RETHINKFUSE: CROSS-WINDOW EVIDENCE FOR MICRO-EXPRESSION EVENT DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video micro-expression analysis requires an event decoder to determine whether a local response is genuine and which temporal boundary best represents it. A single-window decoder couples candidate survival and boundary construction to one temporal scale, leaving no independent evidence for either decision. We introduce **RethinkFuse**, an event decoder for frozen joint spotting-recognition backbones. Its central observation is that cross-window agreement provides evidence of event support, while boundary disagreement exposes complementary localization hypotheses. RethinkFuse decodes the same score sequence with micro and macro windows, associates compatible candidates into event clusters, constructs a stronger single-window candidate c^S and a fused candidate c^F, and uses an Adaptive Gated Fusion router to reject the cluster, keep c^S, or select c^F. The router is fitted using only training subjects in each outer fold and leaves backbone parameters and frame-level predictions unchanged. Under fully nested subject-disjoint evaluation, RethinkFuse improves the spotting-recognition synergy score (STRS) by approximately 9.6% on SAMMLV and 30.1% on CAS(ME)^3 relative to native decoding. Controlled comparisons attribute a positive STRS gain to adaptive routing on SAMMLV. On CAS(ME)^3, the corresponding interval reaches zero at the reported precision.

1 INTRODUCTION

Long-video micro-expression (ME) analysis locates sparse emotion-labeled events in untrimmed facial recordings. Their brief, weak, and variable motion makes both localization and recognition difficult amid long neutral segments (Xie et al., 2023). Unlike clip classification, this setting must identify event boundaries before assigning labels (Yap et al., 2020; Li et al., 2023; Tran et al., 2021).

Multi-scale representations, boundary models, and joint spotting-recognition networks improve features and frame-level predictions (Wang et al., 2021; Guo et al., 2023; Liu et al., 2024; Li et al., 2024; Yu et al., 2024a; Nag et al., 2019; Zou et al., 2025; Guo et al., 2025b). A decoder must still turn those predictions into events. A short window retains local responses; a long window adds context but may blur boundaries. Isolated candidates cannot reveal support from another scale, while unconditional fusion can degrade an already reliable boundary.

Figure 1 illustrates how cross-window agreement signals event support while disagreement offers alternative boundaries. This separates the decision to retain an event from the choice of its boundary: *Are isolated temporal predictions enough for ME event decoding?*

RethinkFuse decodes frozen joint spotting-recognition outputs through micro and macro windows, then associates candidates using temporal, confidence, and emotion evidence. Each paired cluster offers a stronger single-window hypothesis and a confidence-weighted fused hypothesis. Adaptive Gated Fusion (AGF), a multinomial logistic router trained only on non-held-out subjects, chooses `reject`, `single`, or `fuse`; rescue and temporal NMS finalize the events. Routing changes which inherited emotion predictions survive, so it coordinates spotting and recognition without updating the backbone, rather than guaranteeing higher Spotting F1 alone.

We evaluate frozen ME-TST+ outputs under nested subject-disjoint protocols on SAMMLV and CAS(ME)^3. On SAMMLV, relative to Legacy PairFuse, false positives fall from 193 to 150, true positives change from 49 to 46, and STRS rises from 0.1859 to 0.2025.

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.

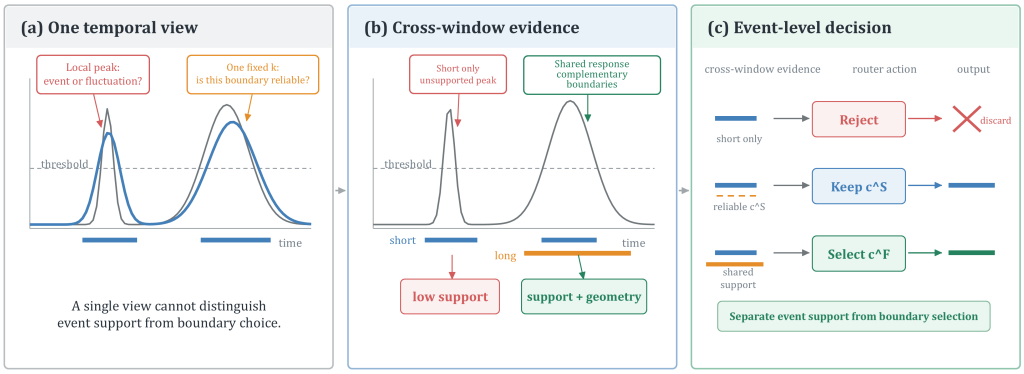

Figure 1: Cross-window evidence distinguishes unsupported peaks from shared responses with complementary boundaries. The router rejects the event or selects the single (c^S) or fused (c^F) hypothesis.

On CAS(ME)^3, STRS rises from 0.0488 for native decoding to 0.0635. Controlled action comparisons isolate routing effects; target-cache adaptation on frozen BoostingVRME outputs tests a second backbone without providing zero-shot evidence.

Our contributions are as follows:

- We formulate ME event decoding as adaptive routing over dual-window event clusters, separating candidate survival from temporal-boundary selection.
- We introduce cross-window association, alternative single and fused hypotheses, and a subject-disjoint AGF router while keeping frame-level predictions frozen.
- We evaluate system-level and controlled routing behavior under nested LOSO and examine compatibility with a second frozen backbone through explicitly identified target-cache adaptation.

2 Related Work

Long-Video Spotting. Long-video spotting locates sparse events among mostly neutral frames. SAMM Long Videos provides spontaneous expression intervals (Yap et al., 2020); CAS(ME)^3 broadens evaluation to a larger, ecologically valid setting (Li et al., 2023). Event localization, class imbalance, and cross-dataset variation remain central challenges (Tran et al., 2021; Xie et al., 2023).

Temporal, Boundary, and Structured Modeling. MESNet and the multi-scale local Transformer model temporal context at different scales (Wang et al., 2021; Guo et al., 2023). Duration-aware modeling and interval-aware embedding adapt to event extent (Liu et al., 2024; Li et al., 2024); optical-flow boundary calibration, hierarchical multi-modal fusion, and causal graphs provide other structured models (Yu et al., 2024a; Chang et al., 2025; Tan et al., 2026). These methods primarily change features, predictions, or boundary representations. RethinkFuse instead tests cross-window support after frame-level prediction.

Joint Prediction and Event-Level Decoding. Joint spotting-recognition systems use visual memory, temporal state transitions, and multi-modal fusion (Nag et al., 2019; Zou et al., 2025; Guo et al., 2025b; Zhang et al., 2026). Their probability sequences still require event extraction; TWNet, for example, uses adaptive peak detection and interval generation (Fang et al., 2026). RethinkFuse decodes frozen outputs through complementary windows, clusters cross-window candidates, and learns whether to reject, retain, or fuse their boundaries. This event-level decision complements representation-level models.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Figure 2 presents RethinkFuse as an event-level decoder above a frozen joint spotting-recognition model. It decodes one score sequence through complementary micro and macro windows, associates their candidates into event clusters, and constructs single-window and fused event hypotheses. AGF then routes each cluster to `reject`, `single`, or `fuse`. This design uses cross-window agreement without forcing fusion, and changes neither frame-level predictions nor backbone parameters.

