higher-precision processing energies in Table 53, before display rounding. These are not measured field visits.

Idle seconds	CPU INT8 J	Encoder+host J	Saving%
0	3.275	2.150	34.35
2	9.275	8.150	12.13
5	18.275	17.150	6.15
10	33.275	32.150	3.38
20	63.275	62.150	1.78
30	93.275	92.150	1.21
60	183.275	182.150	0.61

N NUMERICAL CHECKS AND ILLUSTRATION PROVENANCE

The numerical ledger checks the projection identities, likelihood, metric arithmetic and stage–risk separation. These tests complement, rather than replace, the retinal-data and device evaluations.

Figures 1–3 link the workflow, adaptation mechanism and execution graph. Photographs are illustrative source assets, and matrix textures are schematic.

Table 56: Numerical test ledger.

Check	Outcome	Check	Outcome
two-sided invariance rank 4	PASS	partition sum D3	PASS
two-sided invariance rank 8	PASS	partition sum D4	PASS
two-sided invariance rank 16	PASS	partition sum E1	PASS
two-sided invariance rank 32	PASS	partition sum E2	PASS
one-sided counterexample	PASS	threshold true positives monotone	PASS
rank approximation theorem 0	PASS	threshold false positives monotone	PASS
rank approximation theorem 15	PASS	query rows	PASS
rank approximation theorem 30	PASS	confusion accuracy	PASS
rank approximation theorem 60	PASS	factorial identity	PASS
rank approximation theorem 90	PASS	reference output shapes	PASS
horizon monotonicity	PASS	stage update risk invariance	PASS
competing probability mass	PASS	stage actually changes	PASS
CIF integration	PASS	trainable factor count	PASS
interval likelihood	PASS	host folding equivalence FP32	PASS
participant allocation	PASS	partial censor exposure	PASS
partition sum D0	PASS	administrative survivors included	PASS
partition sum D1	PASS	event/control exclusivity	PASS
partition sum D2	PASS	persistent state within 1 MiB	PASS

Illustration attribution. Fundus photographs: Mikael Häggström, *Fundus photograph of normal right eye* and *Fundus photograph of normal left eye*, and Ktkvtsh, *Fundus geographic atrophy*, via Wikimedia Commons (CC0 1.0). Hardware photograph: Naveen Kumar / Edge Impulse Expert Network, *Multi-camera Video Stream Inference—Brainchip Akida Neuromorphic Processor*, reproduced without alteration under CC BY-ND 4.0; original source. These external images illustrate the diagrams and are not participant data from RetinaPack.

O Additional analyses and evaluation summary

The extended analyses examine protected-basis controls, stronger baselines, model-release histories, D4 workflow evaluation, paired uncertainty and execution cost. They connect the primary results to four complementary analyses: same-estimand paired inference, retained-model behavior after rejection, participant-linked D4 evaluation and direct-readout attribution. Clinical evaluation, controlled feature-space experiments, analytical calculations and hardware measurements retain their respective endpoints and denominators.

Reporting convention. Tables prefixed D contain extended results, while tables prefixed R contain reference-control and consistency analyses. The prefixes identify table series rather than evidence status. Mathematical identities and numerical checks are identified by their captions. The primary staging result is 83.78 on 920 queries. Table 3 reports candidate-only F1 of 82.20 with 1.07-point forgetting, while the independent D3 sensitivity result remains 192/200.

Analysis scope. Primary continual-learning inference uses the candidate predictions corresponding to Table 3. The reference feature-space track isolates parameterization and release-policy behavior under its specified projector and bypass check. D4 comparisons use paired predictions from the same evaluated encounters. These analyses are reported separately because a common sample count alone does not imply identical model settings or evaluation endpoints.

Table D1: Organization of additional evaluations R01-R07.

ID	Evaluation question	Analysis	Location
R01	Clinical basis versus generic spans	Random, PCA, stable-PCA, gradient-weighted, shuffled and no-repeat controls, a rank-three LDA family and an independent lesion test.	Appendix P.
R02	PEFT and full-tuning staging	RETFound, VisionFM and RetiZero under each tuning regime with separate model selection.	Appendix R.
R03	Allocation versus staging test size	Task-specific denominators separate the 13,040-participant allocation from the 920 staging queries.	Appendix U.
R04	Candidate versus retained model	Reference candidate and retained histories, all eight release gates and the original D3 failure.	Appendix S.
R05	D4 query evaluation	Linked workflow accounting, paired predictions, uncertainty and relative chronology.	Appendix T.
R06	Predictive and deployment trade-offs	Stronger baseline regimes and separately scoped risk analyses retain unfavorable comparisons.	Appendices R and V.
R07	Paired continual inference	Same-estimand paired inference and input checks with one participant cluster across all runs.	Appendix U.

O.1 Mechanism, robustness, execution and availability

Table D2: Organization of additional evaluations R08-R13.

ID	Evaluation question	Analysis	Location
R08	Algorithmic mechanism	Direct MQz control, minimum-norm lift, induced penalties, spectra and finite-budget attribution.	Appendix Q.
R09	Basis settings	Protected rank, repeat weight, shrinkage, clipping and pair-window sensitivity with fixed residual capacity.	Appendix P.
R10	Quality evidence	Failure AUROC, sensitivity and specificity, retake recovery, six failure types, unsafe passes and rater agreement.	Appendix T.
R11	Energy boundary	Processing, acquisition and idle energy are distinguished from local update, release evaluation and cold start.	Appendix W.
R12	Reproducibility checks	Basis and adapter specification, cached-feature optimization, paired scoring and numerical tests.	Reproducibility Statement.
R13	Governed benchmark access	Manuscript and code availability distinguished from conditional data release under custodian and ethics review.	Ethics Statement.

Table R1: Analysis map and associated records. Each comparison uses its stated prediction source, model state, denominator and evaluation settings.

Analysis	Component	Associated records
1. Paired statistics	One prediction tensor generates means, paired differences, intervals and tests (Appendix U).	Exact original-run prediction join, checkpoint IDs and regenerated tables.

Table R1 (continued)

Analysis	Component	Associated records
2. Retained state	Optimize from retained parents; preserve teacher and optimizer-reset semantics (Appendix S).	All candidate and retained checkpoints, release inputs and predictions for every order/seed.
3. D4 provenance	Explicit paired IDs, source-matrix arithmetic, and relative-time audit (Appendix T).	Original paired probabilities and dated protocol, acquisition, adjudication and release records.
4. Attribution	Direct logit parameterization and induced-loss control (Appendix Q).	Actual frozen W, Q, matched optimization runs and held-out feature/error diagnostics.

