Università degli Studi di Genova

An EST-Based Generic Event Boundary Detector

Amir Arsalan Serajoddin Mirghaed

Supervisor:
Prof. Gualtiero Volpe *(Università degli Studi di Genova)*

Co-supervisors:

Prof. Giovanna Varni *(Università degli Studi di Trento)*
Prof. Eleonora Ceccaldi *(Università degli Studi di Genova)*

July 12, 2026

Abstract

This thesis proposes a computational approach to Generic Event Boundary Detection (GEBD). GEBD involves identifying *when* meaningful events occur in an untrimmed video, without predefined action classes. The proposed approach is grounded in the Event Segmentation Theory (EST), which posits that humans detect boundaries by predicting errors when their internal model of the ongoing situation fails.

EST is operationalised through a 15-channel feature representation aligned to EST perceptual categories of time, space, objects, characters, and interactions. Such a feature representation is fed to three different machine-learning architectures: *ReconViT-PD* (RVpd), a Vision Transformer (ViT) with Nyström-based attention, a frozen-encoder GRU predictor (RV-PL), and a combination of the former so that the GRU receives the sequence of the embeddings the ViT produced. Event boundaries are detected by applying peak detection to a change-score time series. Results are compared with those obtained by an established baseline model (BL).

The approaches were tested on two datasets with distinct application scenarios: Assembly 101, a multi-view video-recording dataset featuring 53 participants assembling and disassembling 101 toy vehicles, and Breakfast, a dataset of multi-view video recordings of cooking activities. MCC with a $\pm$15-frame tolerance was adopted as the primary performance metric. On Assembly 101, results show MCCs of 0.405 (RVpd), 0.390 (EST), 0.228 (RV-PL, tying the baseline) and 0.228 (BL). For Breakfast, the results show MCCs of 0.593 (RVpd), 0.514 (EST), 0.425 (RV-PL), and 0.256 (BL). A cross-corpus experiment was also conducted by training the models on Assembly 101 and testing them on Breakfast, yielding MCCs of 0.577 (RVpd), 0.522 (EST grid), and 0.425 (RV-PL).

A deeper investigation of the results reveals that the per-frame change score the models emit *does* carry boundary information (ground-truth frames score on average 1.75 times higher than non-boundary frames), but the peak-detection readout cannot recover it because the score's noise has a heavier tail than its signal, causing the highest peaks to be false positives. The binding constraint is therefore neither the encoder nor an absence of boundary structure in the features, but rather the change-score time series, which discards information that is demonstrably present. This analysis opens new directions for future research, such as scoring boundaries on the temporal self-similarity manifold, in line with the recent convergence of the GEBD literature toward this representation.

Acknowledgements

This thesis is the product of eighteen months of work that would have been impossible without the patient guidance, technical mentorship and personal support of many people, and I am grateful to all of them.

My deepest thanks go to my supervisors. To Professor Gualtiero Volpe at the Università degli Studi di Genova, who agreed to supervise this project, kept it pointed at the right scientific question through every reframing, and granted the freedom to follow the honest negative result wherever it led. To Professor Giovanna Varni at the Università degli Studi di Trento, who asked the questions that made the empirical centre of the thesis sharper at every turn, and read every draft with the same generous attention as the first. Their willingness to revise the framing after the post-publication diagnostic re-anchored the central contribution is the reason this document carries an honest result rather than a marginal improvement. To Nadia Benini, whose previous work introduced me to the research for this thesis.

To the the Università degli Studi di Genova and to Casa Paganini - InfoMus and the Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS) for providing the institutional home, the computational resources and the academic environment in which the work was carried out.

To my family, for the support that does not show up in any commit log but is what made the long stretches of debugging, training and rewriting possible.

To my friends, for their patience during periods of disappearance and for asking the questions about the thesis that sometimes mattered more than the technical ones.

Finally, to the open-source communities behind PyTorch, PyTorch Lightning, h5py, NumPy, Matplotlib, scikit-learn and LaTeX, whose tools form the substrate on which the entire thesis was built and written.

Any errors in what follows are my own.

Contents

Abstract i

Acknowledgements ii

1 Introduction 2
- 1.1 Problem Statement . . . 2
- 1.2 Motivation . . . 3
- 1.3 Empirical contribution after the published paper . . . 3
- 1.4 Contributions . . . 5
- 1.5 Thesis Outline . . . 6

2 Background and Related Work 7
- 2.1 Event Segmentation Theory and Complementary Lenses . . . 7
  - 2.1.1 Event Segmentation Theory . . . 7
  - 2.1.2 Complementary Lenses . . . 8
- 2.2 Generic Event Boundary Detection . . . 9
  - 2.2.1 Task Definition . . . 9
  - 2.2.2 Evaluation Protocol . . . 10
  - 2.2.3 Supervised Approaches . . . 10
- 2.3 Unsupervised Approaches to Event Segmentation . . . 10
- 2.4 Technical Building Blocks . . . 11
  - 2.4.1 Transformers, Vision Transformers and Efficient Attention . . . 11
  - 2.4.2 Temporal Modelling . . . 12
- 2.5 Feature Extraction Tools . . . 12
- 2.6 Related-Work Summary and Research Gap . . . 12
- 2.7 Recent GEBD Methods (2022–2025) . . . 13
- 2.8 The Gap This Thesis Addresses . . . 14

3 Methodology 16
- 3.1 System Overview . . . 16
- 3.2 Feature Engineering . . . 16
  - 3.2.1 Channel inventory . . . 18

3.2.2 Empirical channel characterisation . . . 19
3.3 Assembly101 Segment-C Corpus . . . 19
3.3.1 Strategy C — segment selection . . . 20
3.3.2 Train/test split . . . 20
3.3.3 Verb distribution . . . 21
3.4 Verb-Level Ground Truth Pipeline . . . 21
3.5 ReconViT-PD (RVpd): Reconstruction-Based Boundary Detection . . . 22
3.5.1 Architecture . . . 22
3.5.2 Training Objective . . . 23
3.6 ReconViT-PL (RV-PL): Predictive Learning Extension . . . 24
3.6.1 Architecture . . . 24
3.6.2 Training Objective and Boundary Signal . . . 25
3.7 Baseline: Per-Video Pose MLP . . . 25
3.7.1 Architecture . . . 25
3.7.2 Boundary Detection . . . 26
3.8 The EST Model: Gated GRU Autoencoder . . . 26
3.9 Boundary Detection Post-Processing . . . 27
3.9.1 Published Recipe (RVpd) . . . 27
3.9.2 Unified MAD Detector . . . 28
3.9.3 MCC@15 as Primary Metric . . . 29
3.10 Exploratory Architectures . . . 30

4 Experimental Setup **31**
4.1 Datasets . . . 31
4.1.1 Assembly101 . . . 31
4.1.2 Breakfast . . . 32
4.1.3 Preprocessing . . . 33
4.2 Evaluation Protocol . . . 34
4.2.1 Tolerance-Based Matching . . . 34
4.2.2 Metrics . . . 34
4.2.3 Primary metric (revised): MCC@15 . . . 35
4.3 Experimental Configurations . . . 35
4.3.1 Phase 0: Published and pre-thesis configurations . . . 35
4.3.2 Phase 1: Segment-C baseline chain . . . 36
4.3.3 Phase A: inference-side improvement study . . . 36
4.3.4 Phase B: Tier 2 architectural retrains . . . 37
4.3.5 Phase C: Cross-corpus matrix . . . 37
4.4 Implementation Details . . . 38

5 Results and Discussion **39**
5.1 Historical baseline: the published configuration . . . 39
5.2 Phase 1: Segment-C raw baseline . . . 40
5.2.1 Canonical refresh: the full cross-corpus matrix . . . 41
5.2.2 Per-video and cross-dataset diagnostics . . . 43
5.3 Phase A: Inference-side improvements (negative result) . . . 43
5.4 Phase B: Tier 2 architectural retrains . . . 44
5.4.1 Per-model headroom analysis . . . 45
5.5 The diagnostic: noise floor and top-K-on-score ceiling . . . 46
5.5.1 Chance floors: what a perception-free predictor achieves . . . 46
5.5.2 Noise floor: 77–79% of every model's predictions land at no-near-GT frames . . . 48
5.5.3 Top-K-on-score recall ceiling: below 20% on the test split . . . 50
5.5.4 Training-free TSM novelty baseline: the 2-D readout must be learned 51
5.5.5 Per-verb breakdown rules out class-imbalance artefacts . . . 52
5.6 Statistical analysis and alternate framings . . . 53
5.7 Calibration, agreement and failure modes . . . 55
5.8 Summary of results and headline ranking . . . 55

6 Conclusion and Future Work **57**
6.1 Summary of Contributions (Revised) . . . 57
6.2 Empirical Findings Synthesis . . . 59
6.3 Limitations . . . 60
6.4 Future Work . . . 61

A Feature Channel Analysis **69**
A.1 Per-channel audit summary . . . 69
A.2 Model-side outcomes on the features . . . 70
A.3 Breakfast canonical retraining: training schedule and checkpoints . . . 71
A.4 Full cross-corpus metric grid . . . 72

B Code and Reproducibility **75**

Chapter 1

Introduction

1.1 Problem Statement

Conventional works addressing the identification of meaningful temporal units in video focus on action detection (detecting start and end times), segmentation (labelling every frame with an action class), or parsing (detecting temporal boundaries for sub-actions). However, these approaches are limited to predefined action classes and do not scale to generic situations — a critical limitation when developing intelligent machines capable of supporting and collaborating with human operators (Shou et al., 2021).

Humans maintain a continuous multimodal perception of the world and segment this perception into a sequence of discrete events — what Zacks and Tversky (2001) term *event structure perception*. These events represent meaningful temporal units without requiring predefined target classes. The identification of event boundaries increases the effectiveness of understanding others' actions and planning one's own (Hawkins and Blakeslee, 2004).

This capability is formalised in the literature as *Generic Event Boundary Detection* (GEBD), aimed at developing computational approaches to identify event boundaries (Shou et al., 2021). The most promising approaches are those grounded in *Event Segmentation Theory* (EST) (Zacks et al., 2007), which posits that event segmentation relies on the cognitive event models a human being shapes from ongoing situations. Such models guide the prediction of the future; when predictions fail, humans perceive a boundary marking the start of a new event.

The practical relevance of GEBD spans several application domains in which a system has to follow what a person is doing without being told in advance which actions to expect. In *manufacturing assistance and human–machine collaboration*, where this thesis ultimately sits, a robot or wearable interface that recognises when one assembly step ends and the next begins can hand the right tool, surface the right instruction, or hand control back to the operator at exactly the moment when a new event is starting. In *video summarisation and content retrieval*, boundary detection produces the temporal

segmentation on top of which thumbnail selection, chapter generation, and search indexing are built. In *surveillance and behaviour analytics*, the ability to detect that something has changed in the scene without being told what to look for is the difference between an alerting system and a recording one. In *robotics and embodied agents*, event-boundary detection is the perceptual substrate for sub-goal discovery and for hierarchical imitation learning. All of these applications share the constraint that motivates the GEBD framing: the boundary structure of the world is not known in advance, so a system that can only recognise pre-trained classes is structurally insufficient.

1.2 Motivation

EST identifies seven categories of change that drive event boundary perception (Zacks et al., 2009): *time*, *space*, *objects*, *characters*, *character interactions*, *causes*, and *goals*. No existing unsupervised GEBD system explicitly operationalises all these categories as input features; this thesis operationalises the five *observable* ones — time, space, objects, characters and interactions — that off-the-shelf detectors can supply; the two higher-level categories (causes and goals), which require explicit inference of intent and causality, are not given dedicated channels.

Furthermore, the prediction error mechanism central to EST maps naturally to self-supervised learning: a model trained to reconstruct or predict "normal" feature patterns will produce higher error at event boundaries where the underlying activity changes. This motivates a computational approach to GEBD that implements the EST categories and leverages reconstruction error as a boundary detection signal.

Existing approaches either rely on supervised learning with expensive frame-level annotations (limiting scalability), or use simple hand-crafted features that fail to capture the rich multimodal cues humans use. This thesis proposes a fully unsupervised approach grounded in cognitive theory, tested in a manufacturing-like assembly scenario as a step towards human–machine collaboration.

1.3 Empirical contribution after the published paper

The initial version of this work was presented at CHItaly 2025 (Mirghaed et al., 2025). The published paper introduces a single model — the reconstruction-and-peak-detection ViT (RVpd) — and reports MCC$=$ 0.55 on a 62-video subset of Assembly101 with coarse-grained ground truth at $\pm$15-frame tolerance. Between publication and this thesis, an extended experimental programme was carried out as a development exercise: two further models (RV-PL, a temporal predictive extension; BL, a per-video online pose-only baseline reproducing Basgol et al. (2024)) were added, and the evaluation was widened to both Assembly101 and Breakfast at fine and coarse granularity with $\tau \in \{5, 10, 15\}$,

together with cross-dataset transfer and 9-fold cross-validation. These extended experiments inform the development context but are not part of the published paper and are documented for completeness in Section 5.1.

The thesis itself moves the primary evaluation onto a unified Assembly101 segment-C corpus that compares every paper-relevant model under identical conditions and against a single shared peak detector, designed to remove detector-tuning as a confounder of cross-model comparison. The published RVpd headline does not survive this stricter protocol, and the empirical centre of the thesis is therefore shifted: the contribution is no longer "RVpd reaches MCC = 0.55", but a diagnostic explanation of why no model in the family reaches that figure on the new benchmark, together with a concrete architectural proposal designed to break the diagnosed ceiling.

The unified corpus is Assembly101 segment-C: 250 videos sampled at 5 fps, one 600-frame (two-minute) window per video guaranteed to contain at least one ground-truth boundary, verb-level ground truth with nine verbs collapsed from 199 verb–noun pairs (*attach*, *detach*, *screw*, *unscrew*, *inspect*, *attempt*, *demonstrate*, *position*, *remove*), and a person-disjoint 80/20 split (193 training videos / 57 test videos). Every model is scored with the same Median Absolute Deviation (MAD) peak detector (window = 15, $k_{\text{MAD}} = 3.0$) on its per-frame output, and the primary metric is MCC at a ±15-frame tolerance (MCC@15). On this benchmark, the four paper-relevant models land at MCC@15 of 0.405 (RVpd), 0.390 (EST grid=4), 0.228 (RV-PL k=3, level with the baseline) and 0.228 for the pose-only Baseline (BL). The two reconstruction-based models cluster at the top, only 0.015 MCC@15 apart, while the predictive RV-PL head ties the pose-only baseline. This thesis additionally contributes (a) a comprehensive inference-side improvement study (13 configurations, zero wins over the single-model baselines), demonstrating that score-side post-processing cannot improve performance on this corpus; (b) Tier-2 architectural retrains addressing concrete encoder bottlenecks: the EST grid= 4 spatial-pool fix carries into the v2.1 headline (MCC@15 = 0.390), while the RV-PL k=3 horizon shift recovered +0.012 on the original extraction but does not survive the cleaner v2.1 channels, where the predictor collapses to 0.228 (Section 5.4); (c) the **central negative diagnostic finding** that 77–79 % of every encoder-based model's predicted boundaries fire at frames with no ground-truth boundary in any ±15-frame window — a noise floor of the unsupervised peak-detection paradigm itself — together with a **top-K-on-score ceiling of 18.8 % recall** that no detector observing only the per-frame score can exceed — a ceiling at the level of blind placement at the same budget, with perception-free random chance floors (uniform D1 and train-histogram D2) confirming that the encoders sit clearly above chance at every tolerance (MCC@15 floors of 0.18–0.19 calibrated and ≈ 0.28 encoder-matched, against the encoders' 0.390–0.405); and (d) a **full cross-corpus matrix** (A→A, A→B, B→A, B→B over Assembly segment-C and Breakfast Coarse) on the EST features, demonstrating symmetric encoder transfer for EST and RV-PD and a predictor-cadence

asymmetry for RV-PL. The diagnosis points clearly at the convergent direction of the recent GEBD literature (UBoCo (Kang et al., 2022), FlowGEBD (Zheng et al., 2024), DDM-Net (Tang et al., 2022), EfficientGEBD (Lin et al., 2024)): scoring boundaries on the two-dimensional temporal self-similarity manifold rather than on a per-frame scalar. Translating that direction into a concrete model on the segment-C corpus is left as future work.

1.4 Contributions

This thesis makes the following contributions. The central contribution of this revised version is the **honest negative result and accompanying diagnostic**, not an improvement of the headline numbers reported in the CHItaly paper.

