Supplementary Materials

Physical-Time Signed Residual Calibration for Fixed-View Forest Smoke Detection

S1. Implementation and Objective Details

S1.1. Feature Construction and Stage Settings

Table S1 summarizes the settings of the reported pipeline. The detector, TDNRI, CCEH, and PMTR are frozen successively before the next stage is trained. The final HSRC head contains 1761 trainable parameters; this count excludes the upstream system.

Table S1. Reported final-pipeline settings.

Stage	Setting	Value
S0	Image size / batch / workers	1024 × 1024 / 16 / 8
S0	Training schedule	Up to 100 epochs per stage; seed 42
S0	Candidate extraction	Confidence 0.01; NMS IoU 0.70; maximum 20
Features	Pooled detector tensors	P3: 128 × 16 × 16; P4: 256 × 8 × 8
TDNRI	Time normalization / constants	Seconds ÷ 30; τ_s = 2.5; τ_l = 20; κ = 0.95
TDNRI	Initialization	Mean from first observation; variance 1.0; $\gamma = 10^{-3}$
CCEH	Projection / branch dimensions	Each input → 64; each branch 192 → 64 → 1
PMTR	Optional horizons	5, 15, 30, 60, 120 s
PMTR	Encoder / fusion	1538 → 96 → 32; 33 → 32
PMTR	Optimizer / stopping	AdamW; lr 10^{-3}; wd 10^{-3}; ≤500 epochs; patience 80
PMTR	Optional-slot dropout / bound	0.20 / β_p = 2
PMTR	Trainable parameters	152,739
HSRC	Head / trainable parameters	35 → 32 → 16 → 1 / 1761
HSRC	Optimizer / stopping	AdamW; lr 10^{-3}; wd 10^{-3}; ≤500 epochs; gradient clip 5
HSRC	Pooling / hard negatives	Temperature 0.25; 30% groups; 15% candidates
HSRC	Ranking margin / bound	0.50 / β_h = 2

lr: learning rate; wd: weight decay. Stage-freezing order and loss coefficients are specified in the main text.

The local CCEH descriptor uses aligned 3 × 3 ROIAlign with sampling ratio 2 on each refined feature map. For the neighborhood descriptor, the candidate is expanded by λr = 1.75 and clipped at the image boundary, then pooled by aligned 7 × 7 ROIAlign with the same sampling ratio. The normalized bin centers and central exclusion margin are

$$u_j = \frac{j+0.5}{7}, \qquad j = 0, \ldots, 6, \qquad m_r = \frac{1-\lambda_r^{-1}}{2}. \qquad m_r < u_j < 1 - m_r. \tag{S1}$$

The strict central interval selects the central 3 × 3 cells. The other 40 cells are averaged separately on P3 and P4 and concatenated. Annular neighborhood denotes this central-cell mask on the pooled grid, not an exact geometric mask outside the original candidate box. The same grid rule is retained after boundary clipping.

The internal TDNRI residual contains five blocks, but the candidate descriptor includes only the short residual, long residual, and log-variance ratio. Across P3 and P4, these supply 3 × (128 + 256) = 1152 dimensions. The detector-derived protection mask controls frame-level updates and need not correspond one-to-one with retained CCEH supervision entries.

PMTR standardization uses valid training anchors only. Invalid slots are zeroed and excluded by attention masking. All samples entering the evaluated PMTR/HSRC path have a native preceding observation; no separate all-history-missing branch is evaluated. Optional physical-time slots are dropped independently with probability 0.20 during training. Attention entropy is unnormalized over the valid slots.

S1.2. HSRC Loss Definitions

Let P and N be the positive and negative temporal-group sets. Candidate logits are aggregated using the main-text softmax-weighted rule with temperature 0.25. BCE takes logits and averages per-element losses within each indicated set. Group-balanced classification is

$$\mathcal{L}_{\text{group}} = \frac{1}{2}\text{BCE}(Z_{\mathcal{P}}, 1) + \frac{1}{2}\text{BCE}(Z_{\mathcal{N}}, 0). \tag{S2}$$

The highest-scoring 30% of negative groups, rounded upward, form the hard-negative set. Ranking averages all pairs of a positive group and a selected negative group, with margin m = 0.50:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}||\mathcal{H}_g|}\sum_{p\in\mathcal{P},n\in\mathcal{H}_g} \text{softplus}\left(Z_n - Z_p + m\right). \tag{S3}$$

Candidate supervision retains all supervised positives and the highest-scoring 15% of supervised negatives, rounded upward. With $\ell(z,y)$ = softplus(z) − yz, the balanced candidate objective is

$$\mathcal{L}_{\text{cand}} = \frac{1}{2|\mathcal{P}_c|}\sum_{i\in\mathcal{P}_c} \ell\left(z_i, 1\right) + \frac{1}{2|\mathcal{H}_c|}\sum_{i\in\mathcal{H}_c} \ell\left(z_i, 0\right). \tag{S4}$$

The regularizer is

$$\mathcal{L}_{\text{reg}} = \mathbb{E}[(\Delta z^H)^2]. \tag{S5}$$

The complete objective adds the group term, 0.50 times ranking, 0.25 times candidate supervision, and 0.02 times regularization. These definitions apply to nonempty participating sets. Equal positive and negative set weights are not equivalent to an unweighted mean over all examples.

The weighted group logit lies between the smallest and largest candidate logits. It avoids the explicit count offset of unnormalized log-sum-exp, although changing group composition can still change its value. Inference uses the maximum candidate probability, with zero assigned to a window containing no candidates.

S2. Resource Audit and Reviewed Observations

Table S2. Audited PyroNear2025-I annotation states before source filtering.

Downloaded directory	Images	Positive	Empty label	Missing label	Annotated
Train	3675	3161	137	377	3298
Validation	488	418	23	47	441
Test	536	462	27	47	489
Total	4699	4041	187	471	4228

CCEH candidate supervision distinguishes valid empty-label frames from frames with missing or invalid annotations. Valid empty-label frames contribute negative candidates, whereas missing labels and invalid positive geometry are excluded from candidate supervision.

Table S2 records the downloaded snapshot before source filtering. Images with annotation files number 3298/441/489 in train/validation/test; the experiments retain these directory assignments. The 4699 stored image files include 471 without annotation files. Five static duplicate groups were found, with none spanning the audited static partitions. Original-resource overlap and the separate ordinary temporal-test resource are distinguished from the inputs used in the reported experiments (Sections S7.6 and S7.16).

The onset review identifies 75 high-confidence events: 46 for training and 29 for validation. Their observed endpoints generate 288 causal windows. The no-smoke review identifies 329 temporal units: 292 for training and 37 for validation. In the primary onset set, no observed frame lies strictly between the last reviewed no-smoke endpoint L and the first reviewed visible-smoke endpoint R.

The FIgLib archive contains 522 verified compressed packages. Source-out evaluation uses 127 reviewed positive events. Of 30 candidate negative streams, 22 are retained, two contain visible smoke, and six are invalid for the protocol. The retained negative set contains 1646 windows and 27.8506 camera-hours of database-covered exposure.

Figure 1 in the main manuscript illustrates a real forest sequence, with its temporal reference and annotation limits explained in the caption.

S3. Additional Development Comparisons

S3.1. Intermediate Reference-State Variants

The seed-42 development screen compared R0–R7. Table S3 records the intermediate R2–R5 variants; the principal main-text comparison uses three-seed confirmation of R0, R1, R6, and R7. R6 reached AUC 0.75735 in the development screen. The R7 protection-factor screen selected $\kappa = 0.95$, with AP/AUC of 0.91451/0.76459.

Table S3. Intermediate seed-42 TDNRI development records.

Variant	AP	AUC	F1
R2	0.90734	0.74863	0.89271
R3	0.90453	0.74210	0.89369
R4	0.90413	0.74117	0.89301
R5	0.89466	0.74143	0.89828

These development entries are separate from the three-seed confirmation summarized in main-text Figure 8b and fully reported in Table S8.

S3.2. Available CCEH Screening Results

The CCEH screen favored different configurations on different metrics. Reported C4 results are candidate AP 0.85735, Recall@FPR = 1% of 0.09629, and Recall@FPR = 5% of 0.24068. Reported C5 results are AUC 0.77959 and F1 0.84251. C5-minus-C0 absolute differences are +0.04356 AP, +0.08146 AUC, and +0.04168 F1. Complete archived common-population metrics are reported in Table S20 (Section S7.7).

Later multi-seed comparisons did not establish consistent C5 superiority over C4 in AP or extreme low-FPR recall. C5 was retained for its explicit local–neighborhood and neighborhood–global relations and used as the frozen presence anchor for temporal correction. Both branch outputs are signed, and the final configuration has no nonzero auxiliary competition loss.

C2 was evaluated separately as a four-observation track-level GRU. Label-free geometric association formed tracks within each physical sequence, using a maximum interval of 90 s, minimum IoU of 0.20, maximum normalized center distance of 0.50, and maximum area ratio of 4.0. A 384-dimensional local feature and six normalized box-geometry cues were projected to 128 and 32 dimensions, respectively, and passed to a single-layer GRU with hidden dimension 128. The model has 162017 parameters. Training used the non-FIgLib

training split, seed 42, and validation AP for checkpoint selection. Neighborhood, global, TDNRI residual, CCEH evidence, and detector confidence were excluded from its inputs. Candidate-prefix AP/AUC/F1 were 0.88935/0.74972/0.88240; keeping one final prediction per track gave 0.81779/0.73992/0.82869. These different evaluation units prevent direct ranking against the candidate-level CCEH screen.