3.2 DUAL-WINDOW CANDIDATE EXTRACTION

Let $s = \{s_t\}_{t=1}^{T}$ and $e = \{e_t\}_{t=1}^{T}$ denote the spotting scores and frame-level emotion predictions of the frozen backbone. We write the native peak decoder as $\mathcal{D}(s, e; k)$, where the half-window k determines both peak context and its initial interval. RethinkFuse applies it at two windows, $k_\mu < k_M$:

$$\mathcal{C}^{\mu} = \mathcal{D}(s, e; k_\mu), \qquad \mathcal{C}^{M} = \mathcal{D}(s, e; k_M). \tag{1}$$

The native decoder smooths each view and applies a shared Moilanen threshold, so candidate differences arise from temporal context rather than unequal detection criteria. For each video and window, the smoothed scores are clipped at their 0.05 and 0.99 quantiles and mapped to $[0, 1]$ for cross-window comparison. The micro window favors localized responses and tighter boundaries, whereas the macro window provides broader support. Their differences in peak, confidence, duration, and emotion therefore become evidence about event stability. Ordered pairs (k_μ, k_M) are selected using training subjects only.

A candidate $c_i^a = (I_i, p_i, \hat{y}_i, r_i, q_i, a)$ from window $a \in \{\mu, M\}$ contains an interval $I_i = [u_i, v_i]$, peak p_i, emotion $\hat{y}_i$, response r_i, normalized confidence q_i, and source window. Emotion is the majority frame-level prediction within I_i; exact decoder settings appear in the Experimental Setup and supplementary material.

3.3 CROSS-WINDOW ASSOCIATION AND EVENT HYPOTHESES

Candidates for the same latent event need not share an exact peak or boundary. For a micro–macro pair (c_i^{μ}, c_j^{M}), we therefore measure agreement in peak location, interval overlap, confidence, and

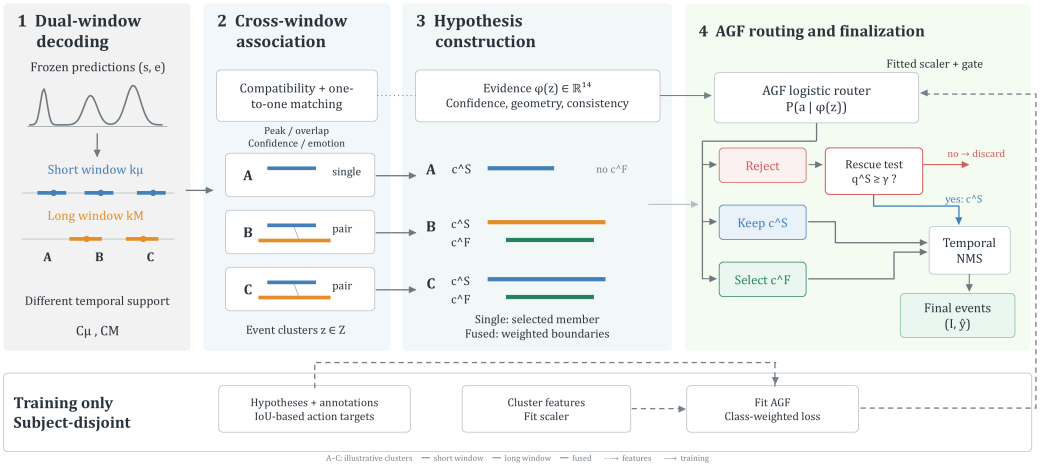

Figure 2: RethinkFuse decodes frozen outputs at two temporal scales, matches candidates into clusters, and constructs single (c^S) and fused (c^F) hypotheses. AGF rejects or selects an available hypothesis; singletons expose only c^S. Solid arrows show inference, while the dashed path fits and transfers the scaler and gate using training subjects only. Optional rescue follows rejection before temporal NMS.

emotion. With $\mathcal{M} = \{p, I, q, e\}$, their compatibility is

$$S_{ij} = \sum_{m \in \mathcal{M}} \alpha_m a_{ij}^m, \qquad \alpha_m \geq 0, \quad \sum_m \alpha_m = 1, \tag{2}$$

where peak agreement is normalized by the macro temporal extent, $a_{ij}^p = \max(0, 1 - |p_i - p_j|/(\lambda_D k_M + 1))$, $a_{ij}^I = \mathrm{IoU}(I_i, I_j)$, $a_{ij}^q = 1 - |q_i - q_j|$, and $a_{ij}^e = \mathbf{1}[\hat{y}_i = \hat{y}_j]$. The weights α_m prioritize temporal evidence, with emotion agreement auxiliary. A pair is eligible when $S_{ij} \geq \tau$ and either $|p_i - p_j| \leq \lambda_D k_M$ or $a_{ij}^I \geq \rho$. Scaling displacement by k_M accommodates temporal scale, and greedy one-to-one association prevents one response from supporting several events. Matched candidates form paired clusters; unmatched ones become singletons.

Each paired cluster exposes two possible temporal interpretations. The *single* hypothesis c^S preserves the member with higher normalized confidence q (the micro member on a tie), while the *fused* hypothesis c^F combines both boundaries. For each temporal coordinate $x \in \{u, p, v\}$,

$$x^F = \mathrm{round}\left(\frac{w^\mu x^\mu + w^M x^M}{w^\mu + w^M}\right), \qquad w^a = \max(\epsilon, q^a). \tag{3}$$

The floor ϵ ensures stable weighting while allowing the stronger member to dominate. Fused confidence combines both member scores, compatibility, and emotion agreement:

$$\begin{aligned} q^F = {} & \omega_1 \max(q^\mu, q^M) + \omega_2 \min(q^\mu, q^M) \\ & + \omega_3 S_{ij} + \omega_4 a_{ij}^e, \qquad \sum_l \omega_l = 1. \end{aligned} \tag{4}$$

The strongest response has the largest weight, while the remaining terms encode corroboration. The fused hypothesis inherits the emotion label of the member with the larger (q, r) in lexicographic order (the micro member on an exact tie); it does not re-vote on the fused interval. A singleton exposes only c^S; every cluster can still be rejected. Thus, matching creates alternatives rather than forcing fusion.

3.4 Adaptive Gated Routing

AGF formulates each event cluster z as an action-routing problem. Its feature vector $\phi(z)$ groups four forms of evidence: member confidence, temporal geometry, cross-window consistency, and emotion consistency. Features use only predicted candidates; ground truth is used solely to construct training labels.