1. **EST-aligned feature representation.** *The EST features*: a 15-channel feature set that explicitly operationalises the EST perceptual categories (time, space, objects, characters, interactions) using off-the-shelf detectors, audited per channel in Appendix B and used for every result in this thesis.

2. **ReconViT-PD (RVpd):** A Vision Transformer with Nyström-based attention (Xiong et al., 2021) trained via self-supervised reconstruction on the EST features, where the reconstruction error signal serves as a computational analogue of EST's prediction error. Boundary detection is performed through signal fusion and peak detection. The CHItaly paper (Mirghaed et al., 2025) reports MCC = 0.55 on the 62-video Assembly101 coarse subset; on the segment-C verb corpus with the shared MAD detector, that number is superseded by MCC@15 = 0.405, which remains the strongest single model in the family on the new benchmark.

3. **ReconViT-PL (RV-PL):** A temporal extension developed after the publication of Mirghaed et al. (2025) that freezes the trained ViT encoder and attaches a causal GRU-based predictor, explicitly modelling temporal context through a 64-frame sliding window. On the segment-C verb corpus RV-PL k=3 reaches MCC@15 = 0.228, level with the pose-only baseline: its predictive head is fragile to the sharpness of its input features and the prediction-error signal flattens on the cleaner v2.1 channels (Section 5.4).

4. **Extended pre-thesis experimental programme (Section 5.1).** Between the CHItaly publication and this thesis, a development-oriented extended programme was carried out: the BL pose-only baseline and RV-PL were added as comparators, evaluation was widened to both Assembly101 and Breakfast at fine and coarse granularity with $\tau \in \{5, 10, 15\}$, and 9-fold cross-validation was run for statistical

significance. These results are not part of the published paper but inform the rest of the thesis as historical baselines.

5. **Post-publication diagnostic.** A unified re-evaluation on Assembly101 segment-C shows that the per-frame score carries genuine boundary information — ground-truth frames score on average 1.75× higher than non-boundary frames, and at every tolerance the encoders sit clearly above the perception-free random chance floors (D1 uniform, D2 train-histogram) — yet the unsupervised peak-detection readout cannot exploit it as fully as it should: 77–79 % of predictions still fall at frames with no ground-truth boundary, and the top-K-on-score recall ceiling caps at 18.8 % because the score's noise has a heavier tail than its signal. Together with the cross-corpus matrix the diagnostic identifies the per-frame-scalar readout — not the encoder, and not an absence of signal — as the binding constraint on the unsupervised peak-detection paradigm; the forward direction this opens (a 2-D readout over the temporal self-similarity matrix, following the convergent pattern of the recent GEBD literature) is set out in Chapter 6 as future work.

6. **A published conference paper** at CHItaly 2025 (Mirghaed et al., 2025), presenting the core approach and initial results.

1.5 Thesis Outline

The remainder is organised as follows. Chapter 2 reviews the background — Event Segmentation Theory, the technical building blocks, and the recent GEBD literature converging on the temporal self-similarity manifold (Kang et al., 2022; Zheng et al., 2024) — and states the research gap. Chapter 3 presents the EST feature pipeline and the four models (RVpd, RV-PL, the EST GRU autoencoder, and the BL baseline). Chapter 4 describes the datasets, the unified segment-C corpus, the shared MAD detector, and the evaluation protocol. Chapter 5 reports the results and the central diagnostic. Chapter 6 draws conclusions, states limitations, and sets out the forward direction.

Chapter 2

Background and Related Work

This chapter reviews the theoretical and technical foundations of this thesis. Section 2.1 introduces Event Segmentation Theory and the prediction-error mechanism that motivates our approach, alongside the representation-learning, signal-processing and neuroscience perspectives that frame the design. Section 2.2 formally defines the Generic Event Boundary Detection task and Section 2.3 surveys unsupervised approaches. Section 2.4 and Section 2.5 cover the technical building blocks and feature extraction tools. Section 2.6 positions this work, while Section 2.7–Section 2.8 survey the recent GEBD literature and state the gap this thesis addresses.

2.1 Event Segmentation Theory and Complementary Lenses

2.1.1 Event Segmentation Theory

Event Segmentation Theory (EST), developed over two decades of cognitive science research (Zacks and Swallow, 2007; Zacks and Tversky, 2001; Zacks et al., 2007, 2009), frames how humans parse continuous sensory experience into discrete, meaningful events. Its central insight is that segmentation is not a post-hoc deliberate process but an automatic, ongoing perceptual mechanism: the brain continuously generates predictions about upcoming sensory input, and event boundaries are registered when these predictions fail. Zacks and Tversky (2001) showed that humans spontaneously segment activity into a hierarchically organised structure at multiple temporal scales, and that this segmentation is remarkably consistent across observers. Zacks et al. (2007) formalised the prediction-error mechanism: the brain maintains an internal *event model* of the current situation that generates expectations; when sensory input deviates sufficiently from these expectations, the system registers a boundary and initiates an *event model update*.

Zacks et al. (2007) further identified seven perceptual dimensions driving segmentation: temporal, spatial, object, character and interaction features (directly observable from visual input), plus causal and intentional features (requiring higher-level inference). This distinction guides our design: we operationalise the five observable categories through explicit feature channels (Section 3.2), while causal and intentional features emerge implicitly from temporal patterns in the learned representations. Prior work from the same group operationalised these EST change categories *semi-automatically* — manual change annotations combined with change-point detection on behavioural time-series (Ceccaldi, 2021); this thesis operationalises them *fully automatically* as learned feature channels. EST also predicts that events are perceived in a nested hierarchy — "picking up a screw" within "assembling the base plate" within "toy assembly" — with boundaries occurring at multiple temporal scales simultaneously; accordingly, our experiments use both fine- and coarse-grained ground truth (Section 4.1.1) and a tolerance-based evaluation ($\tau \in \{5, 10, 15\}$ frames).

The prediction-error framework motivates three design commitments: *self-supervised learning* (a system trained to predict or reconstruct ongoing input should exhibit elevated error at boundaries, without labelled boundary data — the core principle of our approach); *context dependence* (detection depends on accumulated context, motivating RV-PL's causal GRU predictor over a sliding window); and *multi-modal features* (combining motion, pose, objects and interactions rather than a single modality). A key limitation revealed by this lens is that our system is a *flat predictor* lacking a true architectural hierarchy: RVpd processes each frame independently and even RV-PL's GRU window captures a single temporal scale, motivating the multi-scale architecture identified as future work (Section 6.4).

2.1.2 Complementary Lenses

Three further perspectives complete the picture. From a *representation-learning* view, a boundary corresponds to a large distance between feature embeddings of consecutive frames, but only if the encoder maps same-event frames to a cluster and boundaries to large inter-embedding distances; Mounir et al. (2022) showed learned representations outperform raw optical flow. We address this at two levels: an engineered representation (the EST features) and a learned one (the ViT encoder, trained by self-supervised reconstruction). Self-supervised learning (SSL) learns representations from unlabelled data via pretext tasks; the most relevant image-level paradigms are SimSiam (Chen and He, 2021), DINO (Caron et al., 2021) and the reconstruction-based MAE (He et al., 2022). Our system uses **reconstruction error as the boundary signal**: contrastive and self-distillation objectives (SimSiam, DINO) are in principle stronger at separating semantically distinct inputs, but they are collapse-prone without large batches or carefully

tuned momentum encoders, and our own preliminary SimSiam- and DINO-style trials collapsed early under the small batches (batch size 8) imposed by single-GPU memory — well below the scale they need to avoid the trivial constant-embedding solution (the low effective dimensionality of the EST features, Appendix B, may have compounded this). Reconstruction is collapse-free by construction, which is why the paper-scoped models are built on it; the forward direction (Section 6.4) pairs the proposed 2-D self-similarity readout with a collapse-resistant geometry-shaping objective (VICReg-style) rather than naive DINO distillation.

From a *signal-processing* view, the model's output is a noisy time-series whose change-points must be recovered, requiring per-video Z-score normalisation (critical in our system: F1 ≈ 0.39 vs F1 ≈ 0.28 for concatenated evaluation, Section 3.9.1), optional Gaussian smoothing ($\sigma = 3$ for RV-PL), and peak detection via `scipy.signal.find_peaks`. Our published work added **signal fusion** of reconstruction error and temporal feature difference (Section 3.9.1), best at a 90/10 weighting. Finally, from a *neuroscience* view, Hawkins and Blakeslee (2004) casts the neocortex as a hierarchical predictive-coding machine in which predictions descend and prediction errors ascend across cortical levels at different timescales — the biological mechanism behind EST, and a reminder that our purely feedforward architecture lacks the multi-layered hierarchy predictive coding suggests is essential.

2.2 Generic Event Boundary Detection

2.2.1 Task Definition

Generic Event Boundary Detection (GEBD) was formalised as a benchmark task by Shou et al. (2021), who introduced the Kinetics-GEBD and TAPOS datasets with human-annotated event boundaries. Given an untrimmed video $\mathbf{V} = \{v_1, v_2, \ldots, v_T\}$, the goal of GEBD is to produce a set of predicted boundary timestamps $\hat{\mathcal{B}} = \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_K\}$ that align with human-perceived event boundaries.

GEBD differs from related video understanding tasks in several ways. Temporal Action Segmentation (Abu Farha and Gall, 2019; Yi et al., 2021) assigns a category label to every frame, requiring a predefined action vocabulary, whereas GEBD only identifies *where* changes occur — making it more general and applicable to open-world scenarios. Temporal action detection localises instances of specific action classes, while GEBD detects boundaries between *any* types of events; video anomaly detection flags deviations from normal behaviour, while GEBD detects all event transitions.

2.2.2 Evaluation Protocol

The standard GEBD evaluation (Shou et al., 2021) uses **tolerance-based matching**: a predicted boundary $\hat{t}$ is a true positive if $|\hat{t} - t^*| \leq \tau$ for some ground truth boundary t^*, where τ is a tolerance window. Results are typically reported using the F1 score, which is the harmonic mean of precision and recall. The Relative Distance metric, originally proposed by Shou et al. (2021), normalises tolerance by video length. In this thesis, we use absolute frame tolerance ($\tau \in \{5, 10, 15\}$ at 5 FPS) and additionally report the Matthews Correlation Coefficient (MCC), which is more appropriate for imbalanced datasets where boundaries constitute only $\sim$1.2% of frames.

2.2.3 Supervised Approaches

The majority of GEBD methods are supervised, using pre-trained video backbones with temporal classification heads but requiring large-scale annotations (Kinetics-GEBD contains $\sim$54K videos with $\sim$600K boundaries), which are expensive and inherently subjective. Our approach, in contrast, requires no boundary annotations at all.

2.3 Unsupervised Approaches to Event Segmentation

Several works have explored unsupervised or self-supervised event boundary detection, each offering a different computational interpretation of prediction error. Basgol et al. (2024) proposed a self-supervised model directly inspired by EST in which a recurrent network predicts upcoming sensory features and prediction error serves as the boundary signal, validated through psychological experiments; our baseline model (Section 3.7) reproduces their core principle of per-video online training on pose features with surprise-based detection. Aakur and Sarkar (2019) used a recurrent architecture with a dual-memory (short- and long-term) mechanism over pre-extracted CNN features, detecting boundaries when accumulated prediction error exceeds a threshold, though the fixed CNN backbone limits features to generic visual semantics rather than EST-specific categories. Mounir et al. (2022) combined LSTM-based sequence modelling with temporal forecasting and showed that combining different feature representations (optical flow, objects, pose) improves detection — directly motivating our multi-channel EST-aligned features.

Table 2.1 summarises the key differences between our approach and these related unsupervised methods.

Our approach differs from prior work in three key ways: (1) we explicitly design features to map to EST's perceptual categories rather than using generic visual features; (2) we employ a Vision Transformer with Nyström attention, which can model spatial

Table 2.1: Comparison with related unsupervised event boundary detection methods.

	Basgol	**Aakur**	**Mounir**	**Ours**
EST-motivated	✓	Partial	✓	✓
Multi-channel features	—	CNN	LSTM + CNN	15ch EST
Architecture	MLP	RNN	LSTM	ViT + GRU
Attention mechanism	—	—	—	Nyström
Temporal modelling	Per-frame	Dual memory	LSTM	GRU window
Post-processing	Threshold	Threshold	—	Peak detection
Self-supervised	✓	✓	✓	✓

relationships across the feature map more effectively than recurrent architectures; (3) we introduce a two-stage pipeline where a frozen self-supervised encoder provides features for a temporal predictor, combining the strengths of both reconstruction and prediction paradigms.

2.4 Technical Building Blocks

2.4.1 Transformers, Vision Transformers and Efficient Attention

The Transformer (Vaswani et al., 2017) is built on $O(N^2)$ self-attention, and the Vision Transformer (ViT) (Dosovitskiy et al., 2021) adapts it to images by tokenising non-overlapping patches; our implementation replaces the linear patch projection with a single convolutional layer over the multi-channel feature map, a simplified form of the convolutional tokenisation introduced by CvT (Wu et al., 2021). Because standard attention is $O(N^2)$ in sequence length, this work explores four mechanisms, summarised in Table 2.2.

Table 2.2: Comparison of attention mechanisms. N = sequence length, m = landmarks, k = projection dimension.

Mechanism	**Complexity**	**Approximation**	**F1 (our data)**
Standard	$O(N^2)$	Exact	0.4310
Nyströmformer	$O(Nm)$	Matrix	**0.4365**
Performer	$O(N)$	Kernel	0.4323
Linformer	$O(Nk)$	Projection	0.4033

The thesis models use the Nyströmformer (Xiong et al., 2021), which approximates the full attention matrix via $m \ll N$ segment-mean landmark points ($O(Nm)$) — well-suited to our spatial patch tokens. The kernel-based Performer (Choromanski et al., 2021) and projection-based Linformer (Wang et al., 2020) were evaluated in ablations (Table 2.2) but are not used by the paper-scoped models.

2.4.2 Temporal Modelling

In our RV-PL model, a 2-layer Gated Recurrent Unit (GRU) (Cho et al., 2014) processes sequences of frozen ViT embeddings to predict an embedding k frames ahead (the reported model uses k=3). Alternative temporal backbones — MS-TCN (Abu Farha and Gall, 2019), its Transformer hybrid ASFormer (Yi et al., 2021), and the selective state space model Mamba (Gu and Dao, 2024) (the latter in exploratory experiments, Section 3.10) — are not used by the paper-scoped models.

2.5 Feature Extraction Tools

The system builds multi-channel feature images from off-the-shelf detectors: the YOLO framework (Jocher et al., 2022) for person detection, 17-keypoint pose estimation and object masks (earlier iterations used RTMDet (Lyu et al., 2022) and the 133-keypoint RTMPose (Jiang et al., 2023) before consolidating on YOLO), and RAFT (Teed and Deng, 2020) for dense optical flow (channels 1–2). Where the earlier recipe's classical Farnebäck flow (Farnebäck, 2003) was effectively degenerate (near-zero variance, $> 99\%$ of pixels zero), RAFT recovers a live per-pixel flow field, solving the "dead channel" problem; App B confirms optical flow x is the strongest temporal-difference discriminator of the fifteen channels.

2.6 Related-Work Summary and Research Gap

In summary, the four lenses converge: EST supplies the theoretical foundation (boundaries are moments of prediction error), representation learning supplies the embedding space in which boundaries manifest as detectable distances, signal processing supplies the machinery to extract boundaries from a noisy output signal, and neuroscience supplies the hierarchical predictive-coding motivation for the architecture.

The research gap identified at the intersection of these lenses is *twofold*, and the two parts play different roles in this thesis. *Within the per-frame-score paradigm* that all existing unsupervised approaches share, prior work either (1) uses generic visual features (CNN backbones) that do not explicitly encode EST's perceptual categories, or (2) uses simple architectures (MLPs, shallow RNNs) that cannot model the rich spatial structure of multi-channel feature representations, or (3) lacks the signal-processing care to extract clean boundary signals from noisy model output. No prior work combines EST-explicit multi-channel features with Vision Transformer architectures and self-supervised reconstruction/prediction under a unified detection protocol; this thesis fills that within-paradigm gap. *At the paradigm level*, however, the recent GEBD literature surveyed in Section 2.7 has moved away from per-frame scalar scoring altogether, and the empirical

diagnostic of Chapter 5 will show that the scalar abstraction itself is the binding constraint on this corpus. The thesis does not close that second gap — it *diagnoses* it, and Section 2.8 states it precisely as the bridge to the forward direction of Chapter 6.