S3.3. Archival Multiplicative Verification Results

Table S4 records the previously evaluated multiplicative verifier. Its source-out thresholds were frozen through the development protocol, and the event metrics use the same 127-event population. This archival comparator is separate from the suppress-only HSRC ablation, which provides the direct mechanism comparison for signed correction.

Table S4. Archival multiplicative-verifier source-out results.

Method	At R (%)	30 s (%)	60 s (%)	120 s (%)	300 s (%)	FAR (h^{-1})
Historical multiplicative verifier	4.72	4.72	9.45	12.60	18.11	0.107718

S4. Detector Accuracy and Computational Profile

YOLO11s-S0, YOLO26s [36], and RT-DETR-L [19] are compared on a common non-FIgLib PyroNear2025-I test subset of 302 images with 261 annotated smoke instances. Table S5 reports formal detector-validator metrics. These comparisons characterize candidate generators and do not reselect the CITE-R operating threshold.

Table S5. Formal detector-validator metrics on the common 302-image non-FIgLib test subset.

Detector	Precision	Recall	mAP50	mAP50–95
YOLO11s-S0	0.5658	0.5134	0.5221	0.2839
YOLO26s	0.6130	0.3830	0.3987	0.2140
RT-DETR-L	0.5523	0.4904	0.4296	0.2058

S0 has the highest recall and both mAP measures in this comparison; YOLO26s has the highest precision. Table S6 profiles detector inference on the same RTX 4090 D using a 1024-pixel input, FP32, and batch size 1. One hundred images are preloaded, and ten warm-up iterations precede timing. FPS is calculated from mean latency.

Table S6. Detector-only computational profile at 1024-pixel input, FP32, and batch size 1 on an RTX 4090 D.

Detector	Params (M)	GFLOPs	Latency (ms)	FPS	Peak VRAM (MiB)
YOLO11s-S0	9.428	55.152	10.984	91.04	230.4
YOLO26s	9.949	59.195	10.117	98.84	245.1
RT-DETR-L	32.808	274.275	31.098	32.16	434.6

Images are preloaded; 10 warm-up iterations and 100 timing images are used. These values do not measure the complete verification pipeline.

In this detector-only profile, YOLO26s is less than 1 ms faster than S0 per image. RT-DETR-L uses approximately 3.5 times the parameters, five times the GFLOPs, and 2.8 times the latency of S0. The costs of TDNRI, CCEH, PMTR, and HSRC are not included.

An independent size-stratified analysis maps boxes to a canonical 1024 × 1024 coordinate system and uses a common JSON-based COCO-style evaluator. Relative box-area groups are Tiny below 0.05%, Small from 0.05% to below 0.25%, Medium from 0.25% to below 1%, and Large at least 1%. The subset contains 34 Tiny, 123 Small, 55 Medium, and 49 Large instances. The independent overall AP50 is a sanity check for this evaluation path and is not substituted for the formal detector-validator mAP50.

The available size-specific results show RT-DETR-L leading in Tiny and Medium AP50 at 0.5962 and 0.4501. Its overall, Tiny, and Small AR@100 are 0.8391, 0.8824, and 0.9024. S0 leads in overall, Small, and Large AP50. For the 123-instance Small group, AP50 is 0.6572 for S0, 0.5745 for RT-DETR-L, and 0.5426 for YOLO26s. These entries illustrate size-dependent detector behavior and are not a full replacement for the formal comparison.

S5. Additional Operating-Point Diagnostics

The representative boundary thresholds are 0.8958815336 for CITE-R and 0.960521 for the historical CNN–LSTM; each triggers one of 37 reviewed negative validation units. Exact CITE-R R-versus-negative AP/AUC are 0.757907/0.832246. Four events alarm before L and nine at L, with a union of 12. These premature responses are not credited as successful boundary detection.

Table S11a retains the original event- and negative-stream-resampling diagnostics, with 10,000 replicates. These use events and streams as the resampling units. Table S23 reports the added paired camera-cluster intervals for the frozen operating points, accounting for records with the same available camera identifier.

The descriptive common-FAR envelope gives CITE-R a local recall advantage at budgets from 0.20 to 0.40 h^{-1}, a slight disadvantage at 0.10 and 0.50 h^{-1}, and equal recall at 0.60 h^{-1}. Its largest advantage is 0.0157 in recall proportion, or two events. Section S7.21 reports the complete budget grid and paired camera-cluster intervals for this retrospective envelope. The frozen operating-point comparison remains separate: its camera-cluster recall-difference interval reaches zero, while the FAR-difference interval is below zero (Table S23).

The ten-head analysis includes the original seed-44 head and nine additional heads trained under the same protocol with upstream representations fixed. Mean ± standard deviation is 0.754233 ± 0.007821 for validation group AUC, 0.731707 ± 0.006590 for group AP, 0.425000 ± 0.029667 for constrained validation recall, 0.360630 ± 0.056868 for source-out Recall@300 s, and 0.222617 ± 0.109851 h^{-1} for source-out FAR. Relative FAR reduction versus CCEH is 61.25% ± 19.12%.

Real-scene qualitative examples are presented in main-text Figure 10 and discussed in Section 3.5. Each method's box and crop identify its actual maximum-score candidate; the maximizing candidates can differ between methods. Candidate replay and spatial diagnostics are reported in Section S7.9.

S6. Complete Supporting Results

Tables S7–S13 provide the complete supporting values for the main-text result figures, preserving the reported precision and complementary metrics. Tables S7–S12 cover mechanism comparisons, reference and context evidence, boundary evaluation, uncertainty, and nuisance responses. Table S13 reports the physical-time controls.

Table S7. HSRC mechanism comparisons. All variants trigger one false validation group. AP, AUC, and transition-group recall use the validation population; the last two columns use FIgLib.

Variant	Validation			FIgLib	
	Group AUC	Group AP	Transition hits (n/28)	300-s recall (%)	FAR (h^{-1})
Full CITE-R	0.7632	0.7399	13/28	40.16	0.2872
No positive boost	0.7407	0.7082	2/28	35.43	0.1795
No hard-negative ranking	0.7685	0.7417	12/28	25.98	0.0718

On the source-out population, signed rather than suppress-only correction raises Recall@300 s from 35.43% to 40.16% and FAR from 0.1795 to 0.2872 h^{-1}. The positive route preserves more sensitivity at the transferred operating points, with an associated false-alarm increase. The equal validation false-group allowance and differing external FAR refer to separate populations and constraints.

Removing hard-negative ranking slightly improves validation AUC and AP but lowers constrained hits to 12/28. Its source-out point is more conservative, with Recall@300 s of 25.98% and FAR of 0.0718 h^{-1}. Retaining ranking is associated with higher recall and higher FAR at validation-selected thresholds. This result concerns operating behavior and does not imply uniform improvement in ranking metrics or FAR.

The one-false-group allowance is a transition-validation constraint, not a camera-hour FAR limit. Transition comparisons use 28 positive groups; the separate boundary comparison in Table S10 uses 29 events.

Table S8. Three-initialization confirmation of reference-state variants. Values are mean ± standard deviation; the evaluation population is that of the TDNRI comparison.

Variant	Reference mechanism	AP	AUC	F1
R0	No temporal state	0.91330 ± 0.00183	0.75747 ± 0.00528	0.89263 ± 0.00067
R1	Adjacent-frame difference	0.91189 ± 0.00115	0.75391 ± 0.00317	0.89323 ± 0.00181
R6	Dual states and elapsed time	0.90995 ± 0.00380	0.75485 ± 0.00676	0.89219 ± 0.00339
R7	R6 plus candidate protection	0.91622 ± 0.00463	0.76968 ± 0.00529	0.89444 ± 0.00214

Table S9. Inference-time interventions on fitted CCEH representations. AP changes are raw proportion differences from the corresponding intact configuration. The head is not retrained after these interventions.

Intervention	Candidate AP change	Window AP change
Shuffle annular neighborhood	−0.02792	−0.01199
Shuffle TDNRI residual	−0.00857	−0.00594
Remove smoke-oriented branch	−0.02528	−0.01854
Shuffle global feature	−0.00130	−0.00078
Remove nuisance-oriented branch	−0.00437	−0.00099

Table S10. Reviewed validation-boundary evaluation. R denotes the first confirmed visible-smoke observation, and L denotes the last confirmed no-smoke observation.

Method	Triggered negative units[1]	R hits	R recall (%)	AP	AUC	Any alarm at/before L (n/29)
Historical CNN–LSTM	1/37	0/29	0.00	0.4373	0.5377	0/29
CITE-R	1/37	14/29	48.28	0.7579	0.8322	12/29

[1] Triggered units in the reviewed no-smoke set, not false-negative errors. AP/AUC compare R anchors with negative units. CITE-R before-L and at-L counts are 4 and 9; their union is 12. The recurrent checkpoint is reused without retraining.

Table S11a. Original event/negative-stream bootstrap diagnostics for representative CITE-R minus CCEH differences. Camera-cluster intervals are reported separately in Table S23.