P CLINICAL-SUBSPACE CONTROLS AND SENSITIVITY

All protected bases are estimated from immutable post-QAT D0-training features. The controls share the encoder, frozen readout, support IDs, replay candidates, stable-pair labels, optimizer, loss coefficients and seeds. Residual rank is eight with 4,096 exposed factors. Two-sided projection is used unless another operator is named. The one-sided map is $I + UV^\top Q$ and the output-only map is $I + QUV^\top$.

Table D3: Protected-basis and operator comparisons. The first two rows match Table 3. SD is trajectory dispersion rather than participant-level uncertainty.

Basis / operator	Final F1	SD	Forget.	Status
Two-sided clinical	82.20	1.31	1.07	Original
One-sided clinical	81.25	1.62	1.36	Original
Output-only clinical	81.48	1.49	1.32	New
Random orthogonal	79.94	1.66	1.68	New
Base-feature PCA	80.26	1.58	1.54	New
Stage LDA 3 + PCA 5	80.61	1.52	1.34	New
Stable-difference PCA	79.92	1.61	1.47	New
Gradient-weighted activation	80.76	1.47	1.21	New
Within-block shuffled labels	80.06	1.63	1.60	New
Clinical, no repeat moment	81.53	1.41	1.26	New
Clinical, centered moments	80.71	1.55	1.49	New
Clinical, unit-norm differences	80.43	1.61	1.55	New

Generic controls. Feature PCA uses centered D0 feature covariance. Stable PCA uses the uncentered stable-difference second moment. The gradient-weighted activation control takes the top eight eigenvectors of the uncentered moment of $h_i = z_i \odot \nabla_{z_i} \mathcal{L}_{\text{stage}}$. This is not a rank-eight covariance of raw frozen-head gradients. The shuffled control permutes change and stable labels within matching blocks and recomputes the moments without changing acquisition metadata. The hybrid combines three LDA directions with five orthogonalized PCA directions.

Table D4: Equal protected rank of three with residual rank eight. Four-class LDA is not assigned rank eight.

Rank-3 protected basis	Final F1	Forget. pp
Random	79.78	1.72
PCA	80.07	1.64
Stage LDA	80.39	1.52
Clinical	81.42	1.23

The clinical and generic spans have the same information access, residual budget and optimization settings. Comparable performance from random or shuffled spans would weaken a clinical-specificity explanation. Trajectory SD alone does not establish participant-level significance.

P.1 SENSITIVITY UNDER FIXED RESIDUAL CAPACITY

Table D5: Protected-rank sensitivity with residual rank eight. Rank zero reproduces factorial cell 0111 and rank eight reproduces cell 1111.

Protected rank	Final F1	Forget. pp	Status
0	80.21	1.55	Original
4	81.75	1.24	New
8	82.20	1.07	Original
16	82.37	1.03	New
32	81.62	0.97	New
64	79.91	0.91	New

Table D6: One-factor sensitivity around repeat weight 0.5 and shrinkage 0.1. The reference result remains 82.20 F1.

Varied coefficient	Setting	Final F1	Forget. pp
Repeat weight	0.0	81.53	1.26
Repeat weight	0.25	81.94	1.15
Repeat weight	1.0	82.04	1.11
Repeat weight	2.0	81.46	1.19
Shrinkage	0.0	81.48	1.35
Shrinkage	0.05	82.06	1.15
Shrinkage	0.2	82.11	1.12
Shrinkage	0.4	81.72	1.27

Table D7: One-factor sensitivity around the 95th-percentile norm clip and 30–180-day pair window. The matched analysis uses equal eligible pair counts with other settings fixed.

Varied factor	Setting	Final F1	Forget. pp
Norm clip quantile	0.90	81.89	1.15
Norm clip quantile	0.99	82.02	1.12
Norm clip	None	81.31	1.29
Pair window, days	30–90	81.67	1.24
Pair window, days	30–120	82.03	1.15
Pair window, days	30–365	80.92	1.32

Changing the pair window changes eligibility. The fixed-count analysis subsamples each group to the smallest eligible group size before estimating any basis, maintaining acquisition matching and supervision access. The common clipping radius is fitted to pooled change and stable training differences and applied unchanged to repeats. Protected-rank and coefficient sweeps are descriptive analyses. Rank eight remains the primary configuration rather than being replaced by the highest-scoring sweep setting.

P.2 INDEPENDENT LESION SEMANTICS AND NUISANCE LEAKAGE

The independent extension contains 120 adjudicated change-pair participants, 120 stable-pair participants and 120 crossed-repeat participants. These participants are disjoint from the original training, selection and query roles. The scoring rule and threshold are selected on development data and frozen before evaluation. The lesion target is adjudicated change rather than a predicted stage difference.

Table D8: Independent lesion-change results. Stable specificity uses a development-locked threshold targeting 90% lesion-change sensitivity. The table reports point estimates.

Protected basis	Lesion AUC	Repeat energy η_N	Stable specificity
Random	0.619	0.071	0.540
Feature PCA	0.714	0.224	0.675
Stage LDA + PCA	0.749	0.137	0.701
Shuffled clinical	0.640	0.068	0.572
Clinical, no repeats	0.839	0.103	0.780
Clinical + repeats	0.866	0.034	0.807

For a candidate projector P, the protected change score is $s_P(\delta) = \|P\delta\|_2^2$. Nuisance leakage is

$$\eta_N(P) = \frac{\sum_i w_i \|P\delta_i^{\text{repeat}}\|_2^2}{\sum_i w_i \|\delta_i^{\text{repeat}}\|_2^2}, \qquad w_i = 1/6, \tag{8}$$

where six contrasts belong to each repeat participant. All contrasts from a participant stay together in resampling. A zero denominator requires explicit handling. Leakage is interpreted jointly with protected rank because retaining fewer directions can itself reduce projected repeat energy.

Basis stability and scope. Pair resampling is performed within matching blocks before refitting the basis. Within-seed angles use one coordinate system. Cross-seed comparison requires Procrustes alignment fitted to shared D0-training anchors. The relevant diagnostics are eigenvalues, eigengaps, numerical rank, principal angles and pair-label sensitivity. Unaligned coordinates and test-selected ranks are not valid substitutes. Clinical-minus-random, feature-PCA, shuffled, no-repeat, one-sided and output-only comparisons form the six-contrast specificity family. A useful clinical span must predict independent lesion change, limit repeated-acquisition leakage and support adaptation. Stage F1 alone does not establish all three.

Q Analytical mechanism and direct-readout attribution

The frozen risk head bypasses the stage adapter. Its fixed-input invariance is a property of this isolation, shared by every staging method using the same bypass, frozen encoder, risk head and calibrator. No risk-invariance number is used to favor a protected basis or staging optimizer. Stage adaptation, independent lesion semantics, nuisance leakage and finite-budget optimization are separate tests.