2.7 Recent GEBD Methods (2022–2025)

The survey in Section 2.3 and Section 2.6 covers the cognitive and early self-supervised antecedents of our model family. Since the Kinetics-GEBD benchmark of Shou et al. (2021), however, a fast-moving GEBD literature has emerged whose top methods share architectural commitments that diverge from the per-frame scalar paradigm of our BL, EST, RV-PD and RV-PL models.

- **UBoCo** (Kang et al., 2022): the first competitive *unsupervised* GEBD detector, using a Boundary Contrastive loss to reshape a frozen-backbone temporal self-similarity matrix (TSM) into block-diagonal structure and recovering boundaries as corners, beating the contemporary supervised state of the art without a single label.
- **DDM-Net** (Tang et al., 2022): introduced the *Dense Difference Map*, a $T \times T$ matrix of multi-order frame differences fused by progressive attention, establishing that a manifold of pairwise relations beats per-frame scoring even with full supervision.
- **SC-Transformer++** (Li et al., 2022): won the 2022 LOVEU GEBD Challenge at F1@0.05 = 86.49 on Kinetics-GEBD, operating on the TSM with structured spatial-context priors.
- **EfficientGEBD** (Lin et al., 2024): holds the single-model state of the art on Kinetics-GEBD at F1@0.05 = 82.9 (2.8× faster than the prior SOTA), using a *DiffMixer* block over a multi-radius difference manifold.
- **DyBDet** (Wang et al., 2024): reached F1@0.05 = 79.6 with a *dynamic* architecture routing each snippet through detector subnets matched to local difficulty, plus a multi-order difference detector.
- **DiffGEBD** (Wang et al., 2025): reframes detection as generation, a denoising diffusion model refining a noisy TSM towards the clean block-diagonal target, reaching F1@0.05 = 78.4 on Kinetics-GEBD.
- **FlowGEBD** (Zheng et al., 2024): the strongest *non-parametric* detector, a training-free pipeline reading boundaries from discontinuities in an optical-flow-normalised similarity signal, attaining F1@0.05 = 71.3 with zero training.

These winners converge on four design choices the per-frame paradigm of our family does not implement: (1) scoring on a 2-D $T \times T$ manifold rather than per-frame scalars, where each boundary is a corner that survives noise a 1-D scalar would not; (2) multi-order and multi-scale difference operators; (3) generative or iterative refinement instead of single-shot thresholding; and (4) a frozen strong backbone with a lightweight temporal head. These commitments form a design language against which the closing sections situate our own models.

Self-supervised representations relevant to GEBD. Because boundary corners only emerge on the similarity manifold when per-frame embeddings are sufficiently discriminative, the choice of self-supervised pretext task matters as much as the temporal head. Relevant objectives group into geometry-shaping (VICReg (Bardes et al., 2022), the Boundary Contrastive loss of Kang et al. (2022), ShotCoL (Chen et al., 2021)) and temporal-prediction / alignment (TS2Vec (Yue et al., 2022), TS-TCC (Eldele et al., 2021), V-JEPA (Bardes et al., 2024), Temporal DINO (Sayed et al., 2023), and the cycle-consistency of TCC (Dwibedi et al., 2019), directly relevant to procedural datasets such as Assembly101 and Breakfast). Together they sketch a representation toolbox — collapse-resistant regularisation, contrastive geometry shaping, embedding-space prediction, cycle-consistency — compatible with the TSM/DDM scoring paradigm that defines the current GEBD state of the art.

2.8 The Gap This Thesis Addresses

Stepping back, the literature reviewed in Section 2.7 and the diagnostic findings of Chapter 5 jointly identify a single, sharp gap that this thesis confronts.

First, all four models developed within the paper-scoped scope of this thesis — BL, EST, RV-PD and RV-PL — produce a per-frame scalar boundary signal (per-video MLP MSE, gated multiplicative error, reconstruction error, prediction error) and apply a one-dimensional peak detector (MAD-based or otherwise) to that signal. The temporal modelling is sophisticated; the *scoring abstraction is one-dimensional.*

Second, the recent GEBD literature has converged on a different abstraction. Without exception, the methods that have topped the Kinetics-GEBD benchmark since 2022 — UBoCo, DDM-Net, SC-Transformer++, EfficientGEBD, DyBDet, DiffGEBD, FlowGEBD — score boundaries on a $T \times T$ similarity or dense-difference manifold, where each boundary is a 2-D corner rather than a 1-D peak. The transition from 1-D to 2-D scoring is the single most consistent architectural commitment across this body of work.

Third, the empirical diagnosis presented in Chapter 5 finds that the per-frame scalar abstraction is the binding constraint of our pipeline as measured against the segment-C ground truth. Across all four paper-scoped models, 77–79% of predictions fire at frames

with no GT boundary within ±15 frames, and the top-K-on-score recall ceiling caps at 18.8%. A subsequent triangulation against multi-model agreement (Section 6.3) shows that approximately 10 percentage points of the no-near-GT rate is attributable to GT incompleteness rather than model failure, lowering the true model FP rate to roughly 65–70% but leaving the qualitative conclusion intact: at the magnitude at which the per-frame score is surfaced by these encoders, the peak detector cannot consistently distinguish a real boundary from a non-boundary spike. The per-frame-score abstraction is therefore a binding constraint of the pipeline, even if it is not the sole one.

Fourth, Chapter 6 reads the literature consensus and the empirical diagnosis together as a forward direction: the natural next step for a system built on EST-aligned multi-channel features and a Nyströmformer encoder is to replace the per-frame scalar readout with a 2-D readout over a temporal self-similarity matrix, shaped by a contrastive geometry-shaping objective rather than by reconstruction. Translating that direction into a concrete model and running it against the segment-C baselines of this thesis is left as future work.

Chapter 3

Methodology

This chapter presents the proposed methodology: a system overview (Section 3.1); the feature-engineering pipeline (Section 3.2); the Assembly101 segment-C corpus that is the unified evaluation substrate (Section 3.3) and the verb-level ground-truth pipeline underpinning it (Section 3.4); the core ReconViT-PD model (Section 3.5), its predictive-learning extension (Section 3.6) and the baseline reproduction (Section 3.7); the shared boundary-detection post-processing including the unified MAD detector and the move to MCC@15 (Section 3.9); and the exploratory architectures (Section 3.10).

3.1 System Overview

The proposed system follows a four-stage pipeline: (1) **feature extraction** (offline) — raw frames are processed at 5 FPS by off-the-shelf detectors into a multi-channel feature image per frame, stored in HDF5; (2) **self-supervised training** — a model learns "normal" temporal patterns by reconstruction (RVpd) or prediction (RV-PL); (3) **signal generation** (inference) — the trained model produces an error signal on unseen test data (reconstruction error, temporal difference or prediction error); (4) **boundary detection** (post-processing) — signals are normalised, optionally fused, and peak-detected to identify action boundaries.

3.2 Feature Engineering

The feature representation is a central contribution of this work. Guided by EST's categories of change (Section 2.1.1), we designed a multi-channel feature image that explicitly captures each dimension of event change. Throughout this thesis we call it simply *the EST features*: a 15-channel tensor of shape $(15, 135, 240)$ per frame, stacked over time into the $(T, 15, 135, 240)$ input to every paper-scoped encoder in Chapter 5. The published version (Mirghaed et al., 2025) used an earlier 16-channel variant; the thesis re-evaluation

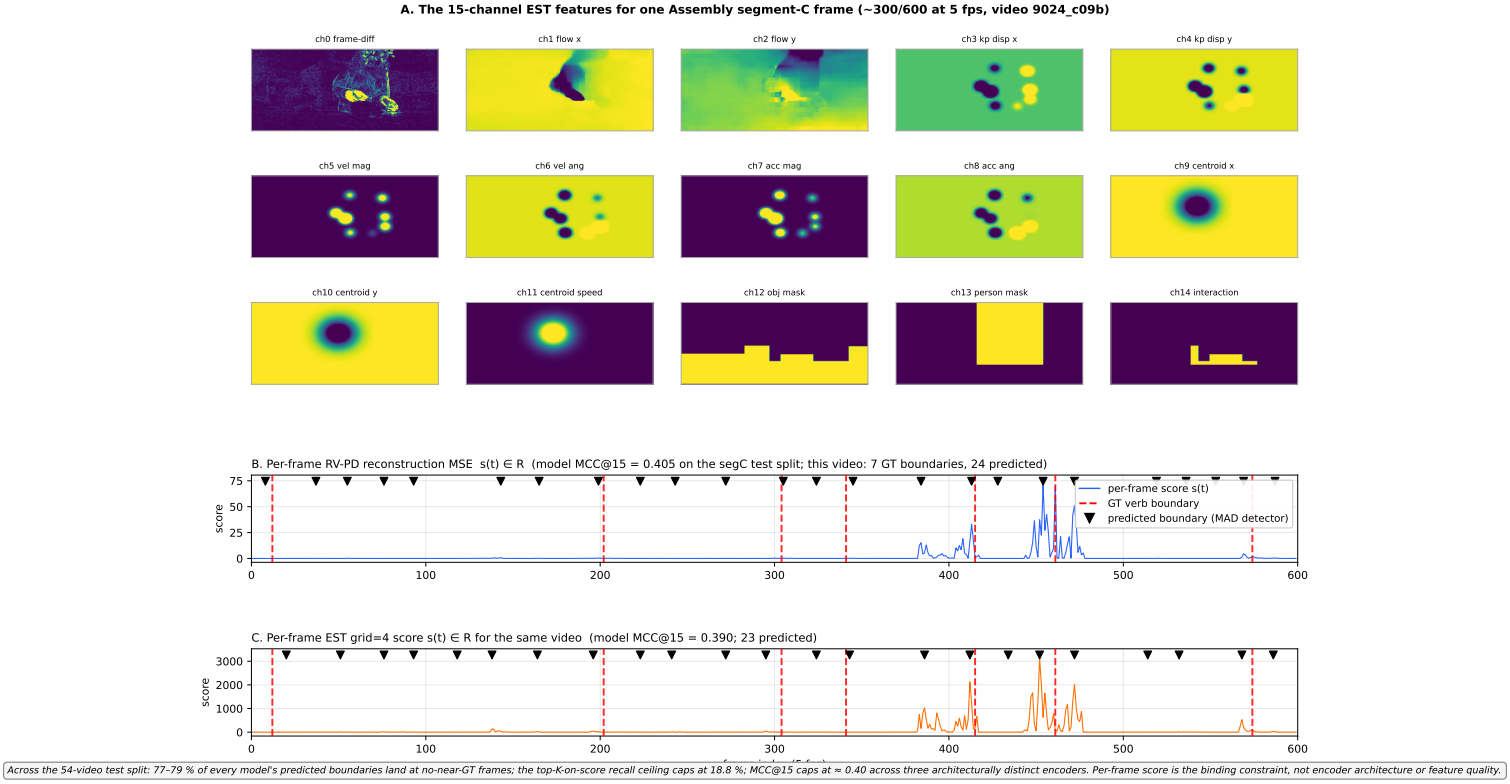

Figure 3.1: End-to-end system pipeline on one Assembly segment-C test video (`9024_c09b`, 600 frames at 5 fps, 7 GT verb boundaries). The figure starts from the derived EST features, not a raw frame (Assembly101 frames are not redistributable). **A:** the 15 EST feature channels for one frame (Section 3.2.1); ch0, the dense RAFT optical-flow fields (ch1–2), the keypoint splats (ch3–8), the centroid heatmaps (ch9–11) and the presence/interaction masks (ch12–14) all carry clear spatial structure, as the audit of Appendix B reports. **B, C:** the per-frame score $s(t)$ from the two strongest encoders (RV-PD reconstruction MSE in B, EST grid= 4 in C), with the seven GT boundaries (red dashed) and the predictions of the shared MAD detector ($W = 15$, $k_{\text{mad}} = 3$, black triangles). Both produce far more boundaries (24 and 23) than the segment has GT (7) — the headline diagnostic of Chapter 5: across all 54 test videos the binding constraint is the per-frame scalar-score abstraction itself, not the encoder or the feature input.

uses the 15-channel refinement described below, audited per channel in Appendix B.

3.2.1 Channel inventory

The 15 channels map onto EST's perceptual categories of change (time, space, characters, objects, interactions), as documented in Table 3.1; each produces a 135×240 image at the 5 fps feature rate.

Table 3.1: The EST features: 15 channels across EST perceptual categories of change.

EST Category	Channel	Count	Method
Time	Frame difference energy	1	Localised absolute difference
Space	Optical flow x, y	2	RAFT-Large dense flow
Characters	Keypoint displacement x, y	2	8 keypoints, Gaussian-RBF splat
Time/Space	Keypoint velocity (mag, angle)	2	Savitzky-Golay 1st derivative
Time/Space	Keypoint acceleration (mag, angle)	2	Savitzky-Golay 2nd derivative
Time/Char.	Centroid displacement x, y, speed	3	Gaussian heatmap modulated by displacement
Objects	Object presence mask	1	YOLO bounding boxes
Characters	Person presence mask	1	YOLO bounding boxes
Interaction	Person–object interaction mask	1	Intersection of person and object masks

Key implementation choices are as follows. Keypoint velocity and acceleration are the first and second derivatives of per-keypoint trajectories smoothed with a 7-frame Savitzky-Golay filter (window = 7, order = 2). Of the 17 COCO keypoints from the YOLOv8-pose detector, only the 8 reliably detected upper-body joints are retained (left/right shoulder, elbow, wrist and hip). Each keypoint is splatted with a Gaussian-RBF kernel ($\sigma \approx 3$–5 px) rather than a single pixel, which would otherwise leave sparsity ≥ 0.9999 and be lost under the encoder's spatial-mean pooling. The optical-flow channels (1–2) use RAFT-Large (Teed and Deng, 2020) dense flow, each component divided by a fixed 20 px scale (no clipping), replacing an earlier Farneback-with-clipping implementation that was numerically degenerate (mean $\sim 1.7 \times 10^{-10}$, sparsity ~ 0.996). Centroid channels (9–11) render per-person displacement as a Gaussian heatmap modulated by displacement magnitude, and the presence channels (12–14) are binary bounding-box maps, the interaction mask being the intersection of the person and object masks.

An exploratory extended set. A development pipeline of up to 35 channels (SAM segmentation masks, distance transforms, RAFT divergence/curl, ResNet-18 content features and scalar object/interaction/acceleration channels) was built but used for none of the thesis numbers; it is retained as a future-work code branch only.

3.2.2 Empirical channel characterisation

Table 3.1 describes the *nominal* content of each channel; a quantitative characterisation on the actual segment-C corpus — statistical profiles, cross-channel correlations and per-channel boundary discriminability — is presented in full in Appendix A. Three findings from that appendix recur in the remainder of the thesis: (i) the 15 nominal channels carry fewer effective degrees of freedom than their count suggests (Appendix A.2); (ii) the maximum per-channel boundary discriminability bounds what any unsupervised head can recover (Appendix A.1); and (iii) the channel density mix is mismatched to the pixel-wise L2 objectives used by EST and RV-PD (Appendix A.1, Appendix A.2).

3.3 Assembly101 Segment-C Corpus

The evaluation substrate differs from the published paper (Mirghaed et al., 2025) in two respects: the action vocabulary is collapsed from the full (verb, noun) inventory to nine verb-only classes, and each video is replaced by a single fixed-length 600-frame window (a *segment*). The combined corpus is denoted *segment-C* ("C" for the window-selection strategy of Section 3.3.1) and serves as the unified Assembly set against which BL, EST, RV-PD and RV-PL are compared on equal footing.

The ChItaly version of this work (Mirghaed et al., 2025) evaluated RVpd on 62 frontal-view Assembly videos with the coarse ground truth (1 261 boundaries). For the thesis, the corpus is redesigned for (i) direct comparability to Breakfast in scale and boundary density, and (ii) annotation at a single action granularity, so cross-model comparisons are not confounded by boundary spacing. Two design constraints follow. First, a **verb-only vocabulary**: the 199 distinct (verb, noun) labels are collapsed to nine verbs (`attach`, `detach`, `screw`, `unscrew`, `inspect`, `attempt`, `demonstrate`, `position`, `remove`) and adjacent same-verb segments are merged into one super-segment; without this merge every noun transition would yield a boundary, fragmenting each video into many more micro-events than the ∼eight verb transitions it now carries (the noun-level inventory spans 199 distinct (verb, noun) labels). Second, **frame-budget parity**: the target of 120 k train / 30 k test frames at 5 FPS (Breakfast's 80 k/20 k inflated ∼50 % for Assembly's higher within-video variance) is realised as 112 957 train frames over 193 videos and 33 605 test frames over 57 videos.[1]

[1] The 80 k/20 k figure is the design-time Breakfast subset that set the segment-C frame budget; the content-disjoint V2.2 Breakfast rebuild used for the cross-corpus matrix of Chapter 5 is larger (∼145 k train / 39 k test frames, Section 4.1.2).