Quantity	Observed difference (CITE-R − CCEH)	95% bootstrap interval	Resampling unit
Recall@300 s	−0.04724	[−0.0945, −0.00787]	Positive event
FAR (h^{-1})	−0.28725	[−0.6035, −0.0355]	Negative stream

Observed differences are calculated from the original event and episode counts; intervals use 10,000 bootstrap replicates. Recall differences are proportions, not percentages.

Table S11b. Ten HSRC initializations with upstream CCEH and PMTR fixed. Each initialization has a non-FIgLib validation-selected threshold.

Metric	Mean ± standard deviation
Validation group AUC	0.754233 ± 0.007821
Validation group AP	0.731707 ± 0.006590
Validation constrained transition recall	0.425000 ± 0.029667
Source-out Recall@300 s	0.360630 ± 0.056868
Source-out FAR (h^{-1})	0.222617 ± 0.109851
Relative FAR reduction versus CCEH	61.25% ± 19.12%

Recall values are proportions; relative FAR reduction is reported as a percentage. FAR is measured in h^{-1}.

Table S12. False alarms on seven reviewed hard-negative streams selected by an alarm from CCEH or CITE-R. Episode counts use the same 120-s merging rule as Table 2.

Nuisance category	Streams	Alert windows CCEH / CITE-R	Episodes CCEH / CITE-R	Episode reduction
Cloud / fog / haze	4	21 / 15	9 / 8	11.11%
Terrain / vegetation	3	9 / 0	7 / 0	100.00%
Total	7	30 / 15	16 / 8	50.00%

Table S13. Physical-time controls at seed 44. Group AUC/AP, positive hits, and false-group counts are validation results on 28 positive and 27 negative transition groups. Each reported point triggers one negative group. Mean valid-anchor counts include the native slot and are averaged over all 1244 training and validation cache records. A3 denotes PMTR; A4 denotes the corresponding HSRC stage.

Variant	Stage	Group AUC	Group AP	Positive hits (n/28)	False groups	Mean valid anchors
Full Physical-Time	PMTR (A3)	0.720899	0.677081	2/28	1	2.624
Full Physical-Time	After HSRC (A4)	0.7632	0.7399	13/28	1	2.624
Same-Anchor No-Time	PMTR (A3)	0.694444	0.657803	2/28	1	2.624
Same-Anchor No-Time	After HSRC (A4)	0.7513	0.7289	2/28	1	2.624
Observation-Indexed	PMTR (A3)	0.695767	0.652019	2/28	1	5.359
Observation-Indexed	After HSRC (A4)	0.7500	0.7261	2/28	1	5.359

Same-Anchor No-Time retains the full historical-anchor mask and coverage while zero-filling explicit elapsed-time and rate-dependent descriptor channels. Observation-Indexed uses observation-order retrieval and the same descriptor ablation. Mean anchor counts are repeated for the two stages because they describe the inputs. These are single-seed validation controls; no FIgLib result belongs to the original Table S13 experiment. Archived thresholds, stopping epochs, candidate discrimination, correction-sign counts, and history-coverage diagnostics are detailed in Section S7.

S7. Additional Protocols and Reproducibility Analyses

S7.1. Development Lists and Supervised Caches

The source-filtered static image lists contain 1834 training, 267 validation, and 302 test paths. The four-observation temporal protocol contains 29,822 non-FIgLib training windows and 7406 validation windows. Image-list, window, candidate-entry, transition-group, and event counts refer to distinct stages and units, which are reported separately below.

The CCEH development cache contains 69,560 training and 16,226 validation entries. Excluding 10,814 training and 2488 validation entries with ignored supervision leaves 58,746 and 13,738 positive/negative supervised candidates, respectively, matching the retained CCEH training records. Table S14 reports these populations. The PMTR/HSRC cache contains 1244 transition records: 1061 training records from 188 groups and 183 validation records from 55 groups. Training comprises 41 positive and 147 negative groups; validation comprises 28 positive and 27 negative groups. Positive records number 41 and 28; the remaining 1020 and 155 records are negative.

Table S14. Archived supervised populations by split. An entry is a cached record; the group columns refer to the PMTR/HSRC grouping, not independent camera or fire counts.

Population	Training	Validation
CCEH positive entries	43,277	9170
CCEH negative entries	15,469	4568
CCEH ignored entries	10,814	2488
CCEH total entries	69,560	16,226
CCEH supervised entries	58,746	13,738
PMTR/HSRC positive records	41	28
PMTR/HSRC negative records	1020	155
PMTR/HSRC total groups	188 (41 positive; 147 negative)	55 (28 positive; 27 negative)

The reviewed onset population contains 46 training and 29 validation events; endpoint reconstruction gives 288 causal windows. The anchor-decision table records 69 confirmed anchors (41 training and 28 validation), three rejected anchors, and three candidate-generation failures. The 69 confirmed anchors correspond to the 69 positive transition groups. Of the 329 reviewed negative units (292 training and 37 validation), 147 training and 27 validation units are represented by negative groups in the transition cache; all retained group identifiers map back to reviewed units. The remaining 145 training and 10 validation units have no record in that cache. The accompanying reviewed_negative_unit_crosswalk.csv lists every reviewed unit and its inclusion state. Boundary evaluation and transition-group evaluation therefore use distinct populations.

S7.2. Definition and Selection of the Time Controls

The full PMTR descriptor concatenates the 384-dimensional appearance difference, its absolute value, the signed and absolute time-scaled differences, cosine similarity, and the logarithmic elapsed-time term. The two no-time controls set both 384-dimensional time-scaled blocks and the one-dimensional elapsed-time block to zero. Thus, 769 of the 1538 channels are zero-filled while the input dimensionality and architecture are retained. This ablates explicit elapsed-time and rate encoding at the PMTR descriptor interface; the frozen upstream pipeline is unchanged.

Same-Anchor No-Time uses the full representation's retained historical anchors and intervals. Observation-Indexed retrieves up to five distinct earlier observations from the same frame-parent directory, excluding the

native predecessor. Timestamps enforce strict causality and support diagnostics, but numerical elapsed times are zero-filled in its descriptor. Indexed retrieval has no matching physical-age limit, so this control changes temporal coverage and encoding together. It is distinct from the matched-anchor encoding comparison.

The full seed-44 run was retained, and each control was trained with its corresponding PMTR and HSRC stage on the non-FIgLib training/validation protocol. The retained summaries provide the validation-constrained thresholds and stopping epochs in Table S15. These are transition-validation thresholds, not thresholds from the separate 37-negative-unit boundary procedure.

Table S15. Archived validation selection records for the physical-time controls. Candidate AP and AUC use candidate records; the one-false-group constraint and hits use transition groups.

Variant and stage	Threshold	Epoch	Candidate AP / AUC	Validation hits
Full Physical-Time / PMTR	0.9879937172	54	0.534904 / 0.874194	2/28; 1 false group
Full Physical-Time / HSRC	0.8958803415	288	0.607851 / 0.879724	13/28; 1 false group
Same-Anchor No-Time / PMTR	0.9920802712	138	0.496067 / 0.847465	2/28; 1 false group
Same-Anchor No-Time / HSRC	0.9867050052	415	0.567349 / 0.870968	2/28; 1 false group
Observation-Indexed / PMTR	0.9920808673	60	0.488602 / 0.857373	2/28; 1 false group
Observation-Indexed / HSRC	0.9846367836	119	0.553522 / 0.867512	2/28; 1 false group

The retained HSRC checkpoint summaries record 1027 negative and 217 positive residuals for Full Physical-Time, 840/404 for Same-Anchor No-Time, and 1020/224 for Observation-Indexed. These sign counts cover all 1244 cached records, not validation alone. They describe the direction of the learned corrections rather than the number of correctly rescued smoke events.

S7.3. Observed History Coverage

Table S16. Coverage diagnostics computed from the retained caches. The native slot is included in mean slot counts and excluded from optional-history counts and age summaries.

Representation	Mean slots all / validation	Valid optional slots in validation	Optional age 95th percentile	Optional slots over 600 s
Full Physical-Time	2.624 / 2.541	282	129.0 s	0
Same-Anchor No-Time	2.624 / 2.541	282	129.0 s	0
Observation-Indexed	5.359 / 5.820	882	2027.6 s	116

The full representation's oldest optional validation observation is 162 s old. The indexed control has more valid slots but includes substantially older observations. Slot count alone therefore does not quantify comparable information about the same recent interval. The diagnostics explain the coverage difference without isolating history age as the sole cause of the performance difference, because indexed retrieval and descriptor encoding change together.

S7.4. Manual Review and Excluded Negative Streams

The first author performed the manual review. Per-unit onset and negative-stream decisions are included in the public reproducibility archive. The review used one reader; inter-rater agreement was not assessed. The exact timing of access to model outputs or official temporal labels at the initial decisions cannot be independently reconstructed from the retained records, so no formal blinding claim is made.

Of the 30 candidate negative streams, 22 were retained as verified no-smoke, two were marked smoke-visible, and six were excluded from the evaluated daytime RGB scope. Table S17 reports the recorded reasons for these six exclusions.

Table S17. Archived reasons for the six negative-stream exclusions. Split labels here identify original resource partitions, not the role of these streams in model training.