Q.1 Capacity equivalence and a stronger matched control

Let $W \in \mathbb{R}^{4\times 256}$ be the frozen stage readout and $Q = I - P$. If $A = WQ$ has full row rank, every $L = LQ$ is realizable by

$$R_*(L) = A^\dagger L, \qquad QR_*Q = R_*, \quad WR_* = L, \quad \operatorname{rank}(R_*) \le 4. \tag{9}$$

Indeed, the columns of $A^\dagger$ lie in $\operatorname{range}(QW^\top)$, so $QR_* = R_*$. Right invariance follows from $LQ = L$, and $AA^\dagger = I_4$ proves $WR_* = L$. A rank-eight factorization can represent this rank-at-most-four map by padding a factorization with zero columns. The corresponding direct control is

$$g_M(z) = Wz + MQz + b, \qquad M \in \mathbb{R}^{4\times 256}. \tag{10}$$

It exposes 1,024 matrix entries, compared with 4,096 factor entries in U, V. Since $\operatorname{rank}(Q) = 248$, MQ has at most 992 linear degrees of freedom; centering four logits leaves at most 744 probability-relevant degrees. These are parameter-space counts, not measured accuracy. Full four-row rank is sufficient rather than necessary: the rank of $(I_4 - \mathbf{1}\mathbf{1}^\top/4)WQ$ is also recorded.

Matching the feature-space penalties. Comparing only an unregularized M against Equation 4 confounds parameterization and regularization. Let $L = MQ$ and $G = (AA^\top)^{-1}$. The minimum-norm lift satisfies

$$\|R_*\|_F^2 = \operatorname{tr}(L^\top G L), \qquad \|R_*\delta\|_2^2 = (L\delta)^\top G (L\delta). \tag{11}$$

Use these two quantities in the residual and stable-pair terms, keeping support/replay cross-entropy, ordinal loss, KL teacher, batches, initialization output and selection protocol fixed. This defines the *induced-metric direct control*, not a logit-only ablation with omitted losses. A raw-direct control with $\|MQ\|_F^2$ and no feature displacement penalty is reported separately.

For any other admissible R with $WR = L$, write $R = R_* + Z$. Then $WZ = 0$, $QZQ = Z$, and each column of R_* is orthogonal to each column of Z. Consequently, both $\|R\|_F^2 = \|R_*\|_F^2 + \|Z\|_F^2$ and $\|R\delta\|_2^2 = \|R_*\delta\|_2^2 + \|Z\delta\|_2^2$. The lift cannot worsen the displayed feature penalties at fixed logits. This statement concerns the objective, not equality of finite-step AdamW trajectories or factor-decay effects. It rules out a strict function-class necessity claim under the stated rank condition.

Q.2 Readout spectra and numerical lift validation

The spectra below are computed from the frozen readouts and projectors used in the reference feature-space track. This track isolates the algebraic and optimization properties of the readout parameterizations rather than re-estimating the clinically selected basis. The five seeds are listed explicitly in Table R2. Singular values and condition numbers identify the applicability of the full-row-rank condition to these checkpoints.

Table R2: Frozen WQ singular values for the five seeds in the reference feature-space experiment. The listed seed set is specific to this control track.

Seed	σ_1	σ_2	σ_3	σ_4	Condition
11	1.21199	1.12035	1.07337	0.93427	1.297
22	1.20247	1.08395	1.03385	0.90805	1.324
33	1.21232	1.12224	1.01529	0.96894	1.251
44	1.20661	1.13769	1.06996	0.96232	1.254
55	1.17963	1.12909	1.06881	0.92130	1.280

Table R3: Numerical lift checks for the reference seed-11, order-123 checkpoint on 920 feature records. The copied-logit lift is an algebraic reconstruction of this checkpoint, not a separately trained model.

Quantity	Computed value
Maximum logit gap	1.33e-15
Maximum softmax gap	3.33e-16
Changed argmax predictions	0
Lifted residual rank	4
$\|PR_*\|_F$	4.71e-18
$\|R_* - QR_*Q\|_F$	4.71e-18
Induced-norm identity error	1.78e-15

Rank and identity edge cases. The raw feature gradient of a shift-invariant four-logit loss lies in a space of dimension at most three. A rank-eight gradient-weighted *activation* span is therefore not described as eight raw-gradient directions. The one-sided and output-only counterexamples in Appendix F remain valid. If the measured clinical WQ is rank-deficient, Equation 9 applies only to L whose columns lie in its image; truncating small singular values changes the comparator and must be disclosed.

Q.3 Finite-budget parameterization and error patterns

The reference controls are optimized on 256-dimensional feature records with 40 balanced support examples per domain, 512 replay records, six domain orders and five seeds. They share the readout, feature inputs, sampled batches and candidate or retained endpoint. The reference projector protects the first eight coordinates and is fixed rather than learned from lesion pairs. This controlled setting isolates parameterization and finite-budget optimization from clinical basis estimation, so its F1 values are not substitutes for the primary Table 3 results.

Table R4: Candidate-only results in the reference feature-space track. F1 is domain-balanced and SD summarizes 30 trajectories. The 500-step condition tests optimization sensitivity rather than a query-selected operating point.

Parameterization	Steps	Entries	F1	SD	Forget.	Max $\|PR\|_F$
Two-sided factors	100	4096	90.69	1.38	1.11	0.00e+00
One-sided factors	100	4096	90.29	1.41	2.10	3.05e+00
Direct MQ, raw penalty	100	1024	90.69	1.05	1.17	0.00e+00
Direct MQ, induced metric	100	1024	90.69	1.05	1.17	4.36e-17
Direct MQ, induced metric	500	1024	88.93	1.02	3.72	1.44e-16

Table R5: Paired correctness counts for reference seed 11 and order 123 on the same 920 records. Each row sums to 920 across its four correctness cells.

Control	Both right	2S only	Control only	Both wrong	Disagree
1S100	812	7	13	88	23
MQ100	805	14	12	89	32
Mlift100	805	14	12	89	32
Mlift500	783	36	26	75	73

Scope of clinical attribution. The direct induced-metric, one-sided and two-sided controls share cached features and the frozen readout. Readout spectra, held-out logits, candidate and retained F1, paired errors and projected feature drift characterize their differences. Independent lesion and repeated-acquisition analyses in Appendix P address the separate question of clinical-span selection. Matched-step and longer-optimization comparisons answer different questions. Factor decay and direct-coordinate decay have different implicit effects, so equal nominal decay does not imply identical regularization. The 500-step result is a sensitivity condition, not a convergence certificate.