3.3.1 Strategy C — segment selection

For each of the 250 usable videos in the 251-session manifest, Strategy C draws a single 600-frame window. The window is constrained so that, where possible, it contains at least one verb boundary; the placement rule is

> *If the video has* ≥ 1 *ground-truth boundary and is longer than 600 frames:* pick a boundary b uniformly at random and place the window such that start $\in [\max(0, b-599),\ \min(b, T-600)]$, so b may land anywhere inside the window. *If the video is shorter than 600 frames:* use the whole video. *If the video has no boundaries at all:* draw a random 600-frame window and keep it as an action-less negative.

All random draws (boundary choice and window placement) use a fixed seed (42), recorded with the strategy tag in the split manifest, so the corpus is reconstructible bit-for-bit. The rule produces exactly one segment per video over the 250 source videos. Twelve of these become action-less negatives because their only boundaries fell outside the drawn window, leaving 238 segments with at least one boundary; the single already-action-less source session is the 251st, dropped from the manifest to give the 250 usable videos.

Ground-truth boundaries are then shifted into segment-relative coordinates ($e \mapsto e -$ seg.start when inside the window, dropped otherwise). This silent drop removes 53 % of full-video boundaries (2 023 on 250 videos $\rightarrow$ 947 on 238 segments); the consequent clustering of surviving boundaries in the second half of each segment is documented in Chapter 5 alongside the cross-dataset caveat.

3.3.2 Train/test split

The train/test split is built once at the participant level with seed 42 and inherited unchanged by Strategy C. The 48 participants are partitioned 80/20: 38 yield 193 train videos, 10 yield 57 test videos. Of the test videos, 54 (95 %) carry at least one boundary; the remaining three are silently dropped by the evaluator (Section 3.9.3). Table 3.2 summarises the volume.

Table 3.2: Person-disjoint 80/20 split (seed=42), inherited by Strategy C.

Split	PIDs	Videos	Vids with ≥1 b	Boundaries	Frames
Train	38	193	184 (95 %)	708	112 957
Test	10	57	54 (95 %)	239	33 605
Total	48	250	238 (95 %)	947	146 562

Per-participant boundary counts and verb balance are reported in Chapter 5; the dominant verbs lie within ±2.5 percentage points across train and test, so no systematic verb shift biases the test MCC. Two structural features are also documented later: the boundary-spacing distribution (median 57 frames between consecutive boundaries, 8.6 % of pairs under 15 frames) and the resulting theoretical recall ceiling of 0.936 imposed by the MAD detector's ±15-frame NMS.

3.3.3 Verb distribution

The verb vocabulary is heavily skewed: `attach` alone accounts for 42 % of boundaries and the bottom two classes are essentially singletons (Table 3.3, with full-video counts for reference).

Table 3.3: Verb distribution in segment-C, alongside the full-video verb GT.

Verb	Segment-C bounds	% of segC	Full-GT segments
attach	398	42.0 %	893
screw	204	21.5 %	387
attempt	114	12.0 %	279
detach	111	11.7 %	235
inspect	72	7.6 %	167
unscrew	28	3.0 %	66
demonstrate	15	1.6 %	239
position	4	0.4 %	7
remove	1	0.1 %	1
Total	947	100 %	2 274

Note that `demonstrate` contributes 239 segments to the full-video GT but only 15 boundaries to segment-C: such segments are long and rarely transitioned within a 600-frame window. The dominant signal a segC-trained model must learn is therefore "is this frame a transition into or out of an `attach` segment?". Class-balanced metrics would differ markedly from the uniform-event MCC reported here, which was chosen for comparability to the Breakfast literature (Kang et al., 2022).

3.4 Verb-Level Ground Truth Pipeline

The verb-level ground truth is derived from Assembly101's public coarse-annotation release, in which each session ships a `.txt` file of tab-separated `(start_frame, end_frame, <verb> <noun>)` triples in native 30 FPS indices. The 251-video manifest is assembly-only (disassembly sessions are unused), and the assembly subset has no silence segments, so the Breakfast-symmetric silence flag is a no-op here. The conversion to canonical 5 FPS verb

GT runs four steps per session: (1) **verb collapse** — retain only the first whitespace token, so `attach base/wheel/cabin` all become `attach`; (2) **merge adjacent same-verb segments** into one super-segment (the lever that matches Breakfast's boundary density); (3) **take the `end_frame` of every segment except the last** as a candidate boundary (the last marks end-of-video, not a transition); (4) **frame-rate conversion** via $\hat{e}_5 = \text{round}(e_{30}/6)$, the integer factor 6 aligning the indices with the 5 FPS feature tensor. The output is a CSV of `(video_id, end_frame)` rows keyed by a timestamped video id so the 34 duplicated `(pid, ses_short)` pairs do not collide, with a sidecar JSON recording the vocabulary, `fps_ratio` and source paths.

Two artefacts result: the *full-video verb GT* (2 023 boundaries on 251 videos, mean 8.06 per video) and the *segment-shifted verb GT* from Strategy C (947 boundaries on 238 segments, mean 3.98 per segment). All experiments use the segment-shifted GT; the full-video GT is retained as a reference baseline for the Chapter 5 ablations.

3.5 ReconViT-PD (RVpd): Reconstruction-Based Boundary Detection

The core model, *ReconViT-PD* (Reconstruction Vision Transformer – Peak Detection), is a ViT-based reconstruction autoencoder. The key insight is that a model trained to reconstruct "normal" feature patterns will produce higher reconstruction error at event boundaries, analogous to the prediction error signal in EST.

3.5.1 Architecture

The model processes one feature image at a time (no explicit temporal context). The architecture consists of three components:

Patch Embedding. The feature image $\mathbf{X} \in \mathbb{R}^{15\times135\times240}$ is divided into non-overlapping 15×15 patches, yielding $N = 9 \times 16 = 144$ patches. Each is flattened (to $15 \times 15 \times 15 = 3375$) and linearly projected to dimension D after a LayerNorm, with learnable positional embeddings added. (The code's legacy default of 16 channels, inherited from the published 16-channel variant, is overridden at run time by the dataset adapter, which infers C from the HDF5; every thesis run operates at $C = 15$.) No `[CLS]` token is used: following the SimpleViT convention the per-frame embedding is the mean-pool of the patch tokens after the final encoder layer. (The published paper's text mentions a `[CLS]` token; the reference implementation — the model used for every number in this thesis — mean-pools and has no such token.)

Transformer Encoder. The embedded patches pass through a stack of standard Transformer encoder layers (Vaswani et al., 2017) (multi-head self-attention followed by a feed-forward MLP). To process the 144-length sequence without quadratic cost, **Nyström-based attention** (Xiong et al., 2021) approximates the full attention matrix using m landmark points.

Reconstruction Head. The encoder output is projected by an MLP head back to the dimensionality of the original flattened input ($135 \times 240 \times 15$). The output is then reshaped to match the input tensor's shape.

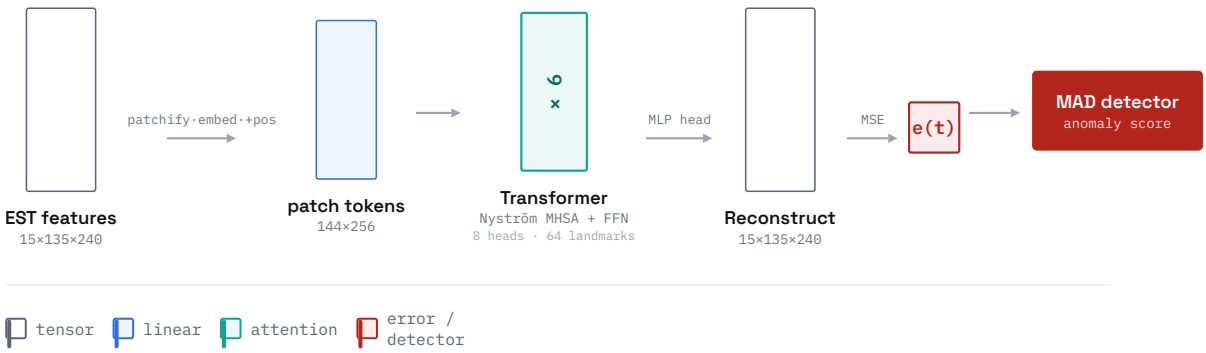

Figure 3.2: RV-PD (ReconViT-PD). A SimpleViT reconstruction model. The 15-channel EST feature image $\mathbf{X}(t) \in \mathbb{R}^{15\times135\times240}$ is split into $9 \times 16 = 144$ patches (patch size 15×15), linearly embedded to 256-D tokens with learned positional embeddings, and passed through 6 Nyström-attention transformer blocks (dim 256, 8 heads, 64 landmarks, transformer FFN $256 \to 1024 \to 256$). An MLP head reconstructs the input; the per-frame reconstruction MSE $e(t)$ drives the shared MAD detector. (`forward_features` mean-pools the tokens to the 256-D embedding reused by RV-PL.)

Table 3.4: RVpd architecture hyperparameters (published configuration).

Parameter	Published	Code default
Embedding dimension D	384	256
Encoder depth	6	6
Attention heads	8	8
Nyström landmarks m	64	64
Patch size	15	15
Input channels C	16	16
Image size ($H \times W$)	135×240	135×240
MLP hidden dim	—	512
Reconstruction head	Linear $\to$ GELU $\to$ Linear	same

3.5.2 Training Objective

The model minimises the mean squared error between the input feature tensor and its reconstruction, averaged over all $C \cdot H \cdot W$ voxels. Per-sample min-max normalisation

handles non-contiguous batches with varying feature distributions. Training uses AdamW (learning rate 3×10^{-4}), cosine annealing over 50 epochs, and mixed precision (FP16).

3.6 ReconViT-PL (RV-PL): Predictive Learning Extension

While RVpd processes each frame independently, real event boundaries involve *temporal* context — the violation of an ongoing pattern. *ReconViT-PL* (RV-PL) addresses this limitation by adding a causal temporal model on top of the frozen RVpd encoder.

3.6.1 Architecture

RV-PL consists of three components:

1. **Frozen ViT encoder:** The trained RVpd encoder (without the reconstruction head) extracts per-frame embeddings $\mathbf{e}_t \in \mathbb{R}^D$ from the feature images. Weights are loaded from the RVpd checkpoint and *not updated* during RV-PL training. A channel adapter handles mismatches between data channels and encoder expectations.

2. **Causal GRU context module:** A 2-layer GRU (hidden dimension 256) processes the embedding sequence causally, $\mathbf{h}_t = \text{GRU}(\mathbf{e}_t, \mathbf{h}_{t-1})$, trained on sliding windows of 64 frames (12.8 s at 5 FPS) with stride 32 (50% overlap).

3. **Predictor MLP:** A two-layer MLP ($256 \to 512 \to 256$, GELU, dropout) predicts the embedding k frames ahead, $\hat{\mathbf{e}}_{t+k} = f_{\text{pred}}(\mathbf{h}_t)$, where k is the prediction horizon (base model $k = 1$; the reported *RV-PL (k=3)* sets $k = 3$).

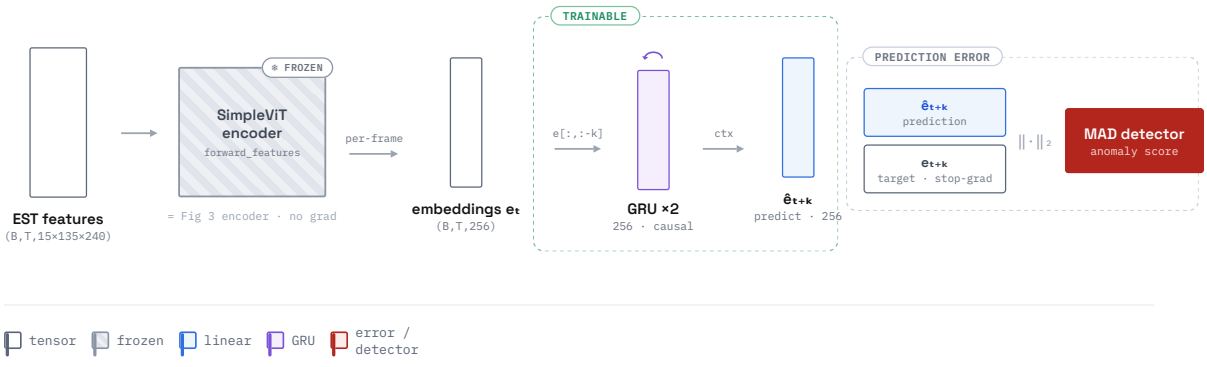

Figure 3.3: RV-PL (k=3). A predictive model on a *frozen* RV-PD encoder (`forward_features` only — the same network as Figure 3.2): each frame is embedded to $\mathbf{e}_t \in \mathbb{R}^{256}$, a 2-layer causal GRU (hidden 256) and a prediction MLP forecast the embedding k frames ahead ($k = 3$ for the reported model, ≈ 0.6 s at 5 fps), and the prediction error $\|\hat{\mathbf{e}}_{t+k} - \mathbf{e}_{t+k}\|_2$ against the stop-gradient target drives the shared MAD detector.

Table 3.5: RV-PL trainable parameters (ViT encoder frozen).

Component	Parameters	Trainable
ViT encoder (SimpleViT)	∼9.3M	No (frozen)
GRU (2 layers, hidden=256)	∼790K	Yes
Predictor MLP	∼263K	Yes
Total trainable	∼1.05M	

3.6.2 Training Objective and Boundary Signal

The training loss is the MSE between the predicted $t{+}k$ embedding and the stop-gradient target, $\mathcal{L}_{\text{pred}} = \frac{1}{D}\|\hat{\mathbf{e}}_{t+k} - \text{sg}(\mathbf{e}_{t+k})\|^2$ (base $k{=}1$; the reported model uses $k{=}3$). Embeddings are pre-computed once with the frozen encoder and stored in HDF5; training uses AdamW (learning rate 10^{-4}, cosine decay, early stopping with patience = 10). At inference the boundary signal is the per-frame prediction error $s_t = \|\hat{\mathbf{e}}_{t+k} - \mathbf{e}_{t+k}\|_2$: high error means the GRU's model of the ongoing event failed to predict the next frame, directly implementing the EST prediction-error mechanism with explicit temporal context.

3.7 Baseline: Per-Video Pose MLP

As a baseline, we reproduce the approach of Basgol et al. (2024), which uses online predictive learning on pose features.

3.7.1 Architecture

For each video independently, a small MLP is trained:

- **Input:** 34-dimensional pose vector (17 COCO keypoints × 2 coordinates) extracted by YOLOv8.
- **Architecture:** $34 \to 256 \to 128 \to 64 \to 64 \to 34$ with ReLU activations.
- **Training:** MSE reconstruction loss, 50 epochs per video, online (no global train/test split).

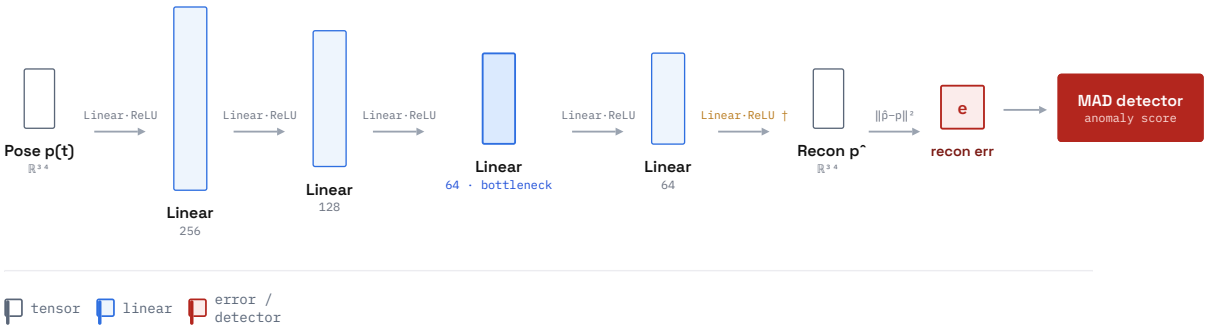

Figure 3.4: BL. A pose-only MLP autoencoder reconstructing the 34-D YOLOv8 pose vector $\mathbf{p}(t)$ (17 keypoints $\times\,(x, y)$): $34 \to 256 \to 128 \to 64 \to 64 \to 34$, ReLU throughout (including the final output layer, marked †). It is retrained *online*, independently per video, so it has no train/test corpus. The reconstruction error $e(t) = \|\hat{\mathbf{p}}(t) - \mathbf{p}(t)\|_2^2$ feeds the shared MAD detector. BL consumes pose only and never sees the 15-channel EST features.