Original split	Sequence	Recorded exclusion reason
train	20200727_Border11Fire_lp-s-mobo-m	Monochrome mobo-m and dusk/night; outside RGB daytime scope.
train	20200727_Border11Fire_om-e-mobo-m	Monochrome mobo-m and dusk/night; outside RGB daytime scope.
train	20201003_structurefire_bh-w-mobo-c	Night sequence; outside daytime scope.
train	20201202_WillowFire-nightime-near-CDF-HQ_lp-w-mobo-c	Night sequence; outside daytime scope.
train	20210302_FIRE_lp-e-mobo-c	Night sequence; outside daytime scope.
val	20191030_CopperCanyon_om-s-mobo-c	Night sequence; outside daytime scope.

S7.5. Public Code Weights and Protocol Archive

Version 1.0.1 of the code and reproducibility materials was published on Zenodo on 12 November 2026 under the MIT License [37]. The archive contains nine bundles covering the environment, evaluation protocols, physical-time controls, supplementary baselines, and S0, TDNRI, CCEH, PMTR, and HSRC. Module README files map historical A3 and A4-v2 names to PMTR and HSRC; historical names beginning with icel are retained for traceability and do not add the superseded ICEL objective to the final method.

Reproduction proceeds by obtaining the third-party datasets, applying the archived source/split protocols, preparing S0 candidate and feature caches, and then following the stage-freezing order S0 → TDNRI-R7 → CCEH-C5 → PMTR → HSRC. The environment bundle records Python, PyTorch, CUDA and Ultralytics versions. Detector training arguments and configuration files document the frozen S0 experiment. Module bundles include the trained weights and reference results; the evaluation bundle supplies source-out, false-alarm, bootstrap, robustness, and mechanism-analysis scripts. Dataset roots and historical absolute project paths must be adjusted to the local installation.

Raw third-party datasets and large generated feature databases are not redistributed. Compact transition caches and small reference outputs are retained. The S0 bundle includes a public candidate-extraction implementation; module README files describe the interfaces and historical naming. Data and archive access are described in the main-text Data Availability Statement.

S7.6. Data Roles and Label States

Table S18 distinguishes image counts from annotation-file states in the retained detector inputs. The final S0 experiment uses batch size 16, seed 42, and 1024-pixel input in both training stages: Pyro-SDIS first, followed by non-FIgLib PyroNear2025-I. These are the settings of the frozen S0 experiment.

Table S18. Detector input images and annotation-file states. Nonempty counts refer to images with valid nonempty labels, not the number of smoke instances.

Dataset and split	Images	Nonempty labels	Empty labels	Missing labels
Pyro-SDIS train	20,568	17,911	2657	0
Pyro-SDIS validation	774	521	253	0
Non-FIgLib image train	1834	1563	137	134
Non-FIgLib image validation	267	224	23	20
Non-FIgLib image test	302	257	27	18

In the retained Ultralytics loader, a valid image with an empty or missing label file is represented by a zero-object target. A missing file alone therefore does not remove an image from the S0 dataset. This detector-loader behavior differs from the CCEH candidate-supervision rule, which excludes missing or invalid annotations. Missing-label images are not described as manually verified no-smoke observations. The counts in Table S18 describe the retained input files; candidate-supervision counts are reported separately in Table S14.

Table S19 maps the original temporal split field to the experimental role recorded in the candidate caches. Every FIgLib row is assigned the external-test role; the original train/validation/test fields remain provenance metadata. FIgLib candidates have no CCEH training-supervision target in this cache. The complete external cache is larger than the reviewed evaluation population of 127 positive events and 22 negative streams.

Table S19. Source, original split, and experimental role in the retained candidate-feature caches. Counts are candidate records, not independent events.

Source	Original split	Experimental role	Candidate records
Non-FIgLib	train	Development training	69,560
Non-FIgLib	validation	Development validation	16,226
FIgLib	train	External test	74,528
FIgLib	validation	External test	6917
FIgLib	test	External test	9971

In the retained four-observation protocol, non-FIgLib training and validation inputs have no shared frame paths, frame filenames, or camera identifiers. No such identifier matches occur between the FIgLib and non-FIgLib subsets. These checks concern the retained protocol identifiers rather than an exhaustive comparison of all re-encoded images or physical fire events.

The archived ordinary non-FIgLib temporal-test protocol retains 281 frame filenames and the ptz-mostazal-01 camera identifier shared with its training counterpart. That protocol and its test cache are separate from both the 302-image static detector test and the FIgLib source-out evaluation. The main source-out results use the latter, with development-selected checkpoints and thresholds. Release-level overlap, cached test resources, and the population used for each reported result are therefore distinguished.

S7.7. Complete Candidate Screening Results

Table S20 reports the retained seed-42 candidate-screening results on the common CCEH validation population. C0 uses the frozen detector score; the learned heads use 58,746 training and 13,738 validation candidates. The GRU alternative in Section S3.2 uses a different track/prefix population and is described separately. The screening metrics below do not constitute a comparison of complete systems on the FIgLib event protocol.

Table S20. Candidate-level screening at seed 42. R at FPR 0.01 is a candidate false-positive-rate diagnostic, distinct from camera-hour episode FAR. Values are proportions.

Head	AP	ROC AUC	F1	R at FPR 0.01
C0_detector_score	0.81007	0.69813	0.80083	0.03097
C1_local	0.85098	0.77178	0.83276	0.06325
C3_concat	0.85536	0.77409	0.83188	0.08037
C4_difference	0.85735	0.77457	0.82976	0.09629
C5_cceh	0.85363	0.77959	0.84251	0.05562

C5 has the highest ROC AUC and F1 in this screen, whereas C4 has the highest AP and recall at FPR 0.01. The seed-42/43/44 summaries are retained alongside these screening results. Section S7.17 and Tables S40–S42 add the corresponding final-alarm comparisons.

S7.8. Score-Only Calibration Control

A residual MLP receiving only the frozen CCEH logit was added as a calibration control. Its hidden widths are 32 and 16, with LayerNorm, ReLU, dropout 0.1 and a correction bounded by 2 tanh. The input dimension and trainable parameter count are 1 and 673; the archived full HSRC head uses 35 inputs and 1761 parameters. Hidden widths, losses, hard-negative mining, optimizer and validation selection order are retained, while parameter counts and initialization histories are not matched.

The new head uses seed 44 and 500 AdamW epochs, learning rate and weight decay 0.001, gradient clipping at 5, group temperature 0.25, hard-negative group/candidate fractions 0.30/0.15 and ranking margin 0.5. Loss weights are 1.0 for group loss, 0.5 for ranking, 0.25 for candidate loss and 0.02 for correction regularization. All 1061 training and 183 validation records retain the audited split and group assignments. Epoch selection maximizes constrained validation group recall, followed by group AP and AUC; ties retain the earlier epoch. The selected new head is epoch 363. Training curves and per-record scores are retained with the revision artifacts.

Table S21. Development validation diagnostics for the score-only control. The common population contains 28 positive and 27 negative transition groups. Each method permits one false-positive negative group at its own threshold; these group proportions are distinct from episode FAR.

Head	Group AUC	Group AP	Positive hits	False groups	Head parameters
CCEH score	0.6098	0.5441	0/28	1/27	0 added
Score-only MLP	0.5992	0.5421	1/28	1/27	673
Archived full HSRC	0.7632	0.7399	13/28	1/27	1761

After epoch selection, the new head's external operating threshold was calibrated on maxima over the original 37 reviewed non-FIgLib validation negative units, permitting at most one unit to trigger. The 37 calibration units differ from the 27 negative transition groups in Table S21. The frozen threshold is 0.8836989998817445, with one false unit. The new head's epoch and threshold were fixed before its FIgLib evaluation. FIgLib scores were computed candidate by candidate before taking window maxima; applying a non-monotone correction to an already maximized window score would implement a different rule.

Table S22. FIgLib results using development-frozen thresholds. The 127 primary events and 22 verified negative streams follow the original R-anchor protocol. FAR uses 120-s episode merging over 27.8506 processed-frame-covered camera-hours. The external FARs are realized values, not matched constraints.

Method	Frozen threshold	Hits at 300 s	Recall	False episodes	FAR (h^{-1})
CCEH	0.977893	57/127	0.4488	16	0.5745
Score-only MLP	0.883699	61/127	0.4803	44	1.5799
CITE-R	0.895882	51/127	0.4016	8	0.2872

Relative to the score-only head, CITE-R reduces false-alarm episodes by 81.82% while losing 7.87 percentage points of Recall@300 s. The two methods jointly detect 50 events; one is detected only by CITE-R and eleven only by the score-only head. The stronger development-group result in Table S21 therefore does not imply uniformly higher external sensitivity. One negative stream has seven CITE-R episodes and six for each comparator, so the total reduction does not imply improvement on every stream.

For CCEH and CITE-R, 50 of the 127 events are detected by both methods, seven by CCEH only, one by CITE-R only, and 69 by neither. These paired counts explain the net difference of six detections in Table 2.

Table S23. Exploratory paired camera-cluster percentile intervals for fixed operating points. Differences are CITE-R minus the comparator. Recall differences are proportions; FAR differences have units h^{-1}.

Comparator	ΔRecall	95% interval	ΔFAR	95% interval
Score-only MLP	−0.07874	[−0.15456, −0.01526]	−1.29261	[−1.81911, −0.69383]
CCEH	−0.04724	[−0.09649, 0.00000]	−0.28725	[−0.61667, −0.03295]

The intervals use 10,000 paired resamples with random seed 20261104. The 127 positive events are grouped by 62 protocol camera identifiers; the 22 negative streams by 18 identifiers. Positive and negative cohorts are sampled separately for recall and FAR, with methods paired within each sampled camera. Intervals condition on the fixed models, thresholds, annotations and coverage. They do not account for full-pipeline training variation or a shared wildfire across different camera identifiers, and they do not constitute a comparison at equal realized external FAR.