If the induced-metric direct control matches two-sided performance, the valid attribution is clinical-span selection plus an efficient parameterization or optimization trade-off, not an exclusive classifier family. A two-sided advantage at matched finite budget supports a finite-budget effect only. A pathology-protection benefit requires the independent lesion and acquisition tests and cannot be inferred from a zero risk-output change or a numerical rank check.

R FOUNDATION TUNING AND NATIVE CONTINUAL-LEARNING COMPARISONS

RETFound, VisionFM and RetiZero use separate PEFT and full-tuning staging regimes with the D0 train, tune and calibration partitions unchanged. Each backbone retains its supported native resolution and the same clinical labels. Larger models are not constrained to the compact model's parameter or compute budget. Each regime uses 24 tuning trials and five confirmation seeds. Query labels do not select checkpoints, augmentations, thresholds or hyperparameters.

Table D9: Foundation-model staging regimes. Compact and probe rows reproduce Table 2. Additional rows use the stated PEFT or full-tuning protocol.

Backbone	Regime	Stage F1	C1 F1	Status
Compact two-sided	As reported	83.78	69.16	Original
RETFound	Probe	83.37	69.33	Original
RETFound	PEFT	85.04	73.26	New
RETFound	Full	85.62	74.91	New
VisionFM	Probe	83.56	69.84	Original
VisionFM	PEFT	85.36	74.18	New
VisionFM	Full	86.08	75.80	New
RetiZero	Probe	83.57	69.81	Original
RetiZero	PEFT	85.29	73.88	New
RetiZero	Full	85.84	75.21	New

Table D10: Native ViT-B/16 continual-learning track with compatible pretrained initialization. This track is not an equal-byte or equal-compute attribution test.

Native comparison	Final F1	Forget. pp	Extra state / access
Image-replay DER++	84.06	1.41	Images
Native GPM	83.74	1.06	Activation memory
Native BiLoRA	84.29	1.12	Native factors
Native CODA-Prompt	83.92	1.27	Prompt pool

The native methods use architecture-level implementations on a compatible ViT-B/16 initialization. DER++ may retain images. GPM activation memory, BiLoRA factors and prompt pools are accounted for separately. These methods provide attainable alternatives rather than equal-byte attribution. The compact FA family remains the controlled feature-space comparison.

R.1 Selection and interpretation

Table D11: Configuration and evidence fields for the tuning regimes.

Item	Procedure	Recorded fields
Model identity	Exact checkpoint, native resolution and CFP-only input. Text use is declared where applicable.	Checkpoint digest, license and preprocessing configuration.
Search	Four learning rates, three weight decays and two augmentation strengths on D0 tune.	All 24 trial scores, completion or failure status, elapsed time and selection rule.
Confirmation	Five seeds with at most 80 epochs and patience 10.	Per-seed checkpoint, training curve, stopping reason and tune score.
Calibration	Fit on D0 calibration after model selection and freeze before queries.	Temperature, referral threshold and calibration-participant digest.
Query reporting	The same 920 participants for the staging snapshot.	Probability vectors, confusion matrix, classwise metrics and paired intervals.
Native CL	The same domain orders and support labels with replay and information access disclosed.	Candidate predictions, persistent bytes, parameters and training cost.
Cost comparison	Execution device and measurement boundary are declared.	Timing and power traces only for measured paths.

PEFT and full tuning outperform the frozen probes and compact model on staging. The deployment question concerns accuracy, adaptation and resource use under the stated information and hardware conditions, not universal predictive superiority. Three backbones and two added regimes use 24 search trials and five confirmation seeds per regime. The native continual-learning track comprises

120 trajectories. A redesigned encoder, new supervision or enlarged support set is a separately identified variant rather than a replacement for the primary RetinaLPA-2S row.

S RELEASE-POLICY VALIDATION AND RETAINED-STATE TRAJECTORIES

The clinical D3 sensitivity result is 192/200, with a one-sided 95% exact lower bound of 92.90%, below the required 95%. That result does not authorize the candidate. Table 3 describes candidate-only trajectories. A gate-aware trajectory with a different post-rejection teacher must be evaluated from its own checkpoints rather than by relabelling the candidate-only means.

The reference release-policy experiment trains the 4,096-factor adapter, evaluates its predictions and follows the resulting checkpoint transitions. It uses the feature-space configuration in Appendix Q, including its fixed reference projector and separate bypass check. Candidate-only and retained-state trajectories are evaluated within that track. Their absolute F1 values describe this control setting, not the primary clinical configuration or an authorization for patient-facing deployment.

S.1 TWO SEPARATELY EXECUTED PATHS

The candidate-only path always advances its candidate and corresponding teacher. The gate-aware path starts each domain from the preceding retained U, V and teacher, resets Adam moments, performs 100 steps and freezes the candidate before validation. It updates the retained state only after the gate decision. Otherwise the parent parameters and teacher remain byte-identical. Later updates are genuinely re-optimized from the retained parent. Query labels never select the candidate or the gate.

Table D12: Candidate-only and retained-state trajectories in the reference feature-space track. Retention is conditional on the specified gate inputs. The primary candidate F1 of 82.20 and forgetting of 1.07 in Table 3 refer to the separate clinical configuration.

Path	Final F1	SD	Forget.	Runs	Accept / try
Reference candidate-only	90.69	1.38	1.11	30	–
Reference retained	90.73	1.39	0.61	30	54/90

Table D13: Order-level summaries over the five reference seeds. Retained-minus-candidate F1 can be positive when rejection prevents a harmful update. Neither endpoint is a universal upper bound on the other.

Order	Candidate	Retained	Δ	Ret. forget.	Accept / try
123	91.05	90.97	-0.08	0.36	10/15
132	91.12	90.97	-0.16	0.65	10/15
213	90.75	90.50	-0.24	0.45	8/15
231	90.57	90.50	-0.07	0.75	8/15
312	90.17	90.97	0.80	0.66	10/15
321	90.50	90.50	-0.00	0.76	8/15

S.2 A REJECTION BEFORE LATER ADAPTATION

The displayed trajectory uses order 312 and seed 11 to show a rejection before subsequent adaptation. Tables D14 to D16 give its scores and gate inputs. Table R6 identifies the parameter states, and Table R7 summarizes all 90 attempts in this reference track. Subsequent candidates are optimized from the retained parent rather than continued from a rejected model.

Table D14: Predictions from the optimized reference checkpoints. A rollback row evaluates the exact retained parent on unchanged query inputs.

Step	Domain	State	D0	D1	D2	D3	Mean
1	D3	candidate	94.29	82.62	76.74	93.59	86.81
1	D3	retained	94.26	83.62	79.23	90.76	86.97
2	D1	candidate	95.39	91.47	75.44	91.28	88.40
2	D1	retained	95.39	91.47	75.44	91.28	88.40
3	D2	candidate	95.97	87.07	79.69	92.14	88.72
3	D2	retained	95.97	87.07	79.69	92.14	88.72

Table D15: Model-side gate values from validation records disjoint from support and queries. The domain-3 condition includes acquisition stress on eight C3 feature records. Its detections, errors and lower bound are evaluated after inference.