Training labels express which action yields the most faithful event hypothesis. Let $\mathcal{G}_v$ be the annotated intervals in a training video and define $I(c) = \max_{g \in \mathcal{G}_v} \mathrm{IoU}(c, g)$, with $I(c) = 0$ when $\mathcal{G}_v$ is empty. Write $I_S = I(c^S)$ and $I_F = I(c^F)$ when fusion is available, and let $V_S = \mathbf{1}[I_S \geq \theta_{\mathrm{pos}}]$ and $V_F = \mathbf{1}[c^F \text{ exists and } I_F \geq \theta_{\mathrm{pos}}]$. The target action is

$$y(z) = \begin{cases} \text{reject}, & V_S = 0 \text{ and } V_F = 0, \\ \text{fuse}, & V_F = 1 \text{ and } (V_S = 0 \text{ or } I_F > I_S + \delta), \\ \text{single}, & \text{otherwise.} \end{cases} \tag{5}$$

The validity threshold follows the evaluation IoU criterion, while margin δ compares localization only when both hypotheses are valid. A valid fused hypothesis is selected directly when the single hypothesis is invalid.

Algorithm 1 summarizes the outer-fold procedure. The fitted gate and finalization steps are detailed below.

Algorithm 1 Subject-disjoint RethinkFuse training and inference

Input: Frozen outputs $\mathcal{O}$; outer-training subjects $\mathcal{S}_{-r}$ with annotations $\mathcal{G}$; held-out subject r; configuration pool $\mathcal{K}$
Output: Event predictions $\mathcal{P}_r$

$\vartheta \leftarrow \mathrm{InnerSelect}(\mathcal{K}, \mathcal{S}_{-r})$
$\mathcal{T} \leftarrow \emptyset$
for each training video v from $\mathcal{S}_{-r}$ **do**
  $\mathcal{Z}_v \leftarrow \mathrm{BuildClusters}(s_v, e_v; \vartheta)$
  for each cluster $z \in \mathcal{Z}_v$ **do**
    Construct c^S and, when available, c^F
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(\phi(z), y(z; \mathcal{G}_v))\}$
  end for
end for
$(N_\phi, g) \leftarrow \mathrm{FitAGF}(\mathcal{T})$
$\mathcal{Z}_r \leftarrow \mathrm{BuildClusters}(s_r, e_r; \vartheta)$
$\mathcal{P}_r \leftarrow \emptyset$
for each cluster $z \in \mathcal{Z}_r$ **do**
  $a \leftarrow \arg\max_{a' \in \{\mathrm{reject}\} \cup \mathcal{A}(z)} P_g(a' \mid N_\phi(\phi(z)))$
  if $a \neq \mathrm{reject}$ **then**
    $\mathcal{P}_r \leftarrow \mathcal{P}_r \cup \{c^a\}$
  else if $\mathrm{Rescue}(z; \vartheta)$ **then**
    $\mathcal{P}_r \leftarrow \mathcal{P}_r \cup \{c^S\}$
  end if
end for
return $\mathrm{NMS}(\mathcal{P}_r; \vartheta)$

After standardizing $\phi(z)$ on training clusters, a multiclass logistic regression estimates a distribution over $\mathcal{A} = \{\mathrm{reject}, \mathrm{single}, \mathrm{fuse}\}$:

$$\begin{aligned} \ell_a(z) &= w_a^\top \widetilde{\phi}(z) + b_a, \\ P(a \mid z) &= \frac{\exp(\ell_a(z))}{\sum_{a' \in \mathcal{A}} \exp(\ell_{a'}(z))}. \end{aligned} \tag{6}$$

Training uses class-weighted cross-entropy and a fixed regularization setting after feature standardization. The gate learns only to interpret existing event evidence, not to produce new frame-level predictions.

3.5 Subject-Disjoint Training and Inference

For held-out subject r, action targets, feature standardization, and gate fitting use only the remaining subjects. Inner LOSO selects the decoder configuration, after which AGF is refit on all outer-training subjects and applied once to the outer-test subject. No held-out annotation affects scale selection, normalization, or gate learning.

At inference, AGF compares rejection with the available boundary actions: singletons can emit only c^S, while paired clusters can emit c^S or c^F. An optional high-confidence rescue preserves exceptional single hypotheses. Scale-aware temporal NMS ranks emitted candidates by confidence, suppresses nearby peaks or overlapping intervals, and extends a retained interval to cover any suppressed interval. Its fixed settings are reported in the Experimental Setup rather than treated as learned modules.

BuildClusters covers dual-window decoding, association, and hypothesis construction; FitAGF includes standardization and logistic regression.

If the two windows produce n_μ and n_M candidates and E eligible pairs, association costs $O(n_\mu n_M + E \log E)$, and the linear gate has constant per-cluster cost; backbone inference is neither repeated nor updated.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Data and evaluation. We use the ME-TST+ protocol on SAMMLV (Yap et al., 2020) and CAS(ME)^3 (Li et al., 2023). All decoders share its frozen frame-level spotting scores and emotion predictions (Guo et al., 2025b). Matching uses IoU ≥ 0.5; we report TP, FP, FN, both F1 scores, and the Spot-Then-Recognize Score:

$$\mathrm{STRS} = F1_{\mathrm{spot}} \times F1_{\mathrm{rec}}. \tag{7}$$

Recognition F1 is the harmonic mean of macro-averaged precision and recall over matched Negative, Positive, and Surprise events, excluding matched *Others*. The result-synergy adjustment removes unmatched *neutral* predictions from FP and moves matched neutral predictions from TP to FN. Because decoders retain different events, recognition reflects both event selection and inherited labels; we report it alongside spotting and STRS.

Nested selection and uncertainty. Each outer LOSO fold holds out one subject. Inner LOSO selects the pair and rescue state from outer-training subjects; scaling and AGF fitting exclude the held-out subject. AGF is refit on all outer-training subjects for testing. SAMMLV uses 20 configurations over ten pairs; CAS(ME)^3 uses $(12, 15)$, $(11, 17)$, $(12, 17)$, $(12, 18)$, and $(12, 19)$, each with and without rescue. STRS-difference intervals (RethinkFuse minus control) use 10,000 paired subject-bootstrap resamples at 95% confidence.

Controls. Native uses its original window; Legacy PairFuse retains its training-free alignment and suppression. Nested Single-Window selects one window without association. Independent Always-Fuse selects its own pair on both datasets; Independent Always-Single does so on CAS(ME)^3.

Shared controls reuse the RethinkFuse-selected pair, clusters, hypotheses, scores, and suppression. Shared Always-Single keeps the single hypothesis; Shared Always-Fuse fuses pairs and retains singletons. No Fuse learns `reject`/`single`; No Reject learns `single`/`fuse`, subject to availability. These comparisons isolate policy differences at the shared configuration.

Decoder settings. Windows use model-output indices. Native has $k = 5$ on SAMMLV and $k = 17$ on CAS(ME)^3; both views use the Moilanen rule with $p = 0.55$. Association uses $\lambda_D = 1$, $\tau = 0.35$, and fusion margin $\delta = 0.01$. A training-subject StandardScaler feeds a multinomial logistic gate with $C = 1$, balanced weights, and at most 2,000 iterations. Selected rescue uses $\gamma = 0.98$; the single-window floor is disabled ($\beta = 0$). NMS ranks by candidate confidence with $\eta_{\mathrm{nms}} = 0.30$ and radius $\max(k_\mu, k_M)$; suppressed intervals extend the retained boundary. Further settings appear in the supplementary material.