S7.9. Spatial Association of Alarm Candidates

The frozen models were replayed on 825 selected windows, retaining 4392 candidate scores and boxes. The selection includes all 743 protocol target windows in [R,R+300 s] for the 127 primary events, onset targets, first pre-R alarms, negative-episode peaks and the original eight Figure 10 windows. The maximum window-score differences from the archive are 2.38×10^{-7} for CCEH and 2.32×10^{-6} for CITE-R; no selected-window frozen-threshold decisions change. Every Figure 10 box and crop now corresponds to that method's actual maximum-score candidate. In panel (a), CCEH uses rank 2 and CITE-R rank 0; the same original source images and decisions are retained.

Available class-0 YOLO boxes from the original source provide annotation-overlap evidence. Each maximum-score candidate is compared with the boxes in its own target frame. Missing files are treated as unavailable spatial evidence, rather than verified no-smoke observations. No new visual annotations are created for these diagnostics. The candidate-replay sample is targeted and does not estimate an overall window-level error rate.

Table S24. Spatial overlap of the first maximum-score candidate triggering each method within 300 s. One first-trigger window per method lacks an annotation file. Counts at each IoU threshold use the available positive annotations.

Method	Temporal hits	Positive GT available	IoU ≥ 0.1	IoU ≥ 0.3	IoU ≥ 0.5
CCEH	57	56	54	42	36
CITE-R	51	50	46	34	27

Among the annotated first-trigger windows, 56/56 CCEH maxima and 49/50 CITE-R maxima have nonzero intersection with a smoke box. The no-overlap CITE-R example is retained: its first trigger is window w00001702 in event F0026. Three other CITE-R candidates have IoU below 0.1 but cover approximately 85.2%, 54.6% and 100% of a supplied smoke box's area. These oversized/partly overlapping boxes illustrate why IoU and annotation-area coverage describe different aspects of localization; coverage does not replace the IoU thresholds.

Table S25. Spatially associated alarm diagnostics over the complete 300-s interval. A matched event has at least one above-threshold alarm whose maximum-score candidate meets the stated IoU. The common denominator is 127 events. These counts supplement the original temporal endpoint.

Method	Temporal hits	IoU ≥ 0.1	IoU ≥ 0.3	IoU ≥ 0.5	Unresolved
CCEH	57	55	45	39	1
CITE-R	51	47	35	30	1

Unresolved denotes a temporal hit without a known spatial match where an alarm window lacks an annotation file. All 127 intervals have some positive annotation evidence, but this does not mean every alarm frame is annotated. Requiring any above-threshold candidate, rather than the maximum candidate, gives the same counts except for CITE-R at IoU 0.3, which increases from 35 to 36 events. Event F0026 has no above-threshold candidate meeting IoU 0.1 during the interval despite an available S0 proposal meeting that threshold; it remains a temporal-hit/spatial-mismatch example. The original Table 2 values are explicitly temporal hits and are not relabelled as universally correct spatial detections.

Table S26. Observed S0 proposal coverage using available smoke boxes. Coverage means at least one cached proposal meets the IoU threshold at the indicated target or at some target in the interval; it does not require an alarm.

Scope	Events	Positive GT available	IoU ≥ 0.1	IoU ≥ 0.3	IoU ≥ 0.5
R or next target	127	119	84	78	56
Complete 300-s interval	127	127	116	113	107

At the first scoreable target at or after R, eight events lack a label file; later targets provide positive annotations in each of those intervals. The complete-interval counts describe observed proposal coverage, not an absolute attainable-recall bound, because some frames remain unannotated. CITE-R verifies the frozen proposal set and cannot generate a smoke proposal absent from that set. Remaining spatial mismatches and incomplete annotations are therefore reported alongside the temporal alarm results.

S7.10. Retained Physical-Time Controls and Threshold Sensitivity

The retained seed-44 Full, Same-Anchor No-Time, and Observation-Indexed checkpoints from Table S13 were evaluated without retraining or checkpoint reselection. Each method was calibrated on the same 37 reviewed non-FIgLib development-negative units, permitting at most one unit to trigger, before its FIgLib outcome evaluation. The common source-out population contains 127 positive events and 22 verified negative streams with 27.8506 covered camera-hours. The event endpoint and 120-s episode-merging rule are unchanged.

Table S27. Retained seed-44 source-out operating points under the one-of-37 development-negative-unit constraint. Hits use a denominator of 127; FAR is episodes per covered camera-hour.

Model	Threshold	300-s hits	Episodes	FAR (h^{-1})
CCEH	0.977892816	57	16	0.5745
Full / PMTR	0.987993717	45	5	0.1795
Full / HSRC	0.895880342	51	8	0.2872

Model	Threshold	300-s hits	Episodes	FAR (h^{-1})
No-Time / PMTR	0.992080271	36	5	0.1795
No-Time / HSRC	0.986705005	24	6	0.2154
Indexed / PMTR	0.992080867	45	8	0.2872
Indexed / HSRC	0.984636784	18	4	0.1436

Full-HSRC detects more events than either retained time control at these frozen points, with a higher realized FAR. Full-minus-No-Time and Full-minus-Indexed recall differences are 0.2126 and 0.2598. Exploratory paired camera-identifier bootstrap intervals are [0.1293, 0.3044] and [0.1915, 0.3359], using 10,000 resamples over 62 positive and 18 negative camera identifiers. These intervals describe sampling uncertainty for the fixed fitted models. The observation-indexed comparison changes both retrieval and descriptor encoding.

Table S28. Complete nearby development-calibration constraints for the retained checkpoints. Each method retains its selected checkpoint; only the threshold allowing 0, 1, or 2 triggered units among the same 37 negatives changes. The full seven-method CSV accompanies this table.

Allowed units	Model	300-s hits	Episodes	FAR (h^{-1})
0/37	Full / HSRC	0	0	0.0000
0/37	No-Time / HSRC	13	3	0.1077
0/37	Indexed / HSRC	12	1	0.0359
1/37	Full / HSRC	51	8	0.2872
1/37	No-Time / HSRC	24	6	0.2154
1/37	Indexed / HSRC	18	4	0.1436
2/37	Full / HSRC	51	10	0.3591
2/37	No-Time / HSRC	37	13	0.4668
2/37	Indexed / HSRC	43	9	0.3232

Under the zero-trigger constraint, Full-HSRC has no 300-s event hits and no negative-stream episodes. Its highest and second-highest development-negative unit maxima are approximately 0.986484 and 0.895880. Crossing this gap changes the transferred operating point substantially. The neighboring-constraint analysis is retrospective, holds checkpoints fixed, and reports the full grid rather than a target-selected optimum. The CPU-recalibrated Full threshold, 0.8958803415, differs from the archived 0.8958815336 by about 1.2×10^{-6}; all 23,689 original Full window decisions are unchanged.

S7.11. Matched PMTR and HSRC Retraining

Full, Same-Anchor No-Time, and Observation-Indexed were trained with seeds 40, 41, and 42, yielding nine PMTR–HSRC pairs. Each HSRC uses its corresponding newly trained PMTR. The archived model definitions, losses, data, and selection rules are retained; training record weights are normalized to mean one as in the original program. Per-stage seed resets match initialization across variants and are distinct from the historical seed-44 random-number trajectory. PMTR uses validation AP/AUC with patience 80, and HSRC uses constrained validation recall followed by group AP and AUC. The maximum is 500 epochs per stage. The accepted runs use CUDA on an RTX 4090 with one CPU thread and the retained 1061/183 development records. FIgLib outcomes do not enter checkpoint selection.

Table S29. Selected HSRC development results after matched retraining. AP/AUC are mean ± sample SD across three seeds; hits follow seed order 40/41/42 and use 28 positive groups with one allowed negative-group trigger among 27 negative groups.

Variant	Group AP	Group AUC	Hits (40/41/42)
Full	0.76674 ± 0.00737	0.78968 ± 0.00867	2/2/2
Same-Anchor No-Time	0.68671 ± 0.01695	0.72928 ± 0.01200	2/1/2
Observation-Indexed	0.68391 ± 0.01193	0.72354 ± 0.01050	2/2/2

Thresholds for all selected models were frozen on the 37 reviewed development-negative units before source-out scoring, with allowed trigger counts 0, 1, and 2 and a primary count of 1. Checkpoints were not reselected for the neighboring constraints. The same 127 events, 22 negative streams, and covered exposure as Table 2 were used. The original CCEH 57-hit/16-episode and CITE-R 51-hit/8-episode results were reproduced separately. The retained seed-44 and fixed-PMTR ten-head experiments are not pooled into the matched three-seed average.

Table S30. All matched-retraining source-out results. Hits and episode counts follow seed order 40/41/42. Recall is mean ± sample SD in percent/percentage points; FAR is mean ± sample SD in h^{-1}. The primary development constraint is 1/37. No-Time denotes Same-Anchor No-Time; Indexed denotes Observation-Indexed.