Step	Domain	TP / n	Lower %	Δ F1	Drift	Decision
1	D3	183/200	87.52	2.29	0.00e+00	V-ROLLBACK
2	D1	197/200	96.17	7.39	0.00e+00	V-ACCEPT
3	D2	196/200	95.48	4.24	0.00e+00	V-ACCEPT

Table D16: Inputs to all eight release gates and the corresponding deployment requirements. The reference state-transition test and physical deployment authorization have separate scopes.

Gate	Reference evaluation	Deployment requirement
Stage drop ≥ -1 pp	Computed from 400 validation records	Frozen candidate and parent on identical consecutive release IDs.
C3 exact lower bound > 0.95	Computed from 200 C3 validation records	Separate prespecified independent enriched release stream.
Protected drift $< 10^{-5}$	Computed from reference parameters	Original feature coordinates and actual precision.
Frozen risk / version	Fixed sigmoid bypass and parameter hash	Same real inputs; risk/encoder/calibration hashes and outputs.
Encoder hardware-only mapping	Mapping-status input to release policy	Physical encoder map log and hardware identifier.
Model/configuration checksum	Parameter hashes and configuration-status input	Verified artifact/configuration digests.
Update time ≤ 300 s	Resource input to release policy	Timed local optimization on the intended host.
Peak memory ≤ 2 GiB	Resource input to release policy	Measured peak memory for the same update boundary.

A fixed four-output sigmoid bypass on immutable features provides the reference invariance check. This numerical control is separate from the clinical competing-risk network. Its outputs and parameter digest are compared before and after every update. The raw-direct MQz control leaves its input features unchanged by definition, so zero feature drift alone is not evidence of learned pathological protection.

Release interpretation. The reference ledger distinguishes policy acceptance (`V-ACCEPT`) and rollback (`V-ROLLBACK`) from physical deployment authorization. Its acceptance decision is conditional on the complete set of supplied gate inputs, while the physical-release field remains `NOT_AUTHORIZED`. A missing required input leads to hold or rollback rather than an assumed pass. Patient-facing release additionally requires the corresponding checkpoint and predictions, measured host resources, hardware mapping evidence and verified version artifacts. No physical release is inferred from a state-transition check.

Table R6: Truncated SHA-256 identifiers for the reference parameter states. Matching identifiers expose teacher, parent and rollback relationships without counting a retained parent as a newly trained checkpoint.

Step	Parent	Teacher	Candidate	Retained
1	45fa3c0865	45fa3c0865	cb262255ca	45fa3c0865
2	45fa3c0865	45fa3c0865	c078bef7ac	c078bef7ac
3	c078bef7ac	c078bef7ac	9fcec3b112	9fcec3b112

Table R7: Whole-chain accounting for the reference release-policy evaluation. Repeated evaluations do not create additional independent participants or authorize physical deployments.

Audit endpoint	Result
Orders / seeds / attempted updates	6 / 5 / 90
Policy accepted / rejected	54 / 36
Rejection fraction	40.00%
Post-rejection next-update evaluations	26
Mean new-domain candidate gain after prior rejection	5.65 pp
Teacher equals retained parent; checked transitions	90 / 90
Incorrect parent or rollback checksum transitions	0
Physical releases authorized by this evaluation	0

S.3 Original failure and separately reported D3B result

The D3B extension is a separate clinical release evaluation. Its support, release and query roles are distinct from the failed D3 cohort and from the reference control histories above. A successful D3B result does not retroactively approve the original D3 candidate. The two evaluations retain separate candidate identities and validation streams.

Table D17: Separate D3B release result. This extension does not supply the original multi-order retained-state trajectories.

Separate release	TP / n	95% lower	Δ F1	Sec.	MiB
D3B	198/200	96.89	1.80	68	392

Table D18: Remaining D3B gate inputs and joint decision. The original D3 decision is unchanged.

New cohort	Drift	Risk/version	Map	Hash	Gates	Decision
D3B	2.9e-07	0 / SAME	PASS	MATCH	8 / 8	ACCEPT

Clinical retained-state evaluation. Clinical retained-state evaluation requires the original encoder features and the five-seed, six-order checkpoint manifests. Candidate-only and gate-aware endpoints are exported separately. After a rejection, subsequent steps start from the retained parent with reset optimizer moments. Each method, order and seed must link the parent, teacher, candidate, retained state, support and replay records, release inputs and query probabilities. Final retained F1, forgetting, acceptance counts and later-domain adaptation are computed from those linked histories. Reusing a fixed research release cohort across trajectories does not constitute 90 independent clinical deployments or permit repeated selection on its labels.

T D4 WORKFLOW RESULTS AND PAIRED EVALUATION

The D4 evaluation comprises 1,200 encounters. Workflow counts, staging predictions, quality dispositions and reliability bins are linked within this cohort. The zero-shot and adapted states are compared on the same 1,150 gradable automatic outputs, while all-encounter workflow accounting retains ungradable outputs and manual reviews. Paired inference uses shared participant identities rather than separate marginal summaries.

The D4 record set links encounter identities, zero-shot and adapted probabilities, quality disposition and timing relative to protocol freeze. Table D21 reports the primary paired comparison and a separate re-pairing sensitivity calculation. The latter preserves marginal predictions while changing their pairing and therefore tests the dependence of interval width on joint errors. It is not a second clinical cohort or the primary significance result.

Table D19: D4 workflow accounting for 1,200 encounters. Expert-ungradable automatic passes remain visible and are excluded only from image-conditional staging.

D4 workflow item	People
Consecutive people	1200
First-capture expert ungradable	180
Retake requests	210
Initial automatic outputs	990
Automatic outputs after retake	168
Repeat failure / manual review	42
All automatic outputs	1158
Expert-gradable automatic outputs	1150
Expert-ungradable automatic outputs	8

Table D20: Staging performance on the same 1,150 gradable automatic outputs before and after D4 adaptation. Candidate and retained predictions coincide under the reported release decision. ECE uses ten equal-width bins.

State	n	Macro-F1	Accuracy	κ_w	C3 recall	ECE %
Frozen zero-shot	1150	75.88	78.09	0.898	83.89	9.86
D4-like candidate	1150	83.80	85.30	0.935	89.44	9.54
D4-like retained	1150	83.80	85.30	0.935	89.44	9.54

Table D21: Paired D4 comparison and re-pairing sensitivity. The first row uses the primary participant pairing. The second is an analytical re-pairing control with unchanged marginals, not a second cohort or an alternative primary confidence interval.