4.2 MAIN RESULTS

Tables 1 and 2 compare RethinkFuse's fixed-reference-profile results with prior methods. S, R, and S&R denote spotting-only, recognition-only, and joint systems. The RethinkFuse reference-profile entries use different settings from the nested system comparison in Table 3. Recognition-only methods evaluate ground-truth intervals, while joint systems evaluate labels on matched predictions. These tables provide context, not a controlled ranking across methods or system types.

Table 3 compares complete decoders. On SAMMLV, RethinkFuse has the highest STRS point estimate, 0.2025 versus Native's 0.1848. FP falls from 171 to 150, TP from 52 to 46, and Spotting F1 from 0.2723 to 0.2592; higher Recognition F1 offsets the spotting loss. Relative to Legacy PairFuse, FP falls from 193 to 150 with TP changing from 49 to 46. Its STRS difference from Independent Always-Fuse remains uncertain ($[-0.0145, 0.0322]$), so the point estimate does not establish superiority over every independently selected strategy.

On CAS(ME)^3, RethinkFuse leads all three reported F1/STRS point estimates. STRS rises from Native's 0.0488 to 0.0635, with interval $[0.0053, 0.0258]$. Its margin over Nested Single-Window (0.0619) is small and uncertain. Joint performance improves relative to Native; advantages over independently selected controls depend on the comparator. Figure 3 plots selected decoders from Table 3 by Spotting F1 and Recognition F1; dashed curves mark equal STRS.

Table 1: Spotting F1 results from prior methods and RethinkFuse's fixed reference profile. The latter is separate from the nested system comparison in Table 3.

Type	Method	CAS(ME)^3	SAMMLV
S	MESNet (Wang et al., 2021)	–	0.0880
	LSSNet (Yu et al., 2021)	0.0653	0.2180
	LGSNet (Yu et al., 2024c)	0.0990	–
	OFCT (Yang et al., 2025)	–	0.2466
	Causal-Ex (Tan et al., 2026)	–	0.2010
S&R	MEAN (Liong et al., 2023)	0.0283	0.0949
	SFAMNet (Liong et al., 2024)	0.0716	–
	USTC-IAT (Yu et al., 2024b)	–	0.1800
	He et al. (He et al., 2024)	–	0.1900
	RethinkFuse (ref.)	**0.1045**	**0.2590**

Table 2: Recognition results on CAS(ME)^3. Recognition-only and joint systems use different evaluation intervals; these entries do not form a controlled ranking.

Type	Method	UF1 ↑	UAR ↑
R	STSTNet (Liong et al., 2019)	0.3795	0.3792
	RCN-A (Xia et al., 2020)	0.3928	0.3893
	FeatRef (Zhou et al., 2022)	0.3493	0.3413
	AlexNet (Li et al., 2023)	0.3001	0.2982
S&R	MEAN (Liong et al., 2023)	0.3894	0.4004
	SFAMNet (Liong et al., 2024)	0.4462	0.4767
	RethinkFuse (ref.)	**0.5097**	**0.4997**

Table 3: Compact-pool nested LOSO comparison. Independent fixed-action decoders select configurations on outer-training subjects; Native and Legacy PairFuse retain fixed rules. Independent Always-Single is evaluated on CAS(ME)^3 only.

Decoder	Configuration	TP	FP	FN	Spot. F1	Rec. F1	STRS
SAMMLV							
Native Decoder	fixed	52	171	107	**0.2723**	0.6787	0.1848
Single Window	nested	50	193	109	0.2488	0.3821	0.0950
Legacy PairFuse	legacy	49	193	110	0.2444	0.7608	0.1859
Independent Always Fuse	nested	54	196	105	0.2641	0.7322	0.1933
RethinkFuse	nested	46	150	113	0.2592	**0.7813**	**0.2025**
CAS(ME)^3							
Native Decoder	fixed	76	727	782	0.0915	0.5338	0.0488
Single Window	nested	95	778	763	0.1098	0.5642	0.0619
Independent Always Single	nested	82	785	776	0.0951	0.5528	0.0526
Independent Always Fuse	nested	78	763	780	0.0918	0.5472	0.0502
RethinkFuse	nested	91	702	767	**0.1102**	**0.5757**	**0.0635**

4.3 CROSS-BACKBONE ADAPTATION

Table 4: Exploratory adaptation on frozen BoostingVRME outputs. Configuration selection and gate fitting use the complete target cache; these results are not an independent generalization test.

Dataset	Decoder	Spot. F1	Rec. F1	STRS
SAMMLV	Native	**0.2906**	0.6882	0.2000
SAMMLV	RethinkFuse	0.2821	**0.9283**	**0.2619**
CAS(ME)^3	Native	**0.0997**	0.5638	0.0562
CAS(ME)^3	RethinkFuse	0.0735	**0.8775**	**0.0645**

On frozen BoostingVRME outputs (Guo et al., 2025a), STRS rises from 0.2000 to 0.2619 on SAMMLV and from 0.0562 to 0.0645 on CAS(ME)^3; spotting F1 falls on both (Table 4). Complete target-cache selection and fitting make this target-selected adaptation.

4.4 ABLATION STUDIES

Table 5: Controlled routing under nested LOSO. Variants share the RethinkFuse-selected pair, clusters, scores, and suppression rules. Always-Fuse retains singletons with the single action. R, S, and F denote `reject`, `single`, and `fuse`.

Router	Available actions	TP	FP	FN	Spot. F1	Rec. F1	STRS
SAMMLV							
Shared Always Single	S	41	236	118	0.1881	0.4738	0.0891
Shared Always Fuse	F	50	223	109	0.2315	0.7282	0.1686
No Fuse	R/S	36	136	123	0.2175	0.5227	0.1137
No Reject	S/F	49	225	110	0.2263	0.7270	0.1645
RethinkFuse	R/S/F	46	150	113	**0.2592**	**0.7813**	**0.2025**
CAS(ME)^3							
Shared Always Single	S	85	747	773	0.1006	0.5780	0.0581
Shared Always Fuse	F	88	751	770	0.1037	0.5629	0.0584
No Fuse	R/S	83	698	775	0.1013	**0.5854**	0.0593
No Reject	S/F	88	747	770	0.1040	0.5672	0.0590
RethinkFuse	R/S/F	91	702	767	**0.1102**	0.5757	**0.0635**

Controlled Routing Comparisons. Under shared candidate evidence, RethinkFuse raises SAMMLV STRS from 0.1686 for Always-Fuse to 0.2025 ($[0.0078, 0.0549]$; Table 5). FP falls from 223 to 150 while TP changes from 50 to 46. On CAS(ME)^3, STRS rises from 0.0584 to 0.0635, but the interval $[0.0000, 0.0111]$ touches zero at the reported precision. The difference from Shared Always-Single is also uncertain ($[-0.0227, 0.0115]$). Thus, evidence for adaptive routing is stronger on SAMMLV under the RethinkFuse-selected shared configuration.

Action-Space Ablation. On SAMMLV, removing `fuse` or `reject` lowers STRS, with difference intervals $[0.0106, 0.1394]$ and $[0.0093, 0.0614]$. On CAS(ME)^3, the No Reject interval is $[0.0006, 0.0090]$, while No Fuse crosses zero. The evidence supports both actions on SAMMLV and more clearly supports rejection on CAS(ME)^3. Aggregate action frequencies appear in the supplementary material but do not establish individual decision accuracy.