Allowed units	Model	Hits (40/41/42)	Episodes (40/41/42)	Recall (%)	FAR (h^{-1})
0/37	Full / PMTR	9/14/6	1/1/1	7.61 ± 3.18	0.03591 ± 0.00000
0/37	Full / HSRC	2/2/2	1/1/1	1.57 ± 0.00	0.03591 ± 0.00000
0/37	No-Time / PMTR	14/15/24	2/0/1	13.91 ± 4.34	0.03591 ± 0.03591
0/37	No-Time / HSRC	11/2/17	2/0/3	7.87 ± 5.94	0.05984 ± 0.05485
0/37	Indexed / PMTR	16/9/11	1/0/0	9.45 ± 2.84	0.01197 ± 0.02073
0/37	Indexed / HSRC	9/10/11	0/2/1	7.87 ± 0.79	0.03591 ± 0.03591
1/37	Full / PMTR	29/18/7	5/1/2	14.17 ± 8.66	0.09575 ± 0.07474
1/37	Full / HSRC	4/7/7	2/3/2	4.72 ± 1.36	0.08378 ± 0.02073
1/37	No-Time / PMTR	15/35/27	2/2/2	20.21 ± 7.93	0.07181 ± 0.00000
1/37	No-Time / HSRC	12/15/22	2/2/7	12.86 ± 4.04	0.13166 ± 0.10365
1/37	Indexed / PMTR	35/16/25	8/3/7	19.95 ± 7.48	0.21544 ± 0.09500
1/37	Indexed / HSRC	20/11/11	2/2/1	11.02 ± 4.09	0.05984 ± 0.02073
2/37	Full / PMTR	48/37/30	10/4/5	30.18 ± 7.14	0.22740 ± 0.11542
2/37	Full / HSRC	50/26/44	13/6/7	31.50 ± 9.83	0.31118 ± 0.13594
2/37	No-Time / PMTR	15/47/30	2/5/3	24.15 ± 12.61	0.11969 ± 0.05485
2/37	No-Time / HSRC	18/44/27	5/8/9	23.36 ± 10.40	0.26331 ± 0.07474
2/37	Indexed / PMTR	52/37/42	14/11/13	34.38 ± 6.01	0.45481 ± 0.05485

Allowed units	Model	Hits (40/41/42)	Episodes (40/41/42)	Recall (%)	FAR (h^{-1})
2/37	Indexed / HSRC	34/16/16	10/6/7	17.32 ± 8.18	0.27528 ± 0.07474

At the primary 1/37 constraint, mean HSRC event recall is 4.72% for Full, 12.86% for No-Time, and 11.02% for Indexed; mean FAR is 0.08378, 0.13166, and 0.05984 h^{-1}. The Full design has higher selected development AP/AUC but does not show a stable source-out recall advantage at this extreme alarm constraint. Equal development-unit constraints do not ensure equal realized external FAR. The original seed-44 results and these newly initialized pairs are reported separately.

Table S31. Paired training-seed differences, Full-HSRC minus the control. Values are mean ± sample SD over matched seeds 40/41/42. Recall differences are percentage points; FAR differences are h^{-1}.

Allowed units	Control	ΔRecall (pp)	ΔFAR (h^{-1})
0/37	No-Time	-6.30 ± 5.94	-0.02394 ± 0.05485
0/37	Indexed	-6.30 ± 0.79	0.00000 ± 0.03591
1/37	No-Time	-8.14 ± 3.18	-0.04787 ± 0.11542
1/37	Indexed	-6.30 ± 5.46	0.02394 ± 0.02073
2/37	No-Time	8.14 ± 20.20	0.04787 ± 0.20730
2/37	Indexed	14.17 ± 7.22	0.03591 ± 0.06219

At 1/37, Full-minus-No-Time recall differences are −6.30/−6.30/−11.81 percentage points; Full-minus-Indexed differences are −12.60/−3.15/−3.15. At 2/37, Full-minus-No-Time differences are +25.20/−14.17/+13.39 points. All budgets and seeds are retained; the primary constraint is not changed after inspecting FIgLib performance.

With all nine fitted PMTR–HSRC pairs held fixed, 10,000 paired camera-identifier resamples give 95% recall-difference intervals of [−12.15, −4.72] and [−11.02, −1.79] percentage points against No-Time and Indexed. The FAR intervals are [−0.15541, 0.04584] and [−0.04956, 0.10667] h^{-1}. These conditional sampling intervals differ from the training-seed SDs in Table S31. Cross-camera physical-fire linkage is unavailable.

S7.12. Calibration Budgets and Alarm Occupancy

The original seed-44 CCEH, score-only calibration head, and CITE-R checkpoints were retained. For each method, thresholds were calculated from the maximum scores of the same 37 reviewed development-negative units, permitting 0, 1, or 2 units to trigger. Each threshold is the next representable value above the corresponding ranked negative-unit maximum. Models and checkpoints are fixed throughout this retrospective sensitivity analysis. No external outcome is used to choose a preferred budget.

Table S32. Three-method source-out results under the complete development-budget grid. Event hits use 127 positive events; episode FAR uses 22 negative streams and 27.8506 covered camera-hours. The primary constraint remains 1/37.

Allowed units	Method	300-s hits	Episodes	FAR (h^{-1})
0/37	CCEH	49	9	0.3232
0/37	Score-only	49	9	0.3232
0/37	CITE-R	0	0	0.0000
1/37	CCEH	57	16	0.5745
1/37	Score-only	61	44	1.5799
1/37	CITE-R	51	8	0.2872
2/37	CCEH	60	19	0.6822
2/37	Score-only	70	60	2.1544
2/37	CITE-R	51	10	0.3591

The complete grid quantifies sensitivity of the frozen alarm decisions to the development budget. Thresholds, paired event outcomes and 10,000-resample camera-identifier intervals accompany all three budgets in the companion CSV files, using the resampling scheme in S7.8. Table S28 gives the corresponding time-control comparison.

Table S33. Alarm-window occupancy at the primary 1/37 constraint. The denominator includes all 1646 target windows in the 22 verified negative streams, including windows with no candidate.

Method	Alarm windows	All windows	Occupancy (%)	Triggering streams
CCEH	30	1646	1.8226	7/22
Score-only	63	1646	3.8275	12/22
CITE-R	15	1646	0.9113	2/22

CITE-R also halves the number of alarming target windows relative to CCEH. Occupancy here describes the sampled observations; it does not assign an alarm state to unobserved wall-clock intervals. Per-stream occupancy and episode counts for every budget are supplied alongside the aggregate results.

S7.13. Retained Supervision Groups and Aggregation

The transition manifest defines positive groups by one confirmed R-anchor candidate and negative groups by the retained transitions of one reviewed no-smoke unit. Groups have variable duration. All 1244 records match the audit, original manifest and multiscale cache in order, split, label, group and native interval. Every retained box has finite positive area, with a strictly earlier native observation in the same frame-parent directory. Companion CSVs give exact frames and boxes.

Table S34. Candidate-bearing supervision groups and first-to-last retained target spans. Native history ages are separate inputs. A zero span denotes a single retained target, not a zero elapsed-time transition.

Split and label	Groups	Records	Distinct targets	Median span (s)	Range (s)
Train negative	147	1020	548	44	0–2961
Train positive	41	41	41	0	0
Validation negative	27	155	92	72	0–507
Validation positive	28	28	28	0	0

The review archive contains 329 negative units; 174 contribute retained transitions. Their exact membership is supplied, so this training population is distinguishable from the 37-unit validation calibration population, which includes units without a retained training transition. Of 75 reviewed onset events, 69 supply a confirmed R candidate. Two training events and one validation event have no R candidate, while three training associations are rejected. Thus training uses 41 positive groups and validation uses 28. The validation event with no R candidate remains an undetected case in the 29-event visibility-boundary evaluation.

PMTR assigns each training record an inverse candidate-frequency class coefficient $n/(2n_y)$, multiplied by the inverse number of retained records in its group, then normalizes record weights to mean one. This equalizes group mass within a class; it does not equalize the total mass of the two classes. HSRC uses all 188 training groups in each epoch, with mean positive-group and mean negative-group BCE each weighted by 0.5. Its ranking term selects the hardest $\text{ceil}(0.30\times147)=45$ negative groups; its candidate auxiliary term selects the hardest $\text{ceil}(0.15\times1020)=153$ negative candidates and all 41 positives. PMTR and HSRC contain 152,739 and 1761 trainable parameters, respectively, excluding frozen upstream modules; zero-filled time controls retain these counts.

Table S35. Fixed seed-44 HSRC aggregation on 55 validation supervision groups. Smooth aggregation uses temperature 0.25. Each rule selects its own threshold allowing at most 1/27 negative groups to trigger.

Aggregation	Group AUC	Group AP	Positive hits	False groups
Maximum	0.76323	0.73990	13/28	1/27
Smooth training rule	0.76587	0.74110	13/28	1/27

All 28 positive groups contain a single candidate, giving identical aggregations. Among negative groups, maximum minus smooth logit has median 0.04153 and 95th percentile 0.13235. Applying the smooth-selected threshold to maximum scores triggers 3/27 negative groups. The surrogate therefore supports differentiable training but does not justify interchanging thresholds. Deployment uses maximum candidate scoring and its separate 37-unit maximum-score calibration, as in the original evaluation.