Record set	Δ F1	Paired 95% CI	p	IDs
D4 primary pairing	7.93	[5.95, 9.86]	0.0005	1150
Re-pairing sensitivity	7.93	[4.47, 11.14]	–	1150

Automatic-output coverage is 96.50%, with 17.50% retake requests and 3.50% final manual review. Eight expert-ungradable automatic passes remain visible. Image-conditional staging uses 1,150 gradable outputs, while clinical escalation uses all 1,200 encounters. The confusion matrices yield 75.87586849 and 83.80139840 macro-F1, a difference of 7.92552991 points, displayed as 7.93. The primary paired interval is calculated from the joint records rather than inferred from these marginal matrices.

T.1 Confusion matrices, release decisions and calibration

Table D22: Zero-shot and adapted confusion matrices with identical reference counts. The corresponding joint prediction records determine paired uncertainty.

State	Reference	C0	C1	C2	C3	n
Zero-shot	C0	359	57	4	0	420
Zero-shot	C1	36	113	39	2	190
Zero-shot	C2	4	50	275	31	360
Zero-shot	C3	0	2	27	151	180
Adapted	C0	378	40	2	0	420
Adapted	C1	22	143	25	0	190
Adapted	C2	2	37	299	22	360
Adapted	C3	0	1	18	161	180

Table D23: D4 inputs to all eight release gates, with mapping and checksum shown together. Release records are separate from the query cohort and do not determine query probabilities.

Gate	Value	Decision
Release Δ F1, consecutive 400	3.44	PASS
C3 lower bound, 198 / 200	96.89	PASS
Protected drift	2.6e-7	PASS
Risk gap / version	0 / SAME	PASS
Encoder map / configuration	PASS / MATCH	PASS
Local-update time, seconds	65	PASS
Peak memory, MiB	384	PASS

Release F1 and C3 inputs are not calculated from query confusion matrices or used to select query probabilities. Table D23 reports passing release checks, so the candidate and retained D4 point estimates coincide. This equality depends on all gates rather than sensitivity alone. After any rejection, query evaluation instead uses the retained parent.

Top-label ECE uses ten fixed equal-width confidence bins. Calibration remains imperfect after adaptation. External calibration uses the unchanged D0 calibrator. Any locally recalibrated variant needs a separate calibration partition rather than D4 query labels.

T.2 Quality discrimination, retake recovery and unsafe passes

Table D24: Quality-failure discrimination. Failure is the positive class. AUROC uses failure scores and threshold metrics use the displayed counts. The reported quality threshold is 0.50.

Capture	n	TP / FN	TN / FP	AUROC	Sens.	Spec.	Cutoff
First capture	1200	175/5	985/35	0.974	97.22	96.57	0.50
Retake only	210	30/3	165/12	0.944	90.91	93.22	0.50

Table D25: Primary reasons for the 180 initial failures. Of 175 flagged failures, 136 recover to gradable automatic outputs. Recovery excludes initially gradable false rejections.

First-failure reason	Reference failures	Retake requested	Recovered gradable	Detection %
Blur	60	59	49	98.33
Exposure	40	39	32	97.50
Field loss	35	34	29	97.14
Media opacity	25	24	12	96.00
Obstruction	15	14	10	93.33
Unsupported file	5	5	4	100.00

Table D26: Annotation agreement calculated from the two-rater matrices. Reference-label provenance and participant clustering remain distinct from agreement arithmetic.

Grader comparison	Items	Agreement %	κ
Four-stage (quadratic weights)	400	89.50	0.956
Quality (unweighted)	100	95.00	0.875

Workflow accounting links first capture, retake and final disposition. All 210 retakes correspond to first-capture records. Of 175 correctly flagged initial failures, 136 become gradable automatic outputs. Another 29 automatic outputs follow initially gradable false rejections, rather than recovery from poor-quality images. Five first-capture false passes and three retake false passes account for the eight unsafe outputs. Each disposition retains its encounter identifier for the all-encounter analysis.

A clinical annotation audit requires pre-adjudication labels, model-masking status, disagreement categories, adjudication and re-read timestamps. The reported agreement arithmetic alone does not establish real grader qualification. Missing clinical-reference labels remain missing. All-encounter sensitivity requires a separate specialist reference rather than model-derived labels.

T.3 Whole-workflow escalation and reliability bins

Table D27: Whole-workflow escalation includes manual review after quality failure. All-encounter C3 reference uses specialist assessment rather than a model-derived stage for ungradable images.

Population	n	TP	FN	TN	FP	Sens. %	Escalated %
Gradable auto only	1150	175	5	906	64	97.22	20.78
All encounters	1200	184	6	910	100	96.84	23.67

Table D28: Nonempty reliability bins for adapted D4 predictions. Counts sum to 1,150 and query calibration is evaluated without refitting.

Confidence bin	n	Mean confidence %	Observed correct %
$[0.4, 0.5)$	17	47.95	70.59
$[0.5, 0.6)$	39	54.47	64.10
$[0.6, 0.7)$	167	67.08	86.83
$[0.7, 0.8)$	305	75.25	87.21
$[0.8, 0.9)$	319	84.96	86.21
$[0.9, 1.0)$	303	94.60	85.15

The all-encounter reference contains 190 C3 cases, comprising 180 in gradable automatic outputs, two in unsafe automatic passes and eight in manual review. Manual review counts as escalation, not

correct stage classification or completed intervention. Workflow sensitivity, argmax recall, thresholded referral and enriched-release sensitivity have different denominators.

T.4 Pairing sensitivity and chronology checks

For the primary D4 pairing, bootstrap samples reuse the same participant identities jointly and each permutation swaps a participant's complete two-model prediction vectors. The analysis uses 1,000 draws and 1,999 swaps with an add-one two-sided test. This exploratory D4 contrast is separate from the four-comparison continual-learning Holm family. The re-pairing sensitivity changes only the joint assignment, making explicit why marginal confusion matrices do not determine a unique paired interval.

Table R8: Joint correctness counts for the primary D4 pairing. The marginals reproduce 898 zero-shot and 981 adapted correct predictions among 1,150 evaluated records.

Zero-shot outcome	Adapted correct	Adapted wrong	Total
Correct	878	20	898
Wrong	103	149	252
Total	981	169	1150

Table R9: Consistency checks for the linked D4 records. Identity, shared-label, quality and checkpoint checks use the same encounter and evaluated-record sets.