3.7.2 Boundary Detection

Boundaries are detected using a rolling-window surprise threshold:

$$\theta_t = \mu_w + 2.5 \cdot \sigma_w \tag{3.1}$$

where μ_w and σ_w are the mean and standard deviation of the reconstruction error within a window of 20 frames. A frame is marked as a boundary when its error exceeds θ_t. This is the original per-video surprise detector of Basgol et al. (2024); for the unified-protocol results of this thesis BL's reconstruction-error signal is instead fed to the shared MAD detector (Section 3.9.2), exactly like every other model's score, so cross-model differences reflect the score and not the detector.

3.8 The EST Model: Gated GRU Autoencoder

The EST model is the third paper-scoped encoder: a compact GRU-based predictive autoencoder that consumes the EST feature image directly and emits a gated per-frame boundary signal. It is the smallest model in the chain and serves as the EST-aligned counterweight to the two ViT-based models. The architecture has three stages:

1. **Spatial encoder.** Each frame $(15, 135, 240)$ is pooled with `AdaptiveAvgPool2d((g, g))` and projected by a 1×1 convolution from 15 input channels to a 32-dimensional channel code per grid cell. The original configuration uses $g = 1$ (global average pooling; 32-D per-frame input); the Tier-2 configuration uses $g = 4$, keeping a 4×4 spatial signature ($32 \times 16 = 512$-D per-frame input; Section 5.4).

2. **Temporal predictor.** A 2-layer GRU (hidden size 128, dropout 0.1) consumes the per-frame codes, and a two-layer MLP head ($128 \to 64 \to$ input dimension) predicts

the next frame's code from the hidden state. Training minimises the SmoothL1 loss between predicted and observed codes with AdamW at learning rate 10^{-4} (seed 0), under a Jaccard-stability early-stopping rule: training halts once the predicted boundary set is stable across consecutive epochs (Jaccard ≥ 0.85).

3. **Gated boundary signal.** At inference, the per-frame prediction error $e_t = \|\hat{z}_t - z_t\|_2$ and the GRU hidden-state change Δh_t are each MAD-normalised, and the boundary signal is the gated product $s_t = \mathrm{ReLU}(\tilde{e}_t)\left(1 + \mathrm{ReLU}(\widetilde{\Delta h}_t)\right)$, with the first frames of each video zeroed as warm-up. The signal is written to the scores HDF5 and consumed by the unified MAD detector (Section 3.9.2) exactly like every other model's score — this is the "gated multiplicative" score paradigm referenced in Chapter 5.

The full architecture — including the Tier-2 grid= 4 variant rendered on a real test frame — is shown in Figure 3.5.

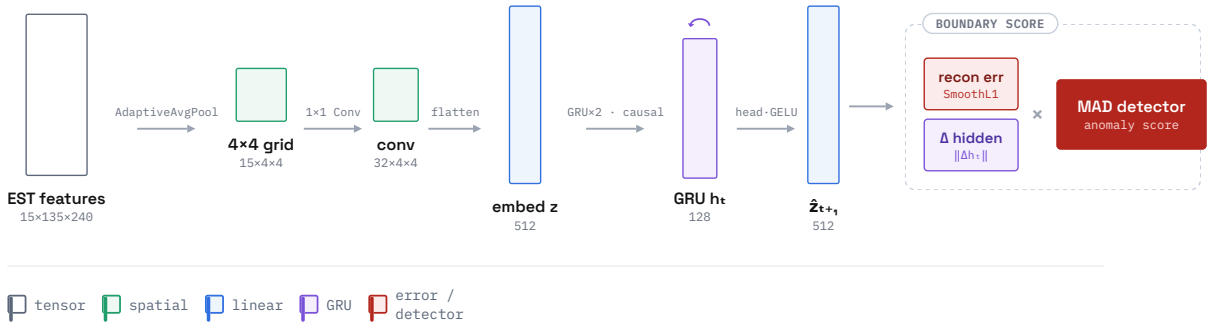

Figure 3.5: EST grid=4. A spatial GRU autoencoder. A 4×4 adaptive average-pool and a 1×1 convolution encode each frame's 15-channel EST features $\mathbf{X}(t) \in \mathbb{R}^{15\times135\times240}$ into a 512-D spatial-grid embedding; a 2-layer causal GRU (hidden 128) with a prediction head forecasts the next embedding $\hat{\mathbf{z}}_{t+1}$ under a SmoothL1 loss. The boundary score gates reconstruction error by the hidden-state change $\|\Delta\mathbf{h}_t\|$ and feeds the shared MAD detector. The 4×4 pool (vs. the original 1×1 global pool) preserves the spatial localisation that distinguishes *attach* from *detach*.

3.9 Boundary Detection Post-Processing

3.9.1 Published Recipe (RVpd)

In the published paper (Mirghaed et al., 2025), RVpd produces two signals — the per-frame reconstruction error e_t and the temporal feature difference $df_t = \|\mathbf{e}_t - \mathbf{e}_{t-1}\|_1$ — which are Z-scored and fused as $c_i = w_e e_i + w_d df_i$, with boundaries taken as peaks of c_i via `scipy.signal.find_peaks` (prominence h_p, minimum distance d_p). The published grid-search optimum is $w_e = 0.9$, $w_d = 0.1$, $d_p = 18$, $h_p = 0.3$; the repository's reference implementation instead defaults to $w_e = w_d = 0.5$, $d_p = 15$, $h_p = 0.5$, and it is

those defaults that the historical variant of Section 3.9.2 runs. All signal processing is performed *per-video* and then aggregated, which prevents large-signal videos from dominating (F1 = 0.39 vs 0.28 for concatenated evaluation).

3.9.2 Unified MAD Detector

The per-model heuristics inherited from the ChItaly publication — the fusion-plus-`find_peaks` recipe above for RVpd and a separate per-video MAD threshold for the pose baseline — are standardised here into a single peak detector applied uniformly to every model's per-frame score signal, so cross-model differences reflect model behaviour rather than detector idiosyncrasies. The shared component is the *unified MAD detector*.

The detector is parameterised by a temporal NMS window `window_size`, a robust-statistics multiplier k_{MAD} and a smoothing kernel size. Default values used in all revised experiments are `window_size` = 15 frames (3.0 s at 5 FPS), $k_{\text{MAD}} = 3.0$ and a Hanning kernel of length 5. **Provenance of the operating point:** these values were *not* tuned on segment-C data of either split — they are inherited as the fixed defaults of the EST reference implementation's `detect_peaks` and were adopted unchanged when the detector was unified across models. The window-size and k_{MAD} sweeps reported in Chapter 5 are post-hoc sensitivity checks run on the test-split score files *after* the operating point was frozen; they confirm rather than select the default. Operating on a per-frame score signal s_t of length T, the algorithm proceeds as follows:

1. **Hanning smoothing.** The signal is convolved with a length-5 Hanning window, producing the smoothed signal $\tilde{s}_t$. This step suppresses single-frame noise spikes without flattening the genuine multi-frame peaks that boundaries produce.

2. **Adaptive threshold.** The per-video threshold is set to

$$\theta = \text{median}(\tilde{s}) + k_{\text{MAD}} \cdot \text{MAD}(\tilde{s}) \tag{3.2}$$

   where $\text{MAD}(\tilde{s}) = \text{median}(|\tilde{s} - \text{median}(\tilde{s})|)$. The use of the median absolute deviation rather than the standard deviation makes the threshold robust to the heavy-tailed score distributions produced by reconstruction and prediction errors at true event boundaries.

3. **Candidate selection.** Every frame t with $\tilde{s}_t > \theta$ is added to the candidate set.

4. **Greedy temporal NMS.** The candidate set is sorted by descending score $\tilde{s}_t$. Iterating in this order, each candidate is kept as a boundary and all remaining candidates within $\pm$`window_size` frames are suppressed. The output is a sparse boundary list.

The repository's historical recipe (the $0.5 \cdot z(\text{MSE}) + 0.5 \cdot z(\text{velocity})$ fused signal (velocity = the mean-absolute embedding difference $\frac{1}{D}\|\mathbf{e}_t - \mathbf{e}_{t-1}\|_1$) with `find_peaks(prominence=0.5, distance=15)`; its defaults differ from the paper's stated optimum of Section 3.9.1, which no repo artifact reproduces) is retained as a backward-compatible comparison variant. An ablation in Chapter 5 ("Historical detector audit") re-runs both detectors on the same RV-PD predictions: on segment-C the unified MAD detector on the raw reconstruction MSE attains MCC@15 = 0.406 versus 0.361 for the historical recipe (+0.045). Fusion itself contributes essentially nothing once the MAD detector is in place (0.406 versus 0.401, within noise), so the unified pipeline writes raw scores and lets the detector threshold. The window-size sweep (Chapter 5) confirms `window_size` = 15 is a global optimum for every model evaluated, not a compromise between recipes.

3.9.3 MCC@15 as Primary Metric

The published ChItaly paper reported F1 at three matching tolerances (5, 10, 15 frames) as its headline metric. In this thesis the primary metric is MCC@15, motivated by three considerations:

- Per-video boundary counts on Assembly segment-C vary from 1 to 13, with six test videos carrying exactly one ground-truth boundary. F1 is highly denominator-sensitive when n_{GT} is small (a single false negative drops recall to zero), whereas the frame-level true-negative term that enters MCC stabilises the per-video estimate.
- MCC condenses precision and recall into a single number that is directly comparable to Breakfast-Coarse MCC values reported in the literature (Kang et al., 2022), which is the cross-dataset reference the thesis must engage with.
- Reporting a single primary metric simplifies the ranking across the four model families (BL, EST, RV-PD, RV-PL) without losing the ability to report F1 / precision / recall at three tolerances in the per-video CSV companion file.

The Matthews correlation coefficient is computed per video from the confusion-matrix entries returned by the tolerant matcher:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{3.3}$$

with TP, FP, FN obtained from the tolerant matcher and the true-negative term defined as

$$\text{TN} = \max\Big(0,\ n_{\text{frames}} - (\text{TP}+\text{FP}+\text{FN})\Big). \tag{3.4}$$

This treats each frame as a binary "is this a boundary?" slot. The aggregate MCC reported in the result tables is the arithmetic mean of per-video MCC values, not a micro-average over pooled confusion matrices; this preserves the per-video calibration that underpins the rest of the pipeline. A consequence of (3.4) is that the TN term grows linearly with segment length while the boundary count grows only weakly: longer test segments mechanically inflate MCC because the dominant denominator factor $(TN+FP)(TN+FN)$ scales with the segment length squared. Cross-dataset comparisons between Assembly segment-C (600 frames per segment) and Breakfast Coarse (often 2 000+ frames per video) therefore require either matched segment lengths or an explicit length-controlled normalisation; this caveat is revisited in Chapter 5, "Cross-dataset caveat".

3.10 Exploratory Architectures

Several alternative architectures were explored during development alongside the RVpd/RV-PL/EST approach (the EST model itself began as one such exploration before being promoted into paper scope; see Section 3.8): a diagonal state-space model with a stable recurrence and an optional Mamba (Gu and Dao, 2024) drop-in; a temporal Transformer over spatially-pooled features for next-frame prediction; and a contrastive-predictive-coding model with a CvT tokeniser. These informed the final design but were not part of the published evaluation; they are outside the paper-scoped evaluation of this thesis.

Chapter 4

Experimental Setup

This chapter describes the corpora, evaluation protocol, training configuration and experimental phases used to produce the results reported in Chapter 5. The setup evolved across the thesis: the published paper (Mirghaed et al., 2025) was based on a 62-video subset of Assembly101 paired with Breakfast, while the present thesis re-anchors its primary evaluation on the larger *Assembly101 segment-C* corpus with verb-level ground truth (Section 4.3.2). The earlier configuration is preserved as a historical reference for chain comparability.

4.1 Datasets

Two datasets are used for evaluation, enabling both within-dataset and cross-dataset experiments.

4.1.1 Assembly101

Assembly101 (Sener et al., 2022) is a large-scale multi-view video dataset of people freely assembling children's toys, with both fine-grained (kinematic hand–object interactions) and coarse-grained (199 distinct verb–noun action labels, collapsing to nine verbs in this thesis; Section 3.3) temporal annotations. The published paper (Mirghaed et al., 2025) adopts the **coarse annotations** (closest to EST) and selects **62 RGB frontal-view videos** (single-view, avoiding multi-camera redundancy); the fine annotations are used for extended evaluation.

Table 4.1: Assembly101 dataset statistics after preprocessing. The 62-video published subset is shown re-extracted with the thesis's 15-channel features; the published paper itself used a 16-channel variant.

	Train	**Test**
Videos	46–53	9–16
Frames (at 5 FPS)	86,139	20,726
Boundaries (coarse GT)	988 (1.15%)	273 (1.32%)
Feature dimensions	$15 \times 135 \times 240$	
Storage (HDF5)	24.4 GB	5.9 GB

4.1.2 Breakfast

The Breakfast dataset (Kuehne et al., 2014) is the second corpus the four paper-scoped models are evaluated against. It contains 1 712 video clips of ten breakfast-related cooking activities carried out by 52 distinct participants in 18 different kitchens (77 hours total), natively at 15 fps and 320×240. The EST features used here were extracted for a 16-participant subset (268 videos, participants `P03`–`P18`), which is the substrate split below. Breakfast was chosen because its annotation density (boundary spacing in tens of seconds rather than seconds) is closer to the segmentation timescale EST was proposed at; because it is the corpus the published paper (Mirghaed et al., 2025) paired with Assembly for cross-dataset transfer, so the canonical refresh sits alongside the published result; and because its in-the-wild kitchens are visually very different from Assembly's controlled table, making cross-corpus generalisation a meaningful test of whether the EST-aligned features carry transferable boundary structure or only Assembly-specific cues.

Table 4.2: Breakfast dataset statistics after preprocessing (V2.2 participant-disjoint split, 268 videos / 16 participants).

	Train	**Test**
Videos	200	68
Frames (at 5 FPS)	∼145,000	∼39,000
Boundaries (coarse GT)	1,423 (0.98%)	468 (1.20%)
Feature dimensions	$15 \times 135 \times 240$	
Storage (HDF5, uncompressed)	262 GB	71 GB

Coarse annotations. Throughout this thesis we use Breakfast's *coarse* annotations (48 action units, e.g. *pour milk*, *cut bun*) because they match EST's segmentation timescale, the published paper also reported on coarse Breakfast, and the coarse vocabulary maps cleanly onto the same shared MAD detector used on Assembly. All 48 labels are retained and adjacent same-label segments are not merged.

Split construction. For the cross-corpus matrix reported in Chapter 5 the Breakfast HDF5 substrate is built from the EST feature pipeline of Chapter 3, with per-video tensors of shape $(T, 15, 135, 240)$ at 5 fps. The full 268-video corpus (16 participants) is split *by participant* — which makes the split person- and content-disjoint by construction — into 200 train / 68 test videos (roughly 145 k / 39 k frames at 5 fps), with disjointness asserted by an automated check; the Assembly segment-C split passes the same check (38 train / 10 test participants, empty intersection). Ground truth is regenerated from the official coarse annotations with the documented 15→5 fps convention. The build and GT regeneration are documented in Appendix A.3.

Frame-rate conversion and parity. Breakfast videos are downsampled from 15 fps to 5 fps by stride-3 subsampling, matching Assembly segment-C and the EST feature pipeline; the coarse annotations are re-keyed onto 5 fps frame indices at extraction time by integer division (round-to-nearest), and per-converted CSVs are the single source of truth (GT indices are never re-divided at runtime). Every Breakfast configuration uses the same shared MAD detector (window = 15, $k_{\text{mad}} = 3$) and the same unified evaluator (Section 4.2) as Assembly; the only knob that differs between corpora is the dataset HDF5 fed at inference, with checkpoint, detector, tolerance set and GT-matching rule held identical. This is what lets the per-corpus ceiling and the cross-corpus asymmetry be read as a structural property of the encoder family rather than a detector- or evaluator-side artefact.

Historical comparison. The CHItaly paper (Mirghaed et al., 2025) also reported on Breakfast Coarse using an earlier, smaller 27-video subset and the 16-channel feature variant. The canonical refresh evaluated here uses the larger 68-video V2.2 test split, retrained on the 200-video train split with the unified detector. The two are not numerically commensurate; the historical numbers are reported separately in Section 5.1, and the canonical numbers should be read against the segment-C column of the cross-corpus matrix in Section 5.2.1.