S7.14. Exposure Timestamps and Source-Out Histories

For each of the 22 verified negative streams, exposure is reconstructed from unique target-frame timestamps recorded in processed_frames. Sorted adjacent timestamps contribute their interval only when $0 < \Delta t \leq 180$ s. No time is added before the first or after the last processed target, and larger gaps contribute zero. Four-observation windows are not added together, so their overlap does not duplicate exposure. The 1646 unique target timestamps sum to 100,262 s, or 27.8505556 covered camera-hours, reproducing every original per-stream denominator. Each retained negative stream has one contiguous group; none of its adjacent processed-target gaps exceeds 180 s. This database-covered measure describes sampling coverage. It does not establish a continuously observed alarm state between images.

The companion files list every negative-stream timestamp, candidate count, adjacent interval, inclusion decision, and per-stream covered exposure. The six invalid-stream exclusions remain as documented in S7.4. Camera identifiers cover 18 negative cameras; dependence from a shared physical fire across different identifiers is not resolved by those labels.

The source-out feature index contains 24,670 frames. The original scorer's two history-lookup functions were replayed using this index and the frozen four-observation protocol, without reading feature tensors or changing scores. An independent full-directory search reproduces all lookup choices. A candidate's native predecessor occupies slot 0. Optional targets of 5, 15, 30, 60, and 120 s use the nearest strictly earlier available frame in the same parent directory, within max(3 s, 0.5×target lag), and a frame already used by another slot is not duplicated.

Table S36. Valid histories after deduplication in the FIgLib protocol. Ages are actual elapsed seconds. Optional 5-, 15-, 30-, and 60-s slots contribute no additional unique frames in this population; the native predecessor already supplies the approximately 60-s history.

Window population	History	Valid / total	Median age (s)	Age range (s)
All protocol targets	Native	23689/23689	60	36–120
All protocol targets	120-s optional	22941/23689	120	114–179
Candidate-bearing	Native	15646/15646	60	36–120
Candidate-bearing	120-s optional	15129/15646	120	114–177

There are two valid slots in 22,941 windows and one in 748, giving a mean of 1.9684 slots per protocol target and 1.9631 when weighted by the 91,416 candidates. Of these 23,689 windows, 8043 have no candidate and retain zero alarm scores. All formed windows, including the 15,646 candidate-bearing windows, have current and native-predecessor features. This denominator concerns the formed protocol and excludes any observations omitted before four-observation windows were constructed. These history statistics complement the development distributions in S7.3; controlled missing observations are evaluated in S7.19.

S7.15. Early Alarms and R Candidate Association

The original frozen seed-44 checkpoints and threshold (0.8958815) were replayed on the 117 endpoint windows of the 29 development validation events. All event decisions in Table S10 were reproduced; the largest absolute score difference was 0.000002. The retained scoring includes 80 temporal windows, 15 CCEH fallbacks without valid native history, and 22 zero-candidate windows. No model or threshold was refitted. These events are separate from the 127 FIgLib source-out events in S7.9.

Table S37. Development-event cross-tabulation. Early means an alarm at or before the reviewed no-smoke boundary L; R is the first confirmed visible-smoke boundary. Counts use temporal alarm criteria.

Early alarm at or before L	R alarm	No R alarm	Total
Yes	9	3	12
No	5	12	17
Total	14	15	29

Of the 14 R alarms, 13 maximum-score candidates agree with the retained reviewed R-event anchor. In courmettes-160:g108, the triggering candidate scores 0.923964 and does not overlap that anchor, which scores 0.095563 and remains below threshold. The original temporal result is therefore 14/29, while the anchor-associated R-alarm count is 13/29. The differing candidate is retained in the companion records.

Table S38. Spatial and sampled-continuity diagnostics. Reviewed-anchor agreement and original-annotation box overlap are distinct checks.

Diagnostic	Count
R-alarm maximum agrees with reviewed R anchor	13/14
R-alarm frames with nonempty original annotation files	9/14
R-alarm frames with empty original annotation files	5/14
Nonempty-file R maxima with class-agnostic source-box IoU ≥ 0.5	9/9
Early-and-R events alarming at every retained observation from first early alarm through R	6/9

The nine nonempty-file overlaps range from 0.7744 to 0.9369. This secondary geometry check preserves source class IDs 0 and 1 and is explicitly class-agnostic; no cross-source semantic remapping is assumed. Empty files are not new visual no-smoke judgments. Reviewed-anchor agreement uses the existing review decisions, without a new blinded relabeling.

Among the nine early-and-R events, six have no observed below-threshold gap and three have at least one. The first-early and R maximum-score boxes have normalized-coordinate IoU ≥ 0.1/0.3/0.5 in 9/6/3 events. Continuity and overlap flag possible persistence but do not establish a shared nuisance identity. The endpoint protocol ends at R; no post-R trajectory or alarm state between sampled observations is inferred. Per-event outcomes, all window maxima, and the nonmatching R candidates accompany this section.

S7.16. Source Records and Study Units

Every retained initial Pyro-SDIS image was matched by its original name to a unique record in the source Parquet metadata: 20,568 training and 774 validation images, with no unmatched records. Training and validation contain 31 and 3 source camera identifiers with no overlap; retained source partners are SDIS-07, SDIS-77, and Force-06, and all capture dates are in 2024. Initial-training frame names and camera labels do not intersect the non-FIgLib temporal validation protocol. Neither initial subset shares these identifiers with FIgLib, whose protocol timestamps span 2016–2021.

Stage-to-stage reuse within a development role is recorded explicitly: 20,418 initial-training frame names recur in temporal training, and 774 initial-validation frame names recur in temporal validation. Counts of images, protocol groups, supervised transitions, and reviewed onset events refer to distinct units (Tables 1, S14, and S18).

Table S39. Available identifiers in the retained input and evaluation protocols. Camera counts are source/protocol labels; groups and streams are defined study units, not verified counts of distinct physical ignitions.

Population	Inputs or evaluation units	Camera labels	Grouping
Initial Pyro-SDIS train	20,568 images	31	Image-source records
Initial Pyro-SDIS validation	774 images	3	Image-source records
Non-FIgLib temporal train	29,822 windows	101	445 sequences; 2363 contiguous groups
Non-FIgLib temporal validation	7406 windows	8	40 sequences; 653 contiguous groups
FIgLib input protocol	23,689 windows	101	315 sequences; 327 contiguous groups
FIgLib positive evaluation	127 reviewed events	62	Sequence and contiguous-group onset units
FIgLib negative evaluation	22 streams; 1646 windows	18	Reviewed no-smoke streams

The image snapshot has 4699 files, including 4228 with annotation files and 471 without (Table S2). The downloaded annotated-image split is 3298/441/489 and remains unchanged. The video inventory has 592 directories, 590 distinct directory names, and 80,727 frames. Storage organization need not match the 1049 sequences summarized in the source publication [9]. A distinct official release tag is absent from the retained download records; the experimental snapshot is specified by its file inventories and input protocols.

A common physical-fire identifier across cameras is not retained for these combined sources. Static image manifests also do not provide an independently verified camera identity for every image; camera counts are therefore reported only where explicit source/protocol labels are available. Disjoint frame paths and camera labels establish the audited identifier separation, without establishing that every different camera observed a different physical fire. Camera-clustered intervals are conditional on these available grouping labels. The companion source mappings retain the relevant identities and counting rules.

S7.17. Matched Temporal and Presence Comparisons

A new causal ResNet50–LSTM verifier adapts the public PyroNear temporal architecture [9] to the frozen S0 proposals and four-observation candidate protocol. ImageNet-pretrained ResNet50 features are newly extracted; the backbone remains frozen. A fresh single-layer LSTM with 256 hidden units, dropout 0.2, and a linear classifier is trained with seed 42, batch 16, Adam (lr 1e-5, wd 1e-4), and a 50-epoch cosine schedule ending at 1e-6. Development candidate AP selects epoch 3. The same 37 reviewed development-negative

units select its threshold, allowing one trigger. The historical reused epoch 7 CNN–LSTM remains a separate endpoint reference.

Table S40. Final alarm comparisons under the shared development-negative calibration rule. All methods use 127 positive events, 22 negative streams, and 27.8506 camera-hours. C4 and CCEH primary head seed is 44; the newly trained LSTM seed is 42.

Method	300-s hits	Recall (%)	Episodes	FAR (h^{-1})
S0 detector	38/127	29.92	4	0.1436
C4 simpler presence	62/127	48.82	13	0.4668
CCEH / C5	57/127	44.88	16	0.5745
Matched ResNet50–LSTM	45/127	35.43	12	0.4309
Original CITE-R	51/127	40.16	8	0.2872

The matched LSTM produces 45 detections and 12 episodes, versus 51 and 8 for CITE-R; within 60 s the counts are 6 and 12. This is a public-architecture comparison adapted to the present candidate/event setting, not a reproduction of the source benchmark population. C4 has 165,139 parameters and C5 has 165,124, with the same frozen R7 inputs. The simpler C4 primary operating point has more detections and fewer episodes than C5; the data do not establish uniform final-alarm superiority of the relation head.

Table S41. All retained C4/C5 head seeds. Thresholds are selected on the common development-negative units; seed 44 is primary.

Head and seed	300-s hits	Episodes	FAR (h^{-1})
C4_s42	76/127	21	0.7540
C4_s43	76/127	24	0.8617
C4_s44	62/127	13	0.4668
C5_s42	35/127	8	0.2872
C5_s43	46/127	10	0.3591
C5_s44	57/127	16	0.5745

A separate protection comparison trains fresh C5, PMTR, and corresponding HSRC heads for the archived R6 (unprotected physical-time adaptation) and R7 (candidate-protected) feature generators. Initialization is matched and training/selection uses development data. The new C5 uses class-balanced BCE, so this experiment is kept separate from the historical main pipeline. R7 presence AP is 0.8611 versus 0.8520 for R6, but final event-level behavior under transferred low-false-alarm thresholds differs.

