This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

1 Supporting Tables

Table 1: Frozen-probe fusion and clinical read-out (single split for (b), 5-fold CV for (a)). (a) Multi-stream probe grid (per-criterion AP; N=SurgeNetXL, D=DepthAnything, S=SAM 2; concatenation fusion). Gated/attention fusion underperform (best 0.366/0.347 mAP) and are omitted; probes use a two-hidden-layer MLP, 30 epochs, AdamW. (b) Causal-bridge operating points; a false "CVS achieved" is the safety-critical error, so the main text uses the high-specificity point. At ∼2% prevalence PPV is prevalence-limited under both, NPV stays ≈0.99.

(a) Multi-stream probe grid.

Streams	C1	C2	C3	mAP
N	0.233	0.474	0.316	0.386
D	0.183	0.383	0.186	0.283
S	0.162	0.316	0.165	0.265
N+D	0.237	0.476	0.374	0.395
N+D+S	**0.231**	**0.476**	**0.374**	**0.398**

(b) Operating points.

Point	Sens	Spec	PPV	NPV	AUROC
Probe (hi-spec)	0.06	0.98	0.07	0.98	0.75
Bridge (Youden)	0.59	0.91	0.12	0.99	0.85
Bridge (hi-spec)	0.29	0.97	0.19	0.99	0.85

Table 2: Temporal sweep and label-propagation ablation (single split; (a) 5-fold CV). (a) Temporal-local causal-window sweep (best epoch per window); mAP peaks at a 2–3 frame window and degrades for larger ones. (b) Propagation and confidence weighting help; the hazard term does not at this label density.

(a) Causal-window sweep.

Setting	Epoch	Acc	F1	mAP
$w = 1$	25	0.723	0.451	0.557
$w = 3$	24	**0.862**	**0.561**	**0.576**
$w = 5$	22	0.802	0.493	0.525
$w = 7$	25	0.866	0.519	0.548

(b) Label propagation.

Configuration	mAP	F1
Anchors only	0.368	–
+ propagation	0.388	–
+ confidence	**0.444**	–
+ hazard term	0.393	–
Causal FT (2 blk, 20 ep)	0.367	0.370
+ layer-wise LR (6 blk)	0.402	0.447

2 Sparse Temporal Attention: Interpretability

The sparse model is trained in two stages under the causal local window and evaluated at fixed checkpoints (Table 2a). Analyzing attention weights at epoch 20 across a 20-video sample (attention matrices, lag profiles, and aggregate statistics), the narrow per-lag range, (lag-0 spans only 0.437–0.544) and strict decay across lags 1 and 2 confirm that short-horizon aggregation is a stable learned behavior: the model prioritizes immediate context over long-range dependence.

3 Real-time System: Details and Analysis

Causal label propagation. Since each 90-second clip has only 18 labeled anchor frames, we make a dense, strictly causal inference from sparse anchors. Each anchor's three ratings are fused with confidence weights into a soft label; consensus strength and mean rater confidence form a persistence confidence γ. The propagated label relaxes exponentially towards 0.5—a stale label should express lack of information, not the negative class—at a rate set by anchor distance and an accumulated visual-instability hazard (λ=0.05). Relaxation is asymmetric (τ^+=100 vs. τ^-=50, both extended by $\beta\gamma$, β=60) because landmarks emerge gradually but occlude abruptly. Each frame's training weight is its confidence $2|\hat{l} - 0.5|$; Table 2b ablates these terms, the hazard term being retained for interpretability only.

Streaming causal bridge and its ceiling. The deployed model keeps SurgeNetXL (DINOv3 ViT-L, 336^2, 1024-d) frozen, encoding each frame once (≈13.5 ms); a causal ring buffer feeds a 3-layer masked Transformer (d_m=256, 4 heads, dropout 0.3), trained with confidence-weighted per-criterion BCE (`pos_weight` = 5) on propagated soft targets and evaluated only on anchors. The head adds <1 ms and, since the optimal window is short, latency is independent of clip length; the unmasked bidirectional model is used only to measure the cost of causality. The best causal fine-tuning configuration unfreezes the final six blocks with layer-wise learning-rate decay after a 5-epoch head warm-up. Two architectures meant to beat the bridge—an anchor-conditioned GRU memory with a learned hazard gate and approach loss (ACM), and the same bridge trained on full-theory labels, both reach 0.46 ± 0.03, below the simple bridge, so the 18-anchor label budget, not the temporal model, is the binding constraint.

Edge deployment and omitted variants. Distilling the frozen bridge into MobileNetV3-Small (0.5 soft + 0.5 hard BCE, 224^2, 20 epochs) yields mAP 0.44 at 4.3× lower latency with better calibration (ECE ≈ 0.07). The variants we omit each fail informatively. Frozen BiGRU / Transformer heads over the 18-frame anchor sequence reach only 0.44 / 0.45 because the anchor stride discards the short-horizon dynamics the native-rate bridge exploits. A temporal student trained on its own propagated pseudo-labels scores well against those labels but only ≈0.40 on true anchors, confirming the propagation adds no information beyond the anchors it is derived from. Most tellingly, end-to-end LoRA+temporal training collapses to constant outputs (0.29): jointly adapting the backbone and learning temporal structure is unstable at this label density, which is precisely why the deployed system decouples the two into a frozen backbone and a separately trained bridge.