Check	Result
Unique encounter IDs / image-conditional paired IDs	1200 / 1150
Duplicate IDs / inconsistent shared labels	0 / 0
Reported confusion matrices reproduced by argmax	2 / 2
Joined retakes / final manual reviews / unsafe automatic passes	210 / 42 / 8
Gradable outputs plus unsafe outputs plus manual reviews	1150 + 8 + 42 = 1200
Candidate/retained differing probability rows	0
Rows with missing image-stage label silently imputed	0

Table R10: D4 chronology expressed as relative days from protocol freeze. The corresponding protocol, capture, adjudication and model records define the ordering. Relative days are not calendar dates.

Milestone	Relative day	Associated record
Protocol / split / thresholds frozen	0	Signed protocol and immutable split/configuration version.
Dedicated support capture and label availability	1–7	Support IDs, source image times and adjudication time.
Adapted candidate frozen	8	Checkpoint and candidate hash before release labels are opened.
Independent release collection and decision	9–17	Release IDs, all gate inputs and signed decision.
Consecutive query acquisition	20–44	Enrollment and native capture timestamps, primary eye and quality links.
Query-reference unblinding and export	50–51	Masked grading/adjudication, lock and export records.

Clinical pairing and chronology. Zero-shot, frozen-candidate and retained probabilities are joined on participant, eye, landmark or capture, and model identifiers. The outer join precedes exclusions, with duplicate keys, inconsistent labels, missing pairs and mismatched states treated as errors. Confusion matrices, F1, paired intervals and referral counts use this common record set. Quality failures and missing references remain in workflow accounting. Protocol freeze precedes query acquisition, and the candidate is frozen before the release decision. Prospective deployment

additionally requires release before query inference. Prospective collection followed by retrospective rescoring is distinguished from prospective deployment.

U SAME-ESTIMAND PAIRED INFERENCE AND CONSISTENCY CHECKS

U.1 ONE ESTIMAND, ONE PREDICTION SOURCE

For model m, run r and domain d, compute F_{mrd} from the common reference labels and the named final candidate checkpoint. With fixed equal weights,

$$\bar{F}_m = \frac{1}{30}\sum_{r=1}^{30}\frac{1}{4}\sum_{d=0}^{3}F_{mrd}, \qquad \widehat{\Delta}_{0m} = \frac{1}{30}\sum_{r}\frac{1}{4}\sum_{d}(F_{0rd}-F_{mrd}) = \bar{F}_0 - \bar{F}_m. \tag{12}$$

The same run list, participant IDs, domain weights, model state, metric implementation and missingness rule are required. Pairing affects uncertainty, not this identity. Nonlinearity of macro-F1 does not change the identity after each F_{mrd} has been computed. Rounding occurs only after all reductions.

Table D29 compares the differences displayed in Table 3 with differences recomputed from the participant-level predictions. The mean difference and the mean paired difference use the same trained runs, model state, participants and equal domain weights. They therefore agree before rounding up to floating-point precision. Pairing determines the uncertainty calculation rather than changing the point-estimate definition.

The primary prediction array contains 920 query identities across 30 runs for the five methods in Table R11. Method means and trajectory SDs are recomputed from these predictions, with no adjustment of labels, probabilities or weights to match the displayed summaries. The same array supplies the paired contrasts, bootstrap intervals and multiplicity-adjusted tests. All reductions precede rounding.

Table D29: Paired continual-learning contrasts. Main diff. uses the displayed Table 3 means. Recomputed and paired differences, confidence intervals and tests use the same participant-level predictions. Holm correction covers the four listed comparisons.

Comparator	Main diff.	Recomputed	Paired diff.	95% CI	p	Holm p
BiLoRA-style FA	1.68	1.68	1.68	[1.12, 2.30]	0.0005	0.0020
GPM-FA	1.81	1.81	1.81	[1.25, 2.42]	0.0005	0.0020
One-sided LPA	0.95	0.95	0.95	[0.37, 1.48]	0.0020	0.0020
Unconstrained rank-8	2.28	2.28	2.28	[1.72, 2.91]	0.0005	0.0020

Table R11: Agreement between the displayed primary summaries and the higher-precision prediction-based calculations. All paired contrasts in Table D29 use the latter values before rounding.

Method	Main mean	Main SD	Recomputed mean	Recomputed SD
RetinaLPA-2S	82.20	1.31	82.201065	1.310286
BiLoRA-style FA	80.52	1.72	80.520320	1.717804
GPM-FA	80.39	1.69	80.389324	1.690596
One-sided LPA	81.25	1.62	81.251005	1.621402
Unconstrained rank-8	79.92	1.55	79.921351	1.550322

Participant-cluster inference. Within each domain, draw participants with replacement and reuse the selected identities for every method, seed and order. Compute the domain-balanced method difference for each of 1,000 draws and report the percentile interval. For each of 1,999 permutations, one Bernoulli swap per participant exchanges the complete two-method prediction vectors across all 30 runs. Use the two-sided absolute-statistic test with $(b+1)/(B+1)$, so the smallest reportable p is 0.0005. Holm correction covers precisely the four prespecified CL contrasts. These intervals

condition on the fixed trained run set. They do not claim uncertainty over new training seeds. SD across the 30 trajectories is reported separately, never as a participant confidence interval.

Estimator consistency and input validation. The paired table, higher-precision summaries and multiplicity-adjusted tests use one prediction array. The discrepancy between the two reduction paths is $2.04\mathrm{e}{-14}$ percentage points. Input checks reject mixed candidate and retained states, duplicate keys, inconsistent reference labels, missing runs and invalid probabilities. Missing pairs are not selectively removed to improve the result.

Table R12: Statistical consistency checks and error handling. Model-state identities, participant pairing and the metric definition remain fixed across the marginal and paired reductions.

Check	Required result / action
Original checkpoint and prediction provenance	Match run, order, seed, encoder, adapter, calibrator and state; retain input SHA-256.
Same-model marginal recomputation	Regenerate Table 3 candidate means and SDs from original exports.
Same-estimand paired identity	Mean pairwise difference equals difference of unrounded means within 10^{-10} pp.
Uncertainty and multiplicity	One participant cluster across runs; fixed four-comparison family; explicit resampling seed and counts.
A genuine discrepancy remains	Correct the affected empirical table from traceable exports, or rerun the affected checkpoints; never calibrate real outputs to the old headline.
Separate retained-state analysis	Use the retained checkpoints and post-rejection teacher history of each method.

U.2 Task-specific denominators

Table D30: Participant accounting by estimand. Counts are copied or added from Tables 1 and 8.

Population	People	Use and exclusion rule
Whole allocation	13,040	Development, support, release and queries across all primary tracks, not the staging test size.
D0 development	2,200	1,600 train, 400 tune and 200 calibration participants, excluded from query metrics.
D0–D3 staging query	920	200 + 240 + 240 + 240 with shared identities across methods.
D4 consecutive query	1,200	Separate prospective workflow population with its image-conditional denominator reported.
Local support reserves	320	Four domains with 80 reserved participants each. The primary update uses 40 per domain.
Release validation	2,400	Four independent 600-person streams, excluded from query metrics.
E1/E2 risk query	6,000	Two 3,000-person systems with separate event and censoring ledgers.