4.5 ANALYSIS AND QUALITATIVE RESULTS

Window Selection Stability. On CAS(ME)^3, 93/94 RethinkFuse folds select $(12, 18)$ with rescue. Other selections and fixed-reference sensitivity, using a different single-window floor, appear in the supplementary material.

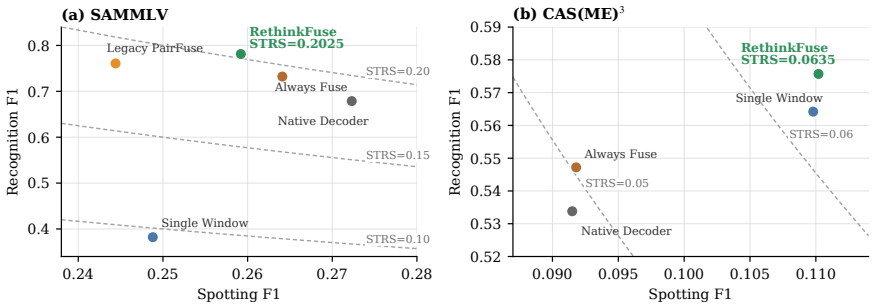

Figure 3: Selected decoders from Table 3 in spotting–recognition F1 space. Dashed curves denote equal STRS.

Qualitative Examples. Figure 4 shows two held-out SAMMLV decisions. In 011/011_5, candidates with confidences 1.000 and 0.902 have different labels and no ground-truth overlap; AGF rejects the cluster with $P_R/P_S/P_F = 0.964/0.013/0.023$. In 006/006_1, both views peak at 514; fusion changes $[511, 517]$ to $[510, 518]$ toward ground truth $[508, 518]$. AGF selects fusion with probability 0.714, raising IoU from 0.636 to 0.818; the event survives NMS. These cases illustrate rejection and localization, without establishing their frequency or superiority to a tuned threshold.

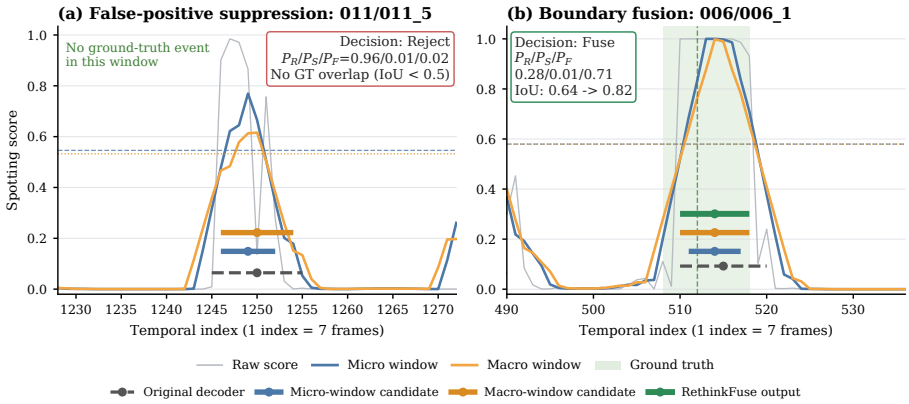

Figure 4: Held-out SAMMLV examples of false-positive rejection and boundary fusion.

5 Conclusion and Limitations

RethinkFuse adapts event decoding across complementary temporal windows, rejecting, retaining, or fusing frozen candidates to improve spotting–recognition coordination. Its limitations include inherited predictions, predefined scales, hand-crafted features, and limited unseen-domain validation. Future work will pursue learnable scales, uncertainty-aware routing, and cross-domain calibration.

AI Use Statement

Generative AI tools were used to provide feedback on the presentation of the experimental methodology, to assist with the presentation and wording of already-computed experimental results, and to prepare LaTeX/TikZ source for the paper's figures. We did not use generative AI tools to generate data, formulate theoretical or mathematical claims, write proofs, design or implement the decoder, or clean and reformat the datasets. We reviewed all AI-assisted text and artifacts against the experimental records; no numerical result or experimental output was generated or altered by AI. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics Statement

This work uses previously released micro-expression benchmark datasets and does not collect new data or involve new interaction with human participants. Micro-expression analysis can nevertheless be misused for surveillance or other sensitive inferences about individuals. We do not advocate deployment in high-stakes or non-consensual settings. Any future use should respect the datasets' licenses and consent conditions and should be assessed for privacy, fairness, and potential harms in its intended context.

Reproducibility Statement

The main paper specifies the decoder, feature construction, and nested subject-disjoint evaluation protocol, and Algorithm 1 states the outer-fold procedure. Appendix A records the candidate pools, metrics, bootstrap construction, software environment, and random seeds. The experiments reuse frozen backbone outputs and previously released benchmarks, so exact replication additionally requires access to those datasets and frame-level predictions.

References

Haolin Chang, Zhihua Xie, and Fan Yang. Micro-expression spotting based on multi-modal hierarchical semantic guided deep fusion and optical flow driven feature integration. *Complex & Intelligent Systems*, 11:238, 2025. doi: 10.1007/s40747-025-01855-3.

Huirong Fang, Bin Chen, Jiayu Zou, Sze Teng Liong, and Kunhong Liu. TWNet: A transfer learning-based framework for facial macro- and micro-expression spotting in long videos using hybrid features. *PeerJ Computer Science*, 12:e3663, 2026. doi: 10.7717/peerj-cs.3663.

Xupeng Guo, Xiaobiao Zhang, Lei Li, and Zhaoqiang Xia. Micro-expression spotting with multi-scale local transformer in long videos. *Pattern Recognition Letters*, 168:146–152, 2023. doi: 10.1016/j.patrec.2023.03.012.

Zizheng Guo, Bochao Zou, Yinuo Jia, Xiangyu Li, and Huimin Ma. Boosting micro-expression analysis via prior-guided video-level regression. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13964–13971. Association for Computing Machinery, 2025a. doi: 10.1145/3746027.3762026.

Zizheng Guo, Bochao Zou, Junbao Zhuo, and Huimin Ma. ME-TST+: Micro-expression analysis via temporal state transition with ROI relationship awareness, 2025b.

Yuhong He, Wenchao Liu, Guangyu Wang, Lin Ma, and Haifeng Li. Enhancing micro-expression analysis performance by effectively addressing data imbalance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11503–11507, 2024. doi: 10.1145/3664647.3689144.

Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2782–2800, 2023. doi: 10.1109/TPAMI.2022.3174895.

Xiaodong Li, Jiajun Li, Wenchao Du, Hu Chen, and Hongyu Yang. Learning interval-aware embedding for macro and micro-expression spotting. In *Computer Vision – ACCV 2024*, pp. 373–390. Springer, 2024. doi: 10.1007/978-981-96-0911-6_22.

Gen-Bing Liong, John See, and Chee-Seng Chan. Spot-then-recognize: A micro-expression analysis network for seamless evaluation of long videos. *Signal Processing: Image Communication*, 110: 116875, 2023. doi: 10.1016/j.image.2022.116875.