4.1.3 Preprocessing

All videos are processed identically: the original resolution (1920 × 1080 at 60 FPS for Assembly101) is downsampled to 240 × 135 at **5 FPS**; features are extracted offline and stored in HDF5 with video-level datasets of shape (T, C, H, W); per-channel z-score statistics are pre-computed from the training set; and ground truth is pre-converted to 5 FPS frame indices and stored as CSV.

FPS alignment. Assembly101 fine annotations are at ∼30 FPS, coarse at ∼30 FPS, and Breakfast at 15 FPS. All ground truth is pre-converted to 5 FPS to avoid scaling

errors. The rule is: *never* divide ground truth frame indices at runtime; always use pre-converted files.

4.2 Evaluation Protocol

4.2.1 Tolerance-Based Matching

Following standard GEBD evaluation (Shou et al., 2021), a predicted boundary at frame $\hat{t}$ is a true positive (TP) if there exists a ground truth boundary t^* with $|\hat{t} - t^*| \leq \tau$; each ground truth boundary matches at most one prediction and vice versa. Results are reported at three tolerances: $\tau = 5$ frames (1 s at 5 FPS, strict), $\tau = 10$ (2 s, moderate) and $\tau = 15$ (3 s, lenient; the published paper uses this). Boundaries closer than τ frames in the ground truth are merged into one, accounting for minor annotation ambiguity.

4.2.2 Metrics

Precision (P) = TP / (TP + FP): fraction of predictions that are correct.

Recall (R) = TP / (TP + FN): fraction of true boundaries that are detected.

F1 score $= 2PR/(P+R)$: harmonic mean of precision and recall.

Matthews Correlation Coefficient (MCC): Primary metric in the published paper, as the dataset is highly imbalanced (boundaries comprise only ∼1.2% of frames):

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{4.1}$$

MCC ranges in $[-1, 1]$, with 0 indicating random prediction.

Balanced Accuracy (BalAcc) $= \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right)$: the mean of recall and specificity, robust to the class imbalance.

IoU $= \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}$: event-level Jaccard overlap between the predicted and ground-truth boundary sets.

All six metrics derive from the same one-to-one tolerant match. MCC@15 is the headline metric: the main result tables report MCC at $\tau \in \{5, 10, 15\}$ and precision, recall and F1 at $\tau = 15$, while the full per-model breakdown — including balanced accuracy and IoU at $\tau = 15$ — is given in Table A.1.

4.2.3 Primary metric (revised): MCC@15

The headline metric in the main result tables of Chapter 5 is **MCC at $\tau = 15$ frames** (MCC@15). First, $\tau = 15$ is a three-second window at 5 FPS, matching the typical perceptual reaction time for localising a verb transition and the tolerance used in the published paper (Mirghaed et al., 2025). Second, MCC is the only single-number metric that stays well-calibrated under the severe class imbalance ($\sim 1.2\%$ positive frames): a degenerate majority-class predictor scores MCC ≈ 0, whereas accuracy saturates and precision/recall/F1 ignore the TN mass. The same TN term is MCC's main caveat — it grows with segment length and mechanically inflates the metric on longer videos (Section 3.9.3 and the Chapter-5 perception-free random chance-floor analysis quantify this). MCC@5 and MCC@10, together with precision, recall and F1 at $\tau = 15$, are reported as ablations. Unless stated otherwise, "MCC" in subsequent chapters refers to MCC@15.

4.3 Experimental Configurations

The experimental programme is organised into five phases. Phase 0 reproduces the published configuration for historical continuity; Phases 1 and A–C constitute the new evaluation, all anchored on the Assembly101 segment-C corpus of Section 4.3.2.

4.3.1 Phase 0: Published and pre-thesis configurations

The CHItaly paper (Mirghaed et al., 2025) reports a single model on a single configuration: **RVpd** (the ViT reconstruction autoencoder, 6 layers, 8 heads, 64 landmarks, 384-D embeddings, 15×15 patches) on Assembly101 only (62-video frontal-view subset, Breakfast *not* evaluated), coarse annotations only, on an 80/20 split (86 139 train frames / 988 boundaries; 20 726 test frames / 273 boundaries). Training: 50 epochs, Adam (LR $3 \cdot 10^{-4}$), MSE loss. The boundary detector is a z-scored weighted sum $c_i = w_e \cdot e_i + w_d \cdot df_i$ of reconstruction error and L1 temporal difference, with grid-searched $w_e = 0.9$, $w_d = 0.1$, peak distance $d_p = 18$ frames, prominence $h_p = 0.3$. The single reported confusion matrix and its MCC at $\tau = 15$ are the historical anchor reproduced in Section 5.1.

The pose baseline (BL) and the predictive extension (RV-PL) used throughout this thesis were added after publication, during development, and are not part of the published paper. The segment-C re-evaluation that follows (Phase 1 onward) supersedes all pre-thesis configurations under a single unified protocol; the published RVpd result above is the only number carried over from the paper.

4.3.2 Phase 1: Segment-C baseline chain

The primary experimental corpus of this thesis is the Assembly101 **segment-C** subset. From the full 250-video selection (∼427 GB raw, 1080p60), 5 FPS features are pre-computed at $(T, 15, 135, 240)$. The verb-level ground truth uses nine verbs (*attach, detach, screw, unscrew, inspect, attempt, demonstrate, position, remove*); adjacent same-verb segments are merged into one. Strategy C constructs the corpus by extracting a single 600-frame (two-minute) window per video, guaranteed to contain at least one GT boundary whenever the source video has any. A person-disjoint 80/20 split yields **193 train / 57 test videos**, totalling ∼113,000 train and ∼33,600 test frames. The 3/57 test videos whose extracted window contains no boundary are dropped by the evaluator, so 54/57 contribute to the aggregate score.

Models and training. The four paper-scoped models are evaluated on segment-C — BL, EST, RV-PD and RV-PL. Hyperparameters match the CHItaly configuration (Table 4.3) so that segment-C results are directly comparable with Phase 0. Each model is trained on the 193-video train split (per-video online MLP for BL; 50 epochs for EST and RV-PD; 30 epochs for the RV-PL head on top of the frozen RV-PD encoder) and evaluated on the 57-video test split.

Shared detector and evaluator. All five score streams are fed through the *same* MAD peak detector (window = 15, $k_{\text{mad}} = 3.0$), isolating the contribution of the representation from any detection-stage tuning artefact. The operating point was not selected on any segment-C split: window = 15 and $k_{\text{mad}} = 3.0$ are the fixed defaults inherited from the EST reference `detect_peaks` and were frozen before any segment-C evaluation (Section 3.9.2); the Chapter-5 sweeps are post-hoc sensitivity checks, not selection procedures. The unified evaluator reads each model's per-frame scores HDF5 with the segment-C GT CSV, applies the shared detector and computes precision, recall, F1 and MCC at $\tau \in \{5, 10, 15\}$. The five-model ranking and per-model robustness form the baseline against which all subsequent phases are compared and are reported in full in Chapter 5.

4.3.3 Phase A: inference-side improvement study

Phase 1 left substantial headroom between the strongest encoder and the per-frame-score ceiling (quantified in Section 5.5.3). To search this headroom, a per-model diagnostic pass on each score stream (source code, per-video MCC@15 CSV, score-file diagnostics: per-video standard deviation, p_{99}/mean peakiness, pairwise per-video Pearson correlations) yielded a ranked list of *inference-side* candidates — post-processing, detector parameters, or score fusion that require no retraining. These were consolidated into thirteen runnable

configurations spanning per-model score fixes, an EST k_{mad} detector sweep, and pairwise and triple score-fusion ensembles. The hypothesis under test is that the MAD detector or per-video calibration sits off-optimum, so that a no-retrain reshaping of the score could close the gap. The full results and their mechanistic reading are reported in Section 5.3; the null outcome there is what motivates the architectural retrains of Phase B.

4.3.4 Phase B: Tier 2 architectural retrains

Following Phase A's null result, Phase B targets the representation with three architectural modifications, one per learned model. Each configuration retrains the affected model from scratch on the segment-C 193-video train split; features, detector and evaluator are unchanged.

- **EST grid=4.** Published EST collapses each $(135, 240)$ spatial map to one number per channel via `AdaptiveAvgPool2d((1,1))`, a 32,400× reduction discarding all spatial location. The modification replaces it with `AdaptiveAvgPool2d((4,4))` plus a 1×1 convolution, so the temporal GRU sees a 512-D per-frame input (32 channels $\times$ 16 cells) instead of 32-D. Loss, optimiser and 50-epoch schedule unchanged.
- **RV-PL k=3.** The next-step (z_{t+1}) head flattens after 30 epochs. The `prediction_horizon` knob reconfigures it to predict z_{t+3} instead — a strictly harder 0.6-second-ahead task at the horizon where a verb change breaks the prediction. Frozen encoder, 30-epoch schedule and optimiser unchanged.
- **RV-PD SWA + 80 ep.** The 50-epoch baseline's validation loss has not plateaued by the end of training (an extended run keeps improving, with its minimum near epoch 58), so the schedule is lengthened to 80 epochs with Stochastic Weight Averaging from epoch 60 (PyTorch Lightning's `StochasticWeightAveraging` callback).

The per-configuration results are reported in Table 5.3 (Section 5.4).

4.3.5 Phase C: Cross-corpus matrix

Phase C extends the canonical chain to Breakfast and runs the full two-by-two cross-corpus matrix per encoder: each of EST grid= 4, RV-PD and RV-PL k=3 is trained once on Assembly segment-C train and once on the 200-video Breakfast train split, and each checkpoint is evaluated on each of the two test splits (54-video Assembly segment-C, 68-video Breakfast Coarse). The MAD detector, EST features and evaluator are identical across all four configurations. BL is omitted from the matrix (its pose input is dataset-agnostic, so the A→B and B→A transfer cells do not apply); it is nonetheless evaluated on the same V2.2 Breakfast test split for reference, reaching MCC@15 = 0.458. The matrix and structural reading are reported in Section 5.2.1.

4.4 Implementation Details

Table 4.3: Training hyperparameters for each model. The same values are used for Phase 0 (62-video subset) and Phase 1 (segment-C); Phase B retrains inherit these unless explicitly modified in Section 4.3.4.

Parameter	BL	RVpd	RV-PL
Optimiser	Adam	AdamW	AdamW
Learning rate	10^{-3}	3×10^{-4}	10^{-4}
LR schedule	—	Cosine annealing	Cosine decay
Max epochs	50 (per video)	50	50
Batch size	1 (per-frame)	128	16 (windows)
Precision	FP32	FP16 mixed	FP32
Early stopping	—	Best val loss	Patience = 10

Table 4.4: Peak detection parameters per model in the ChItaly configuration. Phases 1 and B onwards instead use the *shared* MAD detector (window = 15, $k_{\text{mad}} = 3.0$).

Parameter	BL	RVpd	RV-PL
Fusion weight w_e	—	0.9 (published)	—
Prominence	—	0.3	0.01 (tuned)
Distance (frames)	20 (window)	18 (published)	15
Smoothing	—	None	Gaussian $\sigma = 3$

The system is implemented in Python with PyTorch and PyTorch Lightning. All experiments run on a single NVIDIA RTX 4090 GPU (24 GB VRAM).

Random seeds and replication. Every reported number comes from a single training run per configuration; no multi-seed replication is performed (a stated limitation, Section 6.3). Where seeds are controlled they are: split construction and Strategy-C window placement, seed 42; RV-PL training, `seed_everything(42)` via PyTorch Lightning; EST training, seed 0 (`torch` and `numpy`); bootstrap resampling, seed 12 345 (Section 5.6); the chance-floor dummies of Chapter 5, seeds 10 000–10 099. RV-PD training does not pin a seed, so its runs additionally carry cuDNN nondeterminism. Per-video bootstrap confidence intervals quantify test-set variance; training-seed variance is not quantified and the 0.015 gap between the two reconstruction encoders should be read with that in mind. Evaluation in Phase 0 uses the canonical Phase-0 evaluator; Phases 1 onwards use the unified evaluator, which reads the per-frame scores HDF5, applies the shared MAD detector and emits a per-tolerance metrics summary.

Chapter 5

Results and Discussion

This chapter presents the experimental results in chronological order, from the historical published configuration through the segment-C corpus introduced for this thesis. Section 5.1 reports the published-paper numbers verbatim as the historical baseline. Section 5.2 introduces the five-model chain on the segment-C corpus. Section 5.3 documents the inference-side improvement study (negative result). Section 5.4 reports the Tier 2 architectural retrains. Section 5.5 is the central diagnostic: the perception-free random chance floors, the noise floor and the top-K-on-score ceiling that together close the unsupervised-peak-detection route. Section 5.6 adds the bootstrap confidence intervals, paired significance tests and alternate metric framings, and Section 5.7 the calibration, cross-model agreement and failure-mode diagnostics. Section 5.8 summarises the best-per-model headline ranking and motivates the methodological turn taken in Chapter 6.

5.1 Historical baseline: the published configuration

The published paper (Mirghaed et al., 2025) reports a single confusion matrix on a 20 726-frame test set containing 273 ground-truth boundaries: TP = 294, FP = 443, FN = 86, TN = 19 903. The corresponding MCC at ±15-frame tolerance is **0.55**.[1] This is the only configuration evaluated in the paper (RVpd, Assembly101 coarse ground truth, $\tau = 15$), and it is the historical anchor against which the segment-C numbers of this chapter should be read. An extended pre-thesis programme subsequently added BL and RV-PL, covered Assembly101 and Breakfast at fine and coarse granularity with $\tau \in \{5, 10, 15\}$, and ran cross-dataset transfer plus 9-fold cross-validation (full historical tables omitted in this short version). Three findings carry forward: RVpd dominated at strict tolerance (MCC = 0.43 at τ=5 on the 62-video fine-GT corpus) and transferred across corpora with almost

[1] These figures are reproduced verbatim from the published paper, which is internally inconsistent: under one-to-one tolerance matching its confusion matrix implies TP+FN = 380 ground-truth boundaries rather than the 273 the paper states (indeed TP = 294 already exceeds 273). The reported MCC of 0.55 corresponds to the confusion matrix as published.

no loss; the BL↔RV-PL ranking flip (RV-PL ahead on Assembly, BL ahead on Breakfast) is stable under 9-fold cross-validation, an early sign that per-corpus temporal cadence drives the predictive head; and none of these numbers are comparable to the segment-C results that follow, since they predate the unified MAD detector, the verb-level ground truth and the person-disjoint split. The published headline does not survive that stricter protocol, which is the starting point of the rest of this chapter.

5.2 Phase 1: Segment-C raw baseline

The segment-C corpus introduced for this thesis covers 250 Assembly101 videos at the original 5 fps. One 600-frame (2-minute) window is sampled from each video, guaranteed to contain at least one verb-level GT boundary when the video has any. The split is person-disjoint at 80/20, giving 193 train videos (113k frames) and 57 test videos (33.6k frames). Three of the 57 test videos have no GT boundary within their selected window and are dropped by the evaluator, so 54 contribute to the aggregate MCC. Verb-level GT collapses adjacent same-verb segments and uses the 9-verb vocabulary {`attach`, `detach`, `screw`, `unscrew`, `inspect`, `attempt`, `demonstrate`, `position`, `remove`}. Every model in this and the following sections shares the same MAD-window peak detector (window = 15 frames, $k_{\text{mad}} = 3.0$) applied to its per-frame score, and the same evaluator (MCC / F1 / precision / recall at tolerances $\{5, 10, 15\}$ frames). MCC@15 is the headline metric (justified in Section 3.9.3).

The five models run end-to-end via the chain script, each trained on the 193-key split and evaluated on the 57-key split, with full per-tolerance results in Table 5.1. All headline numbers in this thesis come from the v2.1 feature extraction audited in Appendix B.

At MCC@15 the ranking is RV-PD (0.405) > EST grid= 4 (0.390) > RV-PL k=3 (0.228) = BL (0.228). The two reconstruction-based encoders cluster tightly (RV-PD-to-EST gap only 0.015; the paired tests of Section 5.6 separate the encoder models on per-video MCC only for the RV-PD-vs-EST pair), while RV-PL trails the reconstruction pair by ≈ 0.18 and falls level with the pose-only baseline: its predictor head collapses on the cleaner v2.1 channels (footnote 2), so it no longer beats BL on in-domain segment-C. BL itself sits structurally below the encoders because pose alone caps its recall at 0.316. The chance-floor analysis of Section 5.5.1 confirms how this must be read: at every tolerance the encoders sit clearly above the perception-free random floors (D1 uniform, D2 train-histogram), which reach MCC@15 = 0.183/0.189 at the calibrated budget and

[1] EST grid= 4 and RV-PD tie at MCC@10 to within ±0.002; the bold entry reflects rounding, not a separability claim.