Table S42. Matched R6/R7 downstream retraining at the one-of-37 development-negative-unit constraint. Entries are event hits/127 and false-alarm episodes.

Feature condition	Presence	PMTR	HSRC
R6 without protection	73/127; 22	53/127; 14	21/127; 5
R7 with protection	37/127; 13	2/127; 1	1/127; 1

The final matched comparison gives 21 detections/5 episodes for R6 and 1/1 for R7. It does not establish an independent final-alarm benefit of protection under this new-head training protocol. Candidate-level evidence and final-alarm evidence are therefore reported separately. The original frozen pipeline remains 51/127 with 8 episodes; none of these new comparison heads replaces that model.

S7.18. Complete Online Computational Profile

The complete frozen implementation is profiled on an NVIDIA RTX 4090 at batch 1, from raw-image decoding to final window score. Three development subsequences represent low, median, and high archived candidate loads. After 5 warmup targets per subsequence, 30 consecutive targets per subsequence are measured: 90 targets per resolution. Newly required historical features are computed from images in a bounded causal cache; archived feature and protection-mask databases are not read inside timed windows.

Table S43. Complete-pipeline latency and memory. Throughput counts target windows, including on-demand causal history computation. CPU RSS includes initialization and reference validation.

Metric	1024 pixels	640 pixels
Measured target windows	90	90
Mean latency (ms)	70.62	49.96
Median latency (ms)	71.24	49.41
p95 latency (ms)	73.39	53.63
Target windows/s	14.16	20.02
First-window mean (ms)	207.31	132.99
Peak GPU allocated (MiB)	312.26	235.97
Peak GPU reserved (MiB)	350.00	268.00
Process peak RSS (MiB)	1607.69	1663.23

Table S44. Separate synchronized module timings in milliseconds. Exclusive module measurements use a separate pass; their sum need not equal uninterrupted latency.

Module	1024 mean	1024 p95	640 mean	640 p95
Image decode	2.819	3.056	2.859	3.083
S0 square features/pooling	18.207	19.489	9.648	10.018
Protection proposals	0.620	0.674	0.657	0.671
Target candidate pass	39.121	41.072	25.421	26.215
Target postprocessing	1.233	1.360	1.333	1.357
TDNRI references	4.758	6.568	5.051	5.646
CCEH context/head	1.693	1.984	1.848	1.883
PMTR history/head	1.977	2.316	2.127	2.185
HSRC and window maximum	0.445	0.506	0.483	0.497

Feature extraction uses FP32 with cuDNN TF32; square protection proposals and the separate rectangular target-proposal path use AMP-FP16; downstream heads use FP32. All required S0 passes are included. GPU

memory is PyTorch allocated/reserved memory rather than total driver memory; CPU RSS is a process high-water mark. Raw files are read from local disk with the OS cache intact. These results quantify the tested implementation and hardware. The 640-pixel run is a runtime measurement without a corresponding event-accuracy claim. The detector-only Table S6 is a separate profile on an RTX 4090 D.

S7.19. Controlled Missing Observations

The retained seed 44 Full, Same-Anchor No-Time, and Observation-Indexed models are evaluated with 3 fixed physical-observation masks (seeds 20261106, 20261107, 20261108), each removing 25% or 50% of observations. Masks are shared across variants, and each 50% removal set contains its 25% set. Original weights and development thresholds are fixed. All 127 events and original target instants remain in the evaluation; missing current observations or native predecessors produce no alarm. Optional histories are reselected from retained observations. This tests the temporal stage with frozen upstream candidate/context features, without reconstructing TDNRI states under image loss.

Table S45. Recall and episodes under controlled observation removal. Means ± sample SD are across 3 observation masks, not training seeds; recall uses all 127 events. The 0% row is the retained baseline.

Removal	Method	300-s hits	Recall (%)	Episodes
0%	CCEH	57.00	44.88	16.00
0%	Full CITE-R	51.00	40.16	8.00
0%	No-Time HSRC	24.00	18.90	6.00
0%	Indexed HSRC	18.00	14.17	4.00
25%	CCEH	44.33 ± 1.53	34.91 ± 1.20	12.33 ± 0.58
25%	Full CITE-R	37.00 ± 2.00	29.13 ± 1.57	6.00 ± 1.00
25%	No-Time HSRC	14.00 ± 1.00	11.02 ± 0.79	3.00 ± 1.00
25%	Indexed HSRC	14.33 ± 2.52	11.29 ± 1.98	3.00 ± 1.00
50%	CCEH	25.33 ± 5.13	19.95 ± 4.04	7.00 ± 2.65
50%	Full CITE-R	23.33 ± 4.51	18.37 ± 3.55	2.67 ± 2.08
50%	No-Time HSRC	5.33 ± 1.15	4.20 ± 0.91	2.00 ± 0.00
50%	Indexed HSRC	7.33 ± 2.08	5.77 ± 1.64	2.33 ± 1.15

Full has higher recall than both temporal controls under every tested mask at the frozen operating points. Absolute recall falls as observations disappear. Full detections are 35/37/39 at 25% removal and 19/23/28 at 50%. The main unthinned 51-hit/8-episode result is unchanged. All 49 method/scenario results, including PMTR-stage outputs and per-mask event decisions, accompany the table.

Table S46. Availability and exposure. Values are means across masks except the 0% baseline; conditional recall is calculated per mask for events with an eligible target in [R, R + 300 s], then averaged.

Metric	0%	25%	50%
Current observations removed	0.00	5927.33	11875.67
Missing native among retained (%)	0.00	24.93	49.85
Eligible target windows	23689.00	13333.67	5924.00

Metric	0%	25%	50%
Eligible windows (%)	100.00	56.29	25.01
Events with an eligible target by 300 s	127.00	122.00	93.67
Eligible negative exposure (h)	27.85	17.28	5.35
Full conditional recall (%)	40.16	30.34	24.86
Full episodes / eligible-target hour	0.2872	0.3474	0.4816

The native-predecessor requirement explains why usable windows fall to approximately 56.3% and 25.0% at 25% and 50% removal: both current and native observations must remain. Original scheduled negative exposure stays 27.8506 camera-hours, while eligible-target exposure falls. Both exposure conventions and missing-current/native counts are provided; fewer alarm opportunities during an outage are not interpreted as improved false-alarm control.

S7.20. Training Reproduction

The original seed-44 PMTR loop and HSRC program are replayed in the current environment, retaining the original HSRC random-number call order. PMTR again selects epoch 54, and HSRC with the retained PMTR selects epoch 288 with 13/28 validation hits at one false group. The already-used Full40 pair is separately replayed from the original programs under its stage-matched initialization: it again selects 160/222 and gives 2/28 hits. These are checks of the two recorded training trajectories.

Table S47. Training-program regression checks over all 1244 development records. Full40 head tensors are identical to the previously selected head.

Program and stage	Selected epoch	Maximum prediction difference	HSRC low-FAR hits
Original seed 44 PMTR	54	5.36e-7	—
Original seed 44 HSRC	288	2.98e-7	13/28; 1 false group
Full40 PMTR	160	0	—
Full40 HSRC	222	0	2/28; 1 false group

Saved tensors and predictions are independently checked with a NumPy forward calculation. Full40 group AP/AUC is 0.769211/0.788360. These checks reproduce the retained reference trajectory and the Full40 stage-matched trajectory separately; the lower Full40 constrained recall is retained. The direct original-program replay covers Full40, while the other selected pairs have the checkpoint, pairing, prediction, and count audits described with the matched study.

S7.21. Uncertainty of the Common-FAR Envelope

The archived threshold grids are retained for the retrospective common-FAR comparison. At each FAR upper limit, each method selects the highest observed Recall@300 s. For uncertainty, 62 positive-event camera identifiers and 18 negative-stream camera identifiers are resampled separately with replacement, with both methods paired within each sampled camera. The envelope is reselected within every resample. Table S48 gives percentile intervals from 10,000 resamples (seed 20261106).

Table S48. Descriptive common-FAR envelope and paired camera-cluster uncertainty. Actual FARs are in h^{-1}; recall differences and intervals are percentage points, CITE-R minus CCEH.

FAR upper limit	CCEH actual FAR	CITE-R actual FAR	ΔRecall (pp)	95% camera-cluster interval (pp)
0.10	0.0718	0.0718	-3.94	[-7.69, 3.70]
0.20	0.1795	0.1795	0.79	[-6.60, 3.74]
0.30	0.1795	0.2872	1.57	[-5.30, 3.64]
0.40	0.1795	0.2872	1.57	[-4.62, 3.55]
0.50	0.4668	0.4668	-0.79	[-6.67, 3.28]
0.60	0.5745	0.5745	0.00	[-32.56, 3.08]

All six original envelope point estimates are reproduced. The intervals span zero across the reported budgets. Unlike the development-frozen comparisons, this envelope reselects thresholds on the evaluation sample and does not define a deployment rule. The companion records provide the event maxima, per-stream episode counts on the threshold grid and the paired resampling code. Grouping is conditional on the available camera identifiers; physical fires shared across cameras are not identified.