Gen-Bing Liong, Sze-Teng Liong, Chee Seng Chan, and John See. SFAMNet: A scene flow attention-based micro-expression network. *Neurocomputing*, 566:126998, 2024. doi: 10.1016/j.neucom.2023.126998.

Sze-Teng Liong, Y. S. Gan, John See, Hui-Qi Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–5, 2019.

Jing Liu, Xin Li, Jiaqi Zhang, Guangtao Zhai, Yuting Su, Yuyi Zhang, and Bo Wang. Duration-aware and mode-aware micro-expression spotting for long video sequences. *Signal Processing: Image Communication*, 129:117192, 2024. doi: 10.1016/j.image.2024.117192.

Sauradip Nag, Ayan Kumar Bhunia, Aishik Konwer, and Partha Pratim Roy. Facial micro-expression spotting and recognition using time contrasted feature with visual memory. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2022–2026, 2019. doi: 10.1109/ICASSP.2019.8683737.

Pei-Sze Tan, Sailaja Rajanala, Arghya Pal, Raphaël C.-W. Phan, and Huey-Fang Ong. Causal-Ex: Causal graph-based micro and macro expression spotting. *Pattern Recognition Letters*, 200:52–59, 2026. doi: 10.1016/j.patrec.2025.12.002.

Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, Xiaobai Li, and Guoying Zhao. Micro-expression spotting: A new benchmark. *Neurocomputing*, 443:356–368, 2021. doi: 10.1016/j.neucom.2021.02.022.

Su-Jing Wang, Ying He, Jingting Li, and Xiaolan Fu. MESNet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos. *IEEE Transactions on Image Processing*, 30:3956–3969, 2021. doi: 10.1109/TIP.2021.3064258.

Zhaoqiang Xia, Wei Peng, Hui-Qi Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29:8590–8605, 2020.

Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. An overview of facial micro-expression analysis: Data, methodology and challenge. *IEEE Transactions on Affective Computing*, 14(3):1857–1875, 2023. doi: 10.1109/TAFFC.2022.3143100.

Henian Yang, Shucheng Huang, and Mingxing Li. OFCT: A micro-expression spotting method fusing optical flow features and category text information. *Complex & Intelligent Systems*, 11: 373, 2025. doi: 10.1007/s40747-025-01997-4.

Chuin Hong Yap, Connah Kendrick, and Moi Hoon Yap. SAMM long videos: A spontaneous facial micro- and macro-expressions dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 771–776, 2020. doi: 10.1109/FG47880.2020.00029.

Jun Yu, Yaohui Zhang, Gongpeng Zhao, Peng He, Zerui Zhang, Zhongpeng Cai, Qingsong Liu, Jianqing Sun, and Jiaen Liang. Micro-expression spotting based on optical flow feature with boundary calibration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11490–11496, 2024a. doi: 10.1145/3664647.3689142.

Jun Yu, Gongpeng Zhao, Yaohui Zhang, Peng He, Zerui Zhang, Zhao Yang, Qingsong Liu, Jianqing Sun, and Jiaen Liang. Temporal-informative adapters in VideoMAE V2 and multi-scale feature fusion for micro-expression spotting-then-recognize. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11484–11489, 2024b. doi: 10.1145/3664647.3689141.

W.-W. Yu, J. Jiang, and Y.-J. Li. LSSNet: A two-stream convolutional neural network for spotting macro- and micro-expression in long videos. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4745–4749, 2021.

W.-W. Yu, J. Jiang, K.-F. Yang, H.-M. Yan, and Y.-J. Li. LGSNet: A two-stream network for micro- and macro-expression spotting with background modeling. *IEEE Transactions on Affective Computing*, 15(1):223–240, 2024c.

Jiaonan Zhang, Xinyue Wang, Yiwen Zhang, Tianxu Xu, Zhilin Zhang, Qiang Wang, Zhongqi Pan, and Yang Yue. Multi-modal fusion for robust micro-expression detection: Integrating temporal spotting and emotion classification. *Information Fusion*, 128:103995, 2026. doi: 10.1016/j.inffus.2025.103995.

Ling Zhou, Qirong Mao, Xiaohui Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022. doi: 10.1016/j.patcog.2021.108275.

Bochao Zou, Zizheng Guo, Wenfeng Qin, Xin Li, Kangsheng Wang, and Huimin Ma. Synergistic spotting and recognition of micro-expression via temporal state transition. In *ICASSP*, pp. 1–5. IEEE, 2025.

A Reproducibility Details

RethinkFuse operates on frozen frame-level outputs. The ME-TST+ backbone, its weights, and its frame-level spotting and emotion predictions are not updated by the decoder. The Native Decoder, Legacy PairFuse, fixed-action controls, and RethinkFuse reuse the same cached predictions, isolating the effect of event-level decoding. No new dataset is introduced.

A.1 Subject-Disjoint Protocol

For every outer held-out subject r, candidate-action labels, feature standardization, and AGF fitting use only the remaining subjects. Temporal-pair and rescue-state selection is performed in an inner subject-level LOSO loop. The selected gate is refitted on all outer-training subjects and applied once to the outer-test subject. Test annotations are used only for final evaluation. Temporal matching uses IoU ≥ 0.5.

SAMMLV contains 29 outer folds and evaluates the ten temporal pairs

$$(2,3),(2,4),(2,5),(2,6),(3,4),$$
$$(3,5),(3,6),(4,5),(4,6),(5,6),$$

each with rescue enabled or disabled, giving 20 profiles. CAS(ME)^3 contains 94 outer folds and evaluates the five pairs

$$(12,15),\ (11,17),\ (12,17),\ (12,18),\ (12,19)$$

with rescue enabled or disabled, giving ten profiles. All CAS(ME)^3 profiles use candidate-score NMS. Both pools are frozen before the nested run and are not modified using an outer-test result.

A.2 Metrics and Aggregation

We accumulate TP, FP, and FN over all outer-test predictions and report Spotting F1, three-class Recognition F1, and

$$\mathrm{STRS} = F1_{\mathrm{spot}} F1_{\mathrm{rec}}.$$

Recognition F1 is the harmonic mean of macro precision and macro recall over matched Negative, Positive, and Surprise events; matched *Others* annotations are excluded from this three-class calculation. The result-synergy adjustment removes unmatched *neutral* predictions from FP and moves matched neutral predictions from TP to FN. Uncertainty intervals use 10,000 paired subject-level bootstrap resamples, with competing decoders resampled together. Each resample draws outer subjects with replacement, re-aggregates TP, FP, FN, and matched emotion labels, and then recomputes all three metrics. Recognition uses a fixed three-class macro average; a class absent from a resample contributes zero. The bootstrap seeds are 2027 for SAMMLV and 100 for CAS(ME)^3.

A.3 Software Environment and Randomness

The archived decoder-side environment records Python 3.10, NumPy 1.26.4, SciPy 1.11.4, scikit-learn 1.3.2, and pandas 1.5.3. The SAMMLV nested run uses seed 100. Each reported nested result is one complete pass over all outer folds; repeated stochastic training runs are not averaged. Bootstrap resampling is applied only after outer-test predictions have been collected.