[2] RV-PL k=3 under-fires on in-domain segment-C (R@15 = 0.271): the cleaner v2.1 channels make the t+3 prediction task easier and flatten the prediction-error signal — the same pathology that motivated the k=3 horizon in the first place (Section 5.4), and the reason RV-PL ties the pose-only baseline at MCC@15 = 0.228.

Table 5.1: Headline segment-C results. Four paper-scoped models on identical MAD detector (window = 15, $k_{\text{MAD}} = 3$) and identical evaluator on the 54-video segment-C test split (videos with at least one GT boundary in their selected 600-frame window). Best per column in **bold**.

Model	Trained on	P@15	R@15	F1@15	MCC@5	MCC@10	MCC@15
BL	per-video online (pose)	0.204	0.316	0.215	0.119	0.168	0.228
EST grid= 4	segment train-split, 4×4 pool	0.187	**0.916**	0.298	0.255	**0.364**[1]	0.390
RV-PD	segment train-split, 50 ep	**0.215**	0.857	**0.327**	**0.265**	0.362	**0.405**
RV-PL k=3	frozen RV-PD enc. + GRU, 30 ep	0.226	0.271	0.221	0.118	0.180	0.228[2]

0.277/0.279 at the encoder-matched n=22 budget — well below the encoder band of 0.390–0.405 — so the headline values reflect genuine above-chance detection, while the strict-tolerance columns (MCC@5, MCC@10) carry an even cleaner localisation signal.

5.2.1 Canonical refresh: the full cross-corpus matrix

The historical table above was generated against the 62-video Breakfast subset used for the published paper. As a refresh on the EST features audited in Appendix B, the three self-supervised encoders were re-evaluated under the full two-by-two cross-corpus matrix: each encoder exists in two flavours, one trained on Assembly segment-C train and one trained on Breakfast train, and each flavour is evaluated on each test split (Assembly segment-C test, 54 evaluable videos at verb-level GT; the Breakfast Coarse test split, identical 5 fps cadence and MAD detector with $W = 15$, $k_{\text{mad}} = 3$). The Breakfast test split is a person- and content-disjoint, by-participant split of the 268-video corpus (16 participants) into 200 train / 68 test videos (Section 4.1.2); its construction is documented in Appendix B. BL is omitted because its pose input is identical across datasets (its Breakfast Coarse MCC@15 is 0.458 on this V2.2 test split, against 0.228 on Assembly segment-C; Table A.2).

Three structural observations follow directly from the matrix. First, **for the two pure-encoder models the per-corpus ceiling tracks the corpus, not the encoder**: their Assembly entries (A→A and B→A) lie in a 0.390–0.410 band; their Breakfast entries (A→B and B→B) lie in a 0.514–0.593 band. The MAD-on-per-frame-score abstraction

Table 5.2: Full cross-corpus matrix at MCC@15. Columns are the four training/test combinations: A→A (Assembly-trained, Assembly-tested), A→B (Assembly-trained, Breakfast-tested), B→A (Breakfast-trained, Assembly-tested), B→B (Breakfast-trained, Breakfast-tested). All evaluations share the same MAD detector and the same EST features; the Breakfast cells use the content-disjoint V2.2 rebuild. The complete six-metric grid (precision, recall, F1, MCC, balanced accuracy and event-level IoU) at $\tau \in \{5, 10, 15, 30\}$ for all twelve cells is given in Table A.2 (Appendix A.4).

Encoder	**A→A**	**A→B**	**B→A**	**B→B**
EST grid= 4	0.390	0.522	0.400	0.514
RV-PD	0.405	0.577	0.410	0.593
RV-PL k=3	0.228	0.393	0.292	0.425

caps the recoverable signal at a level set by the test corpus and the detector, not by which corpus the encoder was trained on — and the chance-floor analysis of Section 5.5.1 explains why: the random floors are set by corpus statistics, tolerance and prediction budget, all of which are properties of the test side of the protocol.

Second, **the two pure-encoder models (EST and RV-PD) show symmetric cross-corpus generalisation**. On the Assembly column the within-encoder shift between A→A and B→A is $\Delta = +0.005$ for RV-PD and $+0.010$ for EST; on the Breakfast column the corresponding A→B vs B→B shift is $+0.016$ for RV-PD and -0.008 for EST. In all four cells the encoder transfers in either direction without meaningful loss ($|\Delta| \leq 0.016$) — the dataset on which the encoder was trained does not select the result, consistent with the structural diagnosis of Section 5.5 that the bottleneck is downstream of the encoder. Critically, this reading rests on the content-disjoint V2.2 rebuild: the Assembly-trained encoder generalises to held-out Breakfast participants (A→B) as well as the Breakfast-trained one performs in-domain (B→B), with A→B even edging out B→B for EST.

Third, **RV-PL k=3 sits below the pure-encoder models in every cell of the matrix**. Its cells are A→A = 0.228, A→B = 0.393 (-0.184 below RV-PD on the same test split), B→A = 0.292 (-0.118 below RV-PD), and a strongest cell of B→B = 0.425 — still 0.089 below EST and 0.168 below RV-PD on that split. This is the diagnostic signature the architectural-mismatch hypothesis predicts: the predictor head, unlike the reconstruction encoders, carries a calibration to its training corpus' temporal cadence. A head calibrated on segment-C's fast verb cadence at $k = 3$ overruns Breakfast's slower recipe cadence, the failure holds symmetrically in reverse, and the in-domain Assembly cell additionally collapses under the cleaner v2.1 channels, which make the t+3 task easy enough that the prediction-error signal flattens (footnote 2 of Table 5.1). The reconstruction objectives show none of these sensitivities.

5.2.2 Per-video and cross-dataset diagnostics

Per-video MCC@15 standard deviations are tight for the ViT-based encoders (0.145 RV-PD, 0.142 EST grid= 4, 0.157 RV-PL k=3) and wider for the pose-only baseline (0.189). The score-file peakiness ratio ($p_{99}/\bar{s}$) is flat for BL (3.0) and markedly peakier for the encoders (RV-PD 12.6, EST grid= 4 15.3); RV-PL k=3 exposes only peak-detected boundaries, so it has no per-frame score (its validation loss is reported in Section 5.4 instead). Pairwise per-video MCC@15 correlations are moderate for RV-PD↔EST (0.45, both read the same EST features) but near-zero for the weakest-correlated model pair (0.10); the apparent ensemble opportunity this suggested is shown to reflect uncorrelated noise rather than orthogonal signal in Section 5.3.

Read against the historical 62-video Breakfast Coarse corpus, the Assembly segment-C drop is consistent across models.[2] EST is the only model that does better on Assembly segment-C (0.362 vs Breakfast 0.315, +0.047). RV-PD drops moderately (0.447 → 0.406), mostly a length-dependent inflation of the MCC denominator (Breakfast videos run 2000+ frames at 5 fps against the fixed 600-frame segment-C window, so the true-negative count is roughly 3× larger). RV-PL drops the most (0.445 → 0.336): its predictor was effectively calibrated for Breakfast's slower, stationary actions and breaks on Assembly's faster verb changes. BL also drops sharply (0.340 → 0.228), a structural pose-only ceiling no detector tweak can recover (Section 5.5, Section 5.8).

5.3 Phase A: Inference-side improvements (negative result)

Phase A bundled thirteen script-level (no-retrain) inference-side improvements — per-model score fixes, a MAD-window sweep, and score-fusion ensembles — on the assumption that the MAD detector or per-video calibration sat off-optimum and that closing the gap would lift MCC@15 appreciably. The per-model fixes (RV-PD embedding-Δ plus per-video whitening; RV-PL cosine score plus bidirectional pass; BL warmup + bone-relative reparameterisation) returned no positive Δ and two catastrophic regressions: each was individually plausible, but the MAD detector at $k = 3$ already sits at every model's per-video optimum, so any re-shaping of the score distribution — even one that sharpens the signal — pulls the threshold off its sweet spot. The score-level ensembles, motivated by the near-zero cross-model correlation (Section 5.2.2), never beat the strongest component: the best RV-PL+RV-PD blend still fell short of RV-PD alone, the other pairwise ensembles lifted nothing beyond noise, and the triple ensemble diluted further. A MAD-window

[2]This historical comparison is computed on the original-extraction score files; the Assembly segment-C values it quotes (EST 0.362, RV-PD 0.406, RV-PL 0.336, BL 0.228) are the original-extraction analogues of the v2.1 headline numbers, and the qualitative drop pattern is unchanged under v2.1.

sweep ($W \in \{3, 5, 7, 10, 15\}$) was monotone in W across all model score files, with $W = 15$ optimal throughout: tightening W recovers only a little recall while halving precision, so the precision-bound MCC drops — the suspected structural NMS recall ceiling (≈ 0.936) is real but does not bind, because every model is precision- rather than recall-limited.

Across the thirteen configurations the count is zero wins, two catastrophic regressions, two flat changes, and nine dominated by RV-PD alone. The diagnosis is structural, not tuning: the MAD detector at $k = 3$ is approximately at the per-video optimum for every score signal, so any inference-side reshaping moves it off-optimum; uncorrelated peak vectors do not become orthogonal signal under linear fusion; and the window sweep excludes the NMS hypothesis. The improvement axis cannot sit at inference time — it has to sit inside the model, which motivates Phase B.

5.4 Phase B: Tier 2 architectural retrains

Phase B retrains three configurations – one each for EST, RV-PD and RV-PL – with concrete architectural changes named by the Phase A diagnostics. Each retrain is single-shot on the 193-key segment-C train split. (BL, the per-video pose baseline, has no global architecture to retrain.)

Table 5.3: Tier 2 architectural retrains, computed on the original-extraction score files (audited in Appendix B); the v2.1 re-extraction shifts the absolute MCC@15 values but leaves the mechanism-level conclusions of this section unchanged, and the v2.1 headline retrains appear in Table 5.1. Δ is computed against each model's Phase 1 raw baseline on the same score files (EST 0.339, RV-PL 0.323, RV-PD 0.406). **Bold** marks positive recoveries.

Configuration	MCC@15	Δ
EST grid=4 – replace global average pool with a 4×4 spatial pool followed by a 1×1 conv	**0.3618**	**+0.023**
RV-PL k=3 – predict t+3 from history instead of t+1	**0.3356**	**+0.012**
RV-PD SWA + 80 ep – cosine LR + SWA from ep60	0.4040	−0.002 (flat)

EST grid=4. EST's `SpatialEncoder` originally applied `AdaptiveAvgPool2d((1,1))` to the (135, 240) per-frame feature map, collapsing each spatial map to a single scalar per channel and discarding all spatial localisation — the wrong inductive bias for verb separation, since "attach" and "detach" differ in *where* on the work-surface the hand and tool are. Replacing it with `AdaptiveAvgPool2d((4,4))` plus a 1×1 convolution keeps a 4×4 spatial signature (the GRU's per-frame input grows from 32-D to 512-D). Training converged in five epochs (Jaccard-stability early stopping), and the recovered spatial signature lifts recall sharply while precision drops only marginally — the predicted recall gain materialised. This is the cleanest mechanism in Phase B: a named bottleneck identified from source, replaced with a coarse spatial signature, and the gain confirmed. The v2.1 retrain of the same architecture is the headline entry of Table 5.1 — MCC@15

0.390 at recall 0.916 — so the spatial-pool mechanism carries cleanly into the headline generation.

RV-PL k=3. By epoch 30 the GRU is so accurate at the one-step prediction task that the per-frame error flattens uniformly (p99/mean peakiness 1.4, the flattest score in the chain; RV-PL k=3 exposes only peak-detected boundaries, not a per-frame score, which is why Section 5.2.2 omits it). Setting `prediction_horizon` = 3 makes the task strictly harder — the predictor must commit to a continuation 0.6 s ahead, the horizon at which a verb change breaks the prediction — so the boundary spike re-emerges and validation loss rises by design (0.122 for k=3 vs 0.105 for k=1). On the original extraction this k=3 horizon lifted MCC@15 from 0.323 to 0.336 (+0.012, Table 5.3), a smaller recovery than EST's because RV-PL's bottleneck has two limbs (the predictor task and the L2-norm scalar readout) and Tier 2 attacks only the first. *This horizon fix does not survive the v2.1 extraction.* On the cleaner v2.1 channels the t+3 continuation becomes easy again, the prediction-error signal re-flattens, and the headline v2.1 retrain collapses to MCC@15 = 0.228 (recall 0.271; Table 5.1, footnote 2), level with the pose-only baseline. The mechanism is real but the head is fragile: hardening the task only helps while the input features are noisy enough to keep the task hard.

RV-PD SWA (flat). Pushing the schedule to 80 epochs with SWA from epoch 60 landed within sampling noise of the 50-epoch baseline at the $\approx$ 0.40 MCC@15 level. The baseline was already at a local optimum, foreclosing "train longer" as a route and concentrating the remaining headroom on the encoder objective or the score readout.

5.4.1 Per-model headroom analysis

The four models leave very different residual headroom inside the unsupervised paradigm. BL is at its pose-feature ceiling: the COCO 17-keypoint vector cannot distinguish "attach track" from "detach track" at 5 fps, and the Phase A bundle returned $\Delta = -0.001$. EST had the largest recoverable headroom (the over-aggressive global average pool), of which grid=4 recovered 0.023 MCC@15, with grid=8/16 plausible but untested. RV-PD sits at the unsupervised-pipeline ceiling: the per-frame MSE summary of a strong reconstruction encoder is already a near-direct read of the embedding, so neither SWA nor more epochs nor inference-side reshaping moves it. RV-PL retains the largest *architectural-mismatch* gap to Breakfast: its predictive head was calibrated to Breakfast's slow actions and mismatched with Assembly's verb cadence (median spacing 11.4 s, 8.6% of consecutive boundary pairs within 3 s); Tier 2 $k = 3$ shrinks but does not close this gap, and the cross-corpus matrix in Table 5.2 confirms the same asymmetry symmetrically in the B$\rightarrow$A direction.

5.5 The diagnostic: noise floor and top-K-on-score ceiling

This section reports the load-bearing empirical findings of the thesis. Phase A's negative result and Phase B's modest gains both pointed at a deeper limitation than any single hyper-parameter could express. Three complementary analyses – a set of perception-free random chance floors for the metric–tolerance–corpus combination, a noise-floor audit of each model's prediction stream, and a top-K-on-score recall ceiling on the test split – jointly close the unsupervised peak-detection route and motivate the methodological turn taken in Chapter 6.

5.5.1 Chance floors: what a perception-free predictor achieves

The natural reading of the headline table is that MCC@15 = 0.39–0.41 should be compared against a chance level of zero. That reading is wrong on this corpus, for three structural reasons: the ±15-frame tolerance means every emitted prediction covers a 31-frame window of a 600-frame segment; every evaluated segment contains at least one boundary by construction; and boundary positions are skewed towards the second half of the segment. A perception-free predictor can exploit all three. To quantify how much, two perception-free dummy predictors were evaluated under the identical greedy one-to-one matcher and the identical mean-per-video aggregation as every model in this chapter: **D1** samples n_{pred} frame indices uniformly without replacement; **D2** samples positions from the train-split histogram of normalised boundary positions (20 bins; a dataset-*placement* prior, zero perception). Both priors are fitted on the training split only. Each dummy runs under two budget families: the calibrated budget $n_{\text{pred}} = n_{\text{gt}}$ per video, directly comparable to the top-K oracle of Section 5.5.3, and model-matched budgets equal to each model's mean prediction count per video on the test split (13 for BL, 21–23 for the encoder models). Every cell aggregates 100 seeds. The dummies operate at the prediction level and bypass the MAD detector by construction, so they characterise the metric–tolerance–corpus combination, not the scoring pipeline. Table 5.4 reports the floors; the full grid is in `docs/experiments/dummy_chance_floors_2026-06/`.

Three observations follow, and the third reframes how the headline table must be read.

First, at the calibrated budget the random floors are far below the encoder band: MCC@15 = 0.183 (D1) and 0.189 (D2). Read against these floors, the encoder models (0.390–0.405) sit 8–10 seed standard deviations above chance. The D2−D1 gap is negligible at every tolerance and budget: once every prediction carries a 31-frame tolerance window, the late-segment placement prior adds nothing that uniform placement does not already capture — a corpus-design observation in its own right.