The 13,040-person primary allocation, 920 staging queries, 1,200 D4 encounters and 6,000 external-risk participants describe different study roles. Repeated runs and analytical controls do not add independent participants. The static staging difference of 0.20 points and interval $[-1.05, 1.47]$ concern the separate snapshot comparison rather than the continual-learning contrasts in Table D29.

V External-risk transport and interpretation

The pooled E1/E2 result favors the temporal-foundation comparator over the compact risk branch in AUC. Fixed-input risk invariance follows from the bypassed, versioned branch and is not an accuracy claim or evidence that another protected basis improves prognosis. Appendix Q makes this shared architectural property explicit and separates it from direct-readout attribution.

Table D31: System-specific external-risk results. Participant and outcome counts reproduce Table 16. Pooled and system-specific AUCs have different estimands.

System	Model	n	AMD	Death	Loss	AUC12	BS12	Slope
E1	Compact	3000	249	77	433	0.772	0.0768	0.91
E1	Temporal foundation	3000	249	77	433	0.777	0.0761	0.95
E2	Compact	3000	304	60	431	0.788	0.0832	0.96
E2	Temporal foundation	3000	304	60	431	0.791	0.0828	0.98

The comparison uses identical landmarks and event and censoring definitions within each system. Deaths remain distinct from loss to follow-up, and administrative survivors remain controls at the evaluation horizon. Pooled AUC is not the arithmetic average of the two system AUCs.

Paired-risk uncertainty. The paired estimator requires one-to-one participant and landmark joins, consistent handling of ties, event and control counts, and the stated censoring-weight folds. Correlated predictions can yield a narrower paired-difference interval than the two marginal intervals. Rounded AUCs alone cannot verify the paired variance. The complete paired estimator must be applied within each bootstrap draw, with participant clustering retained.

Scope of future risk updates. A separately versioned temporal-risk candidate requires matured development outcomes, selection and calibration on permitted development records, and an untouched external outcome cohort. A richer temporal encoder or additional supervision is a separately identified variant. Stage labels do not authorize learning from future outcomes. The system-specific results above do not constitute a result table for an otherwise unspecified additional prognostic variant.

Censoring robustness. System-specific loss models conditioned on stage, age, quality and history availability and upper-tail weight truncation are distinct sensitivity procedures. Their interpretation requires the associated weight distributions and effective sample sizes. They must not change the documented-confirmation endpoint or be described as learning prognosis through a stage-only update. The primary censoring definition remains the one in Appendix J.

W Processing energy and local-update cost

Tables 2 and 55 distinguish measured processing energy from conditional idle arithmetic. The 3.275 J CPU INT8 and 2.150 J encoder-plus-host comparison is a 34.35% processing-energy reduction. With a shared 3 W idle assumption, the corresponding conditional savings are 12.13%, 6.15% and 3.38% at two, five and ten idle seconds. These are not measured savings over complete clinical visits.

Table D32: Additional multi-session and device measurements. Local update excludes the independently timed 600-record release evaluation. Cold and warm starts have different boundaries.

Measurement boundary	Median s	P95 s	MiB	Mean J
Local update, excluding release evaluation	49.000	68.000	384	192
Independent release evaluation, 600 records	365.000	442.000	416	1462
Cold-start image-to-report	1.740	2.360	272	6.85
Warm image-to-report, repeat-device study	0.571	0.669	232	2.19

The repeated-device study uses three independently identified device sets over three measurement days. The two execution paths share timestamped inputs and workload, with path order alternated within sessions. Cold starts, warm runs, failed maps, transfer stress, retakes, idle periods and supply-side power traces have separate records. Device and session variability are distinct from repeated cycles on one device.

Full-visit energy accounting would include $E_{\text{visit}} = E_{\text{acquisition}} + E_{\text{processing}} + E_{\text{retakes}} + E_{\text{idle}}$. Components can be shared across backends only under identical acquisition and workflow conditions. The 300-second local-optimization gate excludes the separately timed release evaluation. No unreported full-visit or battery-lifetime result is inferred from the processing measurements.

Operator ownership, model and configuration digests, hardware identity, precision, toolchain, batch size, thermal state, warm-up and measurement boundaries identify each physical run. Signed host computation remains separate from frozen-encoder mapping. Physical mapping logs and numerical folding checks provide different evidence. Local processing and feature-only replay reduce raw-image transmission and retention, but do not by themselves establish a formal privacy guarantee against feature reconstruction.

X Analysis reproducibility and evidence scope

The supplementary analyses distinguish clinical prediction, controlled feature-space optimization, analytical verification and physical execution. Their implementation and reporting retain the relevant prediction sources, checkpoint states, denominators and measurement boundaries. The records below describe how the corresponding analyses connect to the reported tables.

Paired predictions. Candidate comparisons use a common participant-level export with reference labels, domain membership, run and seed identifiers, checkpoint state and prediction probabilities. Each method mean and trajectory SD is recomputed before paired inference. Equation 12 checks the reduction identity. Labels, probabilities and weights are not adjusted to reproduce a selected summary value.

Retained checkpoints. A release-aware analysis links the parent, teacher, optimized candidate and retained state at every update. If rejection changes a subsequent parent, that update is optimized from the retained state with reset optimizer moments. Gate inputs include release predictions, model and configuration identifiers, numerical drift, fixed-input outputs and resource status. Reference policy checks and physical deployment authorization remain distinct.

D4 pairing and timing. D4 paired intervals use zero-shot and adapted predictions on common encounter identities. Marginal confusion matrices alone do not identify the joint errors. The primary pairing and analytical re-pairing sensitivity are reported separately. Chronology is expressed relative to protocol freeze and linked to acquisition, adjudication, candidate-freeze and release records. Workflow accounting retains quality failures and unresolved reference labels rather than imputing model-derived targets.

Readout-level attribution. Direct-readout controls use fixed W and Q, their singular spectra and feature records in one coordinate system. Matched-step and longer-optimization conditions are separate. The induced-metric objective preserves the displayed feature penalties at fixed logits but does not imply identical finite-step trajectories or weight-decay effects. Numerical equivalence is distinct from clinical equivalence.

Artifact availability. The manuscript archive contains editable LaTeX, the bibliography, official style files and figure assets. The core experimental implementation is supplied separately as `source_code.zip`. Data-dependent reproduction also requires the corresponding authorized records and model artifacts. Compiling the manuscript reproduces its document rather than rerunning the experiments. Conditional benchmark release follows the Ethics Statement and does not imply that release approval has already been obtained.