A.4 CONFIRMED IMPLEMENTATION SETTINGS

Table 6 records the fixed settings verified against the paper-aligned implementation. Only the temporal pair and rescue state are selected in the nested loop; all other values are fixed across folds and controls.

Table 6: Confirmed RethinkFuse implementation settings.

Component	Value	Role
Candidate threshold	Moilanen, $p = 0.55$	Shared peak conservativeness
Confidence range	$(0.05, 0.99)$ quantiles	Per-video, per-window calibration
Peak scale	$\lambda_D = 1$	Scale peak tolerance by k_M
Similarity weights	$.35/.25/.25/.15$	Peak, interval, confidence, emotion
Match thresholds	$\tau = .35$, $\rho = .10$	Compatibility and overlap floor
Fusion floor	$\epsilon = .05$	Stable confidence weighting
Fusion weights	$.40/.25/.25/.10$	Max, min, compatibility, emotion
Action labels	$\theta_{\text{pos}} = .50$, $\delta = .01$	Validity and fusion margin
Feature scaler	Training-subject StandardScaler	Standardize AGF inputs
AGF classifier	$C = 1$, balanced, 2,000 iter.	Multinomial logistic gate
Rescue	$\gamma = .98$ when selected	High-confidence exception
Single floor	$\beta = 0$	Disabled
Final NMS	$\eta_{\text{nms}} = .30$	Candidate-ranked interval merging
NMS radius	$\max(k_\mu, k_M)$	Scale-aware consolidation

Peak agreement receives the largest association weight; emotion is auxiliary. The validity threshold matches the evaluation IoU, and the fusion margin avoids rounding-level gains. Balanced class weights address unequal action frequencies; inner validation determines whether to enable rescue.

Each score curve is calibrated per video and window using the 0.05/0.99 quantiles, falling back to min/max and then zero for a constant curve. Final NMS ranks by normalized confidence, breaking ties with the raw peak response. It merges candidates when peak distance is at most $\max(k_\mu, k_M)$ or interval IoU reaches η_{nms}, and expands retained boundaries to their union.

For completeness, the native decoder extracts peaks from smoothed response $\bar{s}^a$ at $\xi^a = \text{mean}(\bar{s}^a) + p(\max_t \bar{s}^a_t - \text{mean}(\bar{s}^a))$. The same p is used for both windows. Peak agreement for micro candidate i and macro candidate j is $a^p_{ij} = \max(0, 1 - |p_i - p_j|/(\lambda_D k_M + 1))$; the normalization by k_M makes admissible displacement scale-aware. These definitions correspond to the candidate threshold and peak scale entries in Table 6.

A.5 AGF FEATURE DEFINITION

The paper-aligned AGF router uses a 14-dimensional vector for every event cluster. Table 7 makes the input to the logistic gate explicit. For a singleton, unavailable pair quantities are set to zero and the single hypothesis supplies both duration and confidence quantities.

Table 7: The 14-dimensional AGF feature vector.

Group	Feature
Structure	Paired-cluster indicator
Confidence	Maximum, minimum, and absolute gap of member confidence
Agreement	Pair similarity, interval IoU, normalized peak distance
Agreement	Emotion agreement and confidence agreement
Geometry	Fused and single duration, each divided by k_M
Scale	Ratio k_M/k_μ
Candidate quality	Fused confidence and maximum confidence $\times$ similarity

B Additional Ablations and Diagnostics

The ablations below isolate configuration selection, routing actions, and fixed decoder settings while preserving the frozen ME-TST+ predictions. Unless otherwise stated, all comparisons retain the subject-disjoint protocol described under Reproducibility Details. Complete-cache analyses are explicitly separated from the nested evaluation evidence.

B.1 Configuration-Selection Stability

Table 8 gives the observed outer-fold selection frequencies. Both datasets concentrate on one profile, but this concentration is a diagnostic of the frozen candidate pool rather than evidence that one temporal scale is universally optimal.

Table 8: Nested configuration-selection frequencies.

Dataset	Selected profile	Folds	Share
SAMMLV	$(3, 4)$, rescue disabled	28/29	96.6%
SAMMLV	$(4, 5)$, $\gamma = .98$	1/29	3.4%
CAS(ME)^3	$(12, 18)$, $\gamma = .98$	93/94	98.9%
CAS(ME)^3	$(12, 19)$, $\gamma = .98$	1/94	1.1%

The independently selected CAS(ME)^3 fixed-action controls show different preferences. Always Fuse selects $(12, 15)$ in 82/94 folds, followed by $(12, 17)$ in 7, $(11, 17)$ in 3, and $(12, 18)$ in 2. Always Single selects $(11, 17)$ in 86/94 folds, $(12, 15)$ in 5, and $(12, 18)$ in 3. All of these selected profiles use $\gamma = .98$. The nested single-window control selects $k = 11$ in all 94 folds. These frequencies verify that the independent controls perform their own inner selection rather than inheriting the RethinkFuse pair.

B.2 Action-Space Ablation and Paired Uncertainty

Table 9 provides the full aggregate action-space ablation. Within each dataset, all rows receive the RethinkFuse inner-selected temporal pair, candidate clusters, confidence scores, and final suppression. The ablation therefore changes the action space while holding the event evidence fixed.

Table 9: Expanded controlled action ablation under nested LOSO.

Dataset	Router	TP	FP	FN	Spot. F1	Rec. F1	STRS
SAMMLV	Shared Always Single (S)	41	236	118	0.1881	0.4738	0.0891
SAMMLV	Shared Always Fuse (F)	50	223	109	0.2315	0.7282	0.1686
SAMMLV	No Fuse (R/S)	36	136	123	0.2175	0.5227	0.1137
SAMMLV	No Reject (S/F)	49	225	110	0.2263	0.7270	0.1645
SAMMLV	RethinkFuse (R/S/F)	46	150	113	0.2592	0.7813	0.2025
CAS(ME)^3	Shared Always Single (S)	85	747	773	0.1006	0.5780	0.0581
CAS(ME)^3	Shared Always Fuse (F)	88	751	770	0.1037	0.5629	0.0584
CAS(ME)^3	No Fuse (R/S)	83	698	775	0.1013	0.5854	0.0593
CAS(ME)^3	No Reject (S/F)	88	747	770	0.1040	0.5672	0.0590
CAS(ME)^3	RethinkFuse (R/S/F)	91	702	767	0.1102	0.5757	0.0635

Before rescue and NMS, the `reject`/`single`/`fuse`/ rescue rates are 33.6/19.6/46.4/0.3% on SAMMLV and 7.4/30.2/43.7/18.7% on CAS(ME)^3. The different rescue rates agree with the dataset-dependent diagnostic reported below.

Table 10 gives paired bootstrap differences for the same controlled comparisons. The SAMMLV rows fix the inner-selected evidence and final suppression to isolate routing; the CAS(ME)^3 native-decoder row evaluates the complete nested pipeline. Intervals that touch zero are retained rather than described as decisive.

B.3 Fixed-Reference-Profile Hyperparameter Sensitivity

The main nested protocol selects temporal pairs and rescue states, but does not independently sweep every fixed decoder constant. We therefore run a compact post-hoc sensitivity diagnostic on the paper-aligned reference profiles: $(4, 5)$ with $\gamma = .98$ and $\beta = .95$ on SAMMLV, and $(12, 15)$ with $\gamma = .98$ and $\beta = .95$ on CAS(ME)^3. For every setting, AGF is still fitted only on outer-training subjects and evaluated on all 29 SAMMLV or 94 CAS(ME)^3 held-out subjects. Only the listed quantity changes; the frozen ME-TST+ outputs, temporal profile, rescue rule, single-hypothesis floor, and all remaining settings are fixed. The middle entries in Table 11 reproduce the reference-profile results reported here.

These runs are not a second hyperparameter-selection procedure and are not used to replace the reported configuration. In particular, they do not claim that $p = .55$ or $\delta = .01$ is globally optimal. The reference profile's $\beta = .95$ is also distinct from the disabled $\beta = 0$ floor in the fully nested configuration. The table reports STRS in the displayed setting order; the archived CSV additionally records Spotting F1 and Recognition F1 for all 24 complete runs.

Candidate threshold p changes the decoded candidate pool and has the largest local effect, with opposite directions on SAMMLV and CAS(ME)^3. This confirms that a fixed reference threshold should not be interpreted as universally optimal. In contrast, the tested τ range leaves the aggregate result unchanged at these profiles, while C and final NMS IoU move STRS only slightly. The tested fusion margins yield small improvements at some reference profiles but do not give a common preference across the two datasets. Removing balanced class weights lowers STRS on both datasets, from 0.2102 to 0.2027 on SAMMLV and from 0.0598 to 0.0516 on CAS(ME)^3, supporting their use in the imbalanced three-action gate. These local observations complement, rather than replace, the subject-disjoint temporal-profile selection in the main experiments.

B.4 Complete-Cache Development Diagnostics

The following complete-cache analyses are development diagnostics, not primary configuration-selection evidence. They inspect operating points using all available frozen outputs and are therefore separated from the fully nested SAMMLV and CAS(ME)^3 comparisons in the main paper.

B.4.1 Temporal-Scale Diagnostics

At the exploratory pair $(3, 4)$ on SAMMLV with rescue disabled, RethinkFuse reduces FP from 195 to 149 relative to Fixed PairFuse and raises STRS from 0.2057 to 0.2275. At $(12, 18)$ on CAS(ME)^3 with rescue disabled, FP falls from 704 to 595 and STRS rises from 0.0570 to 0.0591. Because these pairs were identified through complete-cache exploration, the values diagnose routing at the operating points but are not used as unbiased performance estimates.

B.4.2 Rescue and Final Suppression

Rescue behaves differently across datasets. At the exploratory CAS(ME)^3 operating point, it recovers 14 TP and raises STRS from 0.0591 to 0.0642. On SAMMLV, it recovers no TP, introduces 15 FP, and lowers STRS from 0.2275 to 0.2182. This motivates selecting rescue within the training subjects rather than treating it as universally beneficial. No isolated NMS contribution is claimed from these aggregate comparisons.

Table 10: Paired subject-bootstrap STRS differences, RethinkFuse minus control.

Dataset	Control	ΔSTRS	95% interval
SAMMLV	Shared Always Fuse	$+0.0339$	$[0.0078, 0.0549]$
SAMMLV	No Fuse	$+0.0888$	$[0.0106, 0.1394]$
SAMMLV	No Reject	$+0.0380$	$[0.0093, 0.0614]$
CAS(ME)^3	Native Decoder	$+0.0147$	$[0.0053, 0.0258]$
CAS(ME)^3	Shared Always Fuse	$+0.0051$	$[0.0000, 0.0111]$
CAS(ME)^3	No Reject	$+0.0045$	$[0.0006, 0.0090]$
CAS(ME)^3	No Fuse	$+0.0042$	crosses zero

Table 11: Fixed-reference-profile outer-LOSO sensitivity. Each cell follows the setting order in the preceding column. Values are diagnostic only and are not used to select the main reported configuration.

Parameter	Settings	SAMMLV STRS	CAS(ME)^3 STRS
p	$0.45/0.55/0.65$	$0.2040/0.2102/0.2166$	$0.0626/0.0598/0.0515$
τ	$0.25/0.35/0.45$	$0.2102/0.2102/0.2102$	$0.0598/0.0598/0.0598$
δ	$0.00/0.01/0.03$	$0.2102/0.2102/0.2155$	$0.0605/0.0598/0.0605$
C	$0.1/1/10$	$0.2102/0.2102/0.2096$	$0.0588/0.0598/0.0598$
η_{nms}	$0.20/0.30/0.40$	$0.2102/0.2102/0.2096$	$0.0595/0.0598/0.0597$
Class weight	none/balanced	$0.2027/0.2102$	$0.0516/0.0598$

B.4.3 Exploratory Gate Actions

At the exploratory pairs, `reject` and `single` account for 54.7% of SAMMLV and 56.2% of CAS(ME)^3 decisions before rescue and NMS, showing that the router does not always fuse.

C Artifact Manifest

The frozen SAMMLV and CAS(ME)^3 output caches used by the paper-aligned decoder have the following SHA-256 hashes:

SAMMLV `3B2264BD527BF144DFE10E03528AC5E6`
`37FC30E10A66D8D9575648754B298569`

CAS(ME)^3 `9BDCBB3FC79A3F2A4BF6729B826B549`
`0D8A91AED913B4120C19D68CCEEC67BDA`

The implementation artifact contains the candidate decoder, the paper-aligned AGF evaluator, and the cache-based adaptation evaluator, together with the frozen CAS(ME)^3 candidate configuration and per-subject JSON outputs. Aggregate metrics are recomputed from outer-fold detections rather than averaged over workers.

For CAS(ME)^3, the archive contains 94 complete outer-fold JSON files for both Shared Always Single and independently selected Always Single. The two archives cover the same 94 subjects; summing their fold-level TP, FP, and FN reproduces $(85, 747, 773)$ and $(82, 785, 776)$, respectively. The package-level verification script also checks the aggregate reports, configuration frequencies, report transcriptions, and the SHA-256 manifest without requiring access to the original videos or frozen caches.

D Cross-Backbone Adaptation Details

Table 12 gives target-cache settings for BoostingVRME adaptation (Table 4). Target-cache selection makes these portability results exploratory, not nested or zero-shot.

Table 12: Target-cache settings for BoostingVRME adaptation.

Dataset	Pair	p	NMS	C	Single floor	Rescue
SAMMLV	$(4, 5)$	0.75	0.10	0.05	0.00	0.98
CAS(ME)^3	$(7, 9)$	0.25	0.10	5.0	0.50	0.98

Matching uses one-window peak tolerance, similarity floor 0.35, fusion margin 0.01, and balanced gate weights. Four profiles and $C \in \{0.05, 0.2, 1, 5\}$ are tested per dataset; CAS(ME)^3 further varies C, weights, fusion, matching, NMS IoU, single-window floor, and rescue. These complete-cache searches do not establish unbiased generalization.