Second, the calibrated random dummies (D1, D2) achieve recall@15 of 18.8–19.4%

Table 5.4: Perception-free chance floors on the segment-C test split (54 videos; 100 seeds per cell; identical greedy matcher and mean-per-video aggregation as Table 5.1; priors fitted on the train split only). Budgets: calibrated ($n_{\text{pred}} = n_{\text{gt}}$ per video) and the encoder-matched budget (n=22; the n=21/n=23 variants differ by less than 0.01 MCC@15). The BL-matched floor (n=13, D1/D2 ≈0.26 at τ=15) is quoted in the text. The dummies bypass the MAD detector by construction.

Dummy (budget)	MCC@5	MCC@10	MCC@15 (± std)	95% seed interval	MCC@30
D1 uniform (n_{gt})	0.070	0.130	0.183 ± 0.025	[0.137, 0.231]	0.298
D2 prior-sampled (n_{gt})	0.073	0.134	0.189 ± 0.024	[0.148, 0.240]	0.308
D1 uniform (n=22, encoder-matched)	0.130	0.218	0.277 ± 0.012	[0.250, 0.303]	0.363
D2 prior-sampled (n=22, encoder-matched)	0.138	0.224	0.279 ± 0.014	[0.252, 0.302]	0.359

— numerically the level of the top-K-on-score oracle ceiling of Section 5.5.3 (18.8%). An oracle that reads the per-frame score and picks the $K = n_{\text{gt}}$ highest-scoring frames per video recovers no more GT boundaries than blind placement of the same number of predictions. This is the sharpest statement of the score-abstraction diagnosis available in this thesis: at the calibrated budget, the score's top frames localise boundaries no better than chance.

Third, the encoders also clear the random floors at the encoder-matched budget. At n=22 predictions per video the random dummies reach MCC@15 = 0.277 (D1) and 0.279 (D2), still well below the encoder band of 0.390–0.405 — a margin of 0.11–0.13. The same separation holds at every tolerance: at $\tau = 5$ the encoder models reach 0.255–0.265 against matched D1/D2 floors of 0.130/0.138 (+0.12–0.13, far outside the seed interval), at $\tau = 10$ they reach 0.362–0.364 against 0.218/0.224, and even at the widest $\tau = 30$ the floors (0.363/0.359) do not catch the encoder band. BL, by contrast, sits below its matched random floor at the headline tolerance (0.228 against D1/D2 ≈0.26 at n=13): the pose-only input does not support even chance-level boundary coverage once the budget is matched. *At every tolerance and at both the calibrated and the encoder-matched budgets, the encoders sit clearly above the perception-free random floors, so the headline MCC@15 reflects genuine above-chance detection.*

The random chance floors do not contradict the diagnostic narrative of this section — they bound it from below. The encoders sit clearly above every perception-free floor (D1, D2) at every tolerance and budget, so the headline MCC@15 is genuine above-chance detection, not an artefact of dense prediction placement. The noise-floor audit below nonetheless shows that 77–79% of all predictions land at no-near-GT frames, and the top-K-on-score ceiling caps recoverable recall at 18.8%: the encoders beat chance, but the unsupervised peak-detection readout still leaves most of the available boundary signal on the table. The cross-corpus observation that the per-corpus ceiling tracks the test corpus rather than the encoder (Section 5.2.1) also follows directly: the random floor is set by corpus statistics, tolerance and prediction budget, all of which belong to the test side of the protocol.

Breakfast floors. Repeating the construction on the 68-video Breakfast Coarse V2.2 test split (identical D1 uniform / D2 train-histogram dummies, 100 seeds; `docs/experiments/dummy_chance_floors_breakfast_2026-06/`) gives higher floors than segment-C, as Breakfast's denser per-video boundary statistics predict: MCC@15 = 0.33 (D1) and 0.37 (D2) at the calibrated budget, and 0.34–0.38 at the model-matched budgets (n = 16–19). The two reconstruction encoders clear these decisively even at the headline tolerance — RV-PD (0.593 in-domain, 0.577 transferred) and EST grid= 4 (0.514/0.522) sit 0.15–0.22 above their matched floors — whereas RV-PL (0.425/0.393) only reaches floor level, its transferred cell falling inside the matched floor's 95% interval ($[0.357, 0.405]$). The reconstruction encoders thus clear their perception-free floor on *both* corpora, and the floor's own rise from ≈ 0.28 (segment-C) to ≈ 0.37 (Breakfast) at the matched budget is exactly the corpus-statistics effect that fixes the per-corpus ceiling to the test side of the protocol (Section 5.2.1).

5.5.2 Noise floor: 77–79% of every model's predictions land at no-near-GT frames

For each model, every prediction on the test split was tagged by its distance to the nearest GT boundary. A `no-near-GT` prediction is one whose nearest GT boundary is more than $\tau = 15$ frames away — i.e. a prediction that cannot be a true positive at the headline tolerance. The `no-near-GT` count is the cardinal noise-floor diagnostic: these predictions fire at frames where no verb transition exists within the evaluation window. Table 5.5 reproduces the audit across the four paper-scoped models on the canonical (v2.1) score files. (An earlier audit on the original-extraction score files used a wider $2\tau = 30$-frame exclusion radius and produced rates of 61–65%; the canonical test-split audit at the ±15f radius reported here supersedes it.)

Table 5.5: Noise floor audit per model on the segment-C test split (canonical diagnostic, v2.1 score files). `no-near-GT` predictions fire at frames more than 15 frames away from any GT boundary, i.e. they can never be true positives at the headline tolerance. The % column is essentially constant across models.

Model	Total preds	No-near-GT FPs	%
BL	731	567	77.6%
RV-PD	1234	975	79.0%
EST grid= 4	1332	1046	78.5%
RV-PL k=3	1189	940	79.1%

The noise rate is roughly constant at 77–79% across all four paper-scoped models, two encoder families (the pose-MLP BL pipeline; the ViT-based RV-PD / EST / RV-PL family) and three score paradigms (reconstruction error, prediction error, gated multiplicative). *Every model emits roughly* 3.5× *more no-near-GT predictions than near-GT*

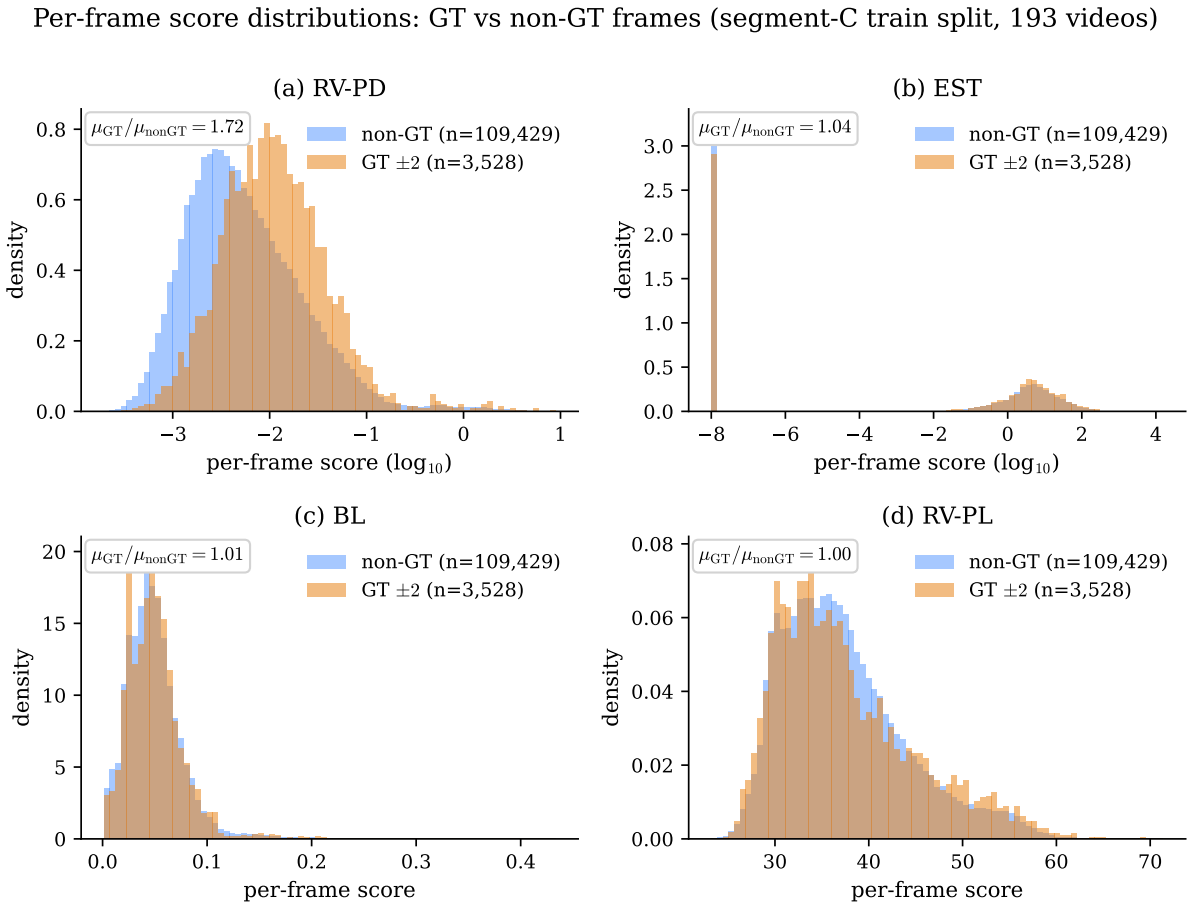

Figure 5.1: Per-frame score distributions at GT frames (orange) vs all other frames (blue) across the 193-video train split. The overlap of the distributions is the mechanism behind the noise floor measured on the test split: 77–79% of model predictions fire at no-near-GT frames because the noise tail of the non-GT distribution exceeds the boundary peaks of the GT distribution.

predictions. This invariance is the load-bearing observation: the noise floor is a property of the unsupervised peak-detection *paradigm* on this corpus, not a property of any one model or any one detector setting.

The Phase A negative result becomes mechanistic in the light of Table 5.5. Score-side post-processing cannot remove a noise spike that locally looks identical to a signal spike, because the MAD detector applies a single per-video threshold and the noise sits at the same local scale as the signal. Ensemble fusion across models cannot help either, because each model's noise spikes fire at *different* frames – summing the scores stacks every model's noise on top of every other model's noise, rather than cancelling out. The window-size sweep is flat for the same reason: shrinking W lets more noise spikes through (each was already above median+3·MAD) and the precision loss dominates the recall gain.

The per-verb F1 numbers in Section 5.5.5 below sit much higher than the global MCC because their denominator only counts predictions tagged with a verb (FPs near a GT). The MCC denominator counts all FPs, including the 975 (RV-PD) or 1 046 (EST grid=4) no-near-GT predictions. *The verb-level signal extraction is working at F1* ≈ 0.65*; what kills the global MCC is the noise rate.*

5.5.3 Top-K-on-score recall ceiling: below 20% on the test split

The noise-floor audit characterises what the MAD detector does. The top-K-on-score analysis characterises what any detector could in principle do given access only to the per-frame score. For each test video we sorted frames by the model's per-frame score and took the K highest-scoring frames, where K = number of true GT boundaries in that video. By construction this is the strict upper bound on what any threshold-based detector consuming only the per-frame score can recover. We then asked how many of those top-K frames land within ±15 frames of a real GT boundary, under the same greedy one-to-one matching as the evaluator. Table 5.6 reports the resulting oracle ceiling per model on the canonical test split; Table 5.7 reproduces the score-magnitude decomposition from the original train-split instance of the same analysis.

Table 5.6: Top-K-on-score recall ceiling per model on the segment-C test split (canonical diagnostic; $K = n_{\text{gt}}$ per video, ±15f greedy matching). RV-PL k=3 exposes only its peak-detected boundaries (no per-frame score), so its ceiling is not computable. Even the oracle that knows the true number of boundaries per video recovers fewer than one in five.

Model	Top-K recall@15
BL	6.0%
RV-PD	15.8%
EST grid= 4	**18.8%**

Three facts in Table 5.6 and Table 5.7 are decisive. First, the top-K-on-score recall on the segment-C test split caps at 18.8% (EST grid= 4, the leader of the encoder-based

Table 5.7: Score-magnitude decomposition of the top-K analysis (RV-PD, 193-video train split, original-extraction score file; top-K recall 211/708 $\approx$ 30% in this configuration). The missed top-K frames have a higher mean score than the matched ones, indicating the score distribution at non-boundary frames has a heavier tail than at boundary frames.

Set	n	**mean RV-PD score**
Frames at GT ± 2	3 528	0.0450
Frames at non-GT	106 872	0.0257
Top-K predictions that HIT a GT	**211**	**0.6903**
Top-K predictions that MISS	**497**	**0.7662**

models; RV-PD 15.8%, BL 6.0%) – so the absolute ceiling on what any detector that observes only the score can recover is fewer than one in five GT boundaries. Second, the GT-frame mean score is roughly 1.75$\times$ the non-GT-frame mean (Table 5.7): there is real signal in the per-frame score magnitude. Third, and most diagnostic, *the missed top-K predictions have a higher mean score than the matched ones.* The noise distribution is not just present in the score; it has a heavier tail than the signal distribution. Per video, the noisiest frames routinely beat the boundary frames in score. (The train-split RV-PD configuration of Table 5.7 caps at $\approx$ 30%; the canonical test-split ceilings of Table 5.6 are stricter, and the heavier-tail structure is the same in both.)

This places the noise-floor finding of Section 5.5.2 in a strictly stronger frame. The 77–79% FP rate is not an artefact of MAD's specific threshold rule – it is a property of the score signal itself. Even a supervised classifier trained directly on (score, GT) pairs cannot beat the top-K ceiling, because the relevant information is no longer in the signal. The journal records one such attempt (a 1-D CNN classifier on the per-frame score with positive-class weighting), which collapsed to a near-uniform output and regressed RV-PD's MCC@15 by 0.180; the diagnosis above explains why no fix to the CNN recipe would help. The information lives upstream of the per-frame score, in the encoder embedding.

5.5.4 Training-free TSM novelty baseline: the 2-D readout must be learned

Chapter 6 argues that the forward direction for this pipeline is to score boundaries on the $T \times T$ temporal self-similarity matrix (TSM) rather than on a per-frame scalar. The cheapest instance of that direction is training-free, and it was run as a sanity check of the claim: for each test video, build the cosine self-similarity matrix of the frame representations, score every frame with a Foote-style Gaussian-tapered checkerboard novelty kernel convolved along the diagonal (Foote, 2000), and feed the resulting 1-D novelty signal to the same shared MAD detector and the same evaluator as every model in this chapter (nothing is fitted on any split, and the kernel half-widths $w \in \{8, 15, 30\}$ are fixed a priori). Three input spaces were tested: the frozen RV-PD embeddings of both the original

and the v2.1 extraction, and the raw EST features spatially pooled to a 4×4 grid (240-D per frame). Table 5.8 reports the best width ($w = 8$; the larger widths are uniformly lower).

Table 5.8: Training-free TSM novelty baseline on the segment-C test split (kernel half-width $w = 8$). Each variant is read against the encoder band at MCC@15 (0.390–0.405) and against the perception-free random chance floors of Section 5.5.1 (D1 uniform / D2 train-histogram): MCC@15 = 0.183/0.189 at the calibrated budget and 0.277/0.279 at the encoder-matched n=22 budget. No per-TSM-budget D1 floor is tabulated, so the comparison uses the calibrated and encoder-matched random floors rather than a budget-matched per-variant floor.

TSM input	$\bar{n}_{\text{pred}}$	**MCC@5**	**MCC@15**
RV-PD embeddings (orig.)	14	0.108	0.279
RV-PD embeddings (v2.1)	16	0.098	0.313
Raw EST features (4×4 pool)	9	0.176	0.331

The outcome is decisive for the forward-direction question, in two parts. First, *the hand-crafted 2-D readout does not approach the encoder band*: the best variant reaches MCC@15 = 0.331 against the band's 0.390–0.405, and the embedding-based variants (0.279/0.313) fall short of it by an even wider margin — cosine similarity over the reconstruction embeddings carries no boundary structure that closes the gap to the learned encoders at this scale. The FlowGEBD existence proof (training-free F1@0.05 = 0.71 on Kinetics-GEBD (Zheng et al., 2024)) does not transfer to segment-C's verb granularity and evaluation regime. Second, and more constructive: all three TSM variants clear the calibrated random floor (0.189 at $\tau = 15$, 0.073 at $\tau = 5$), but the *raw-feature* TSM (0.331 at $\tau = 15$, 0.176 at $\tau = 5$) carries more boundary structure than the reconstruction embeddings computed *from those same features* (0.279/0.313 at $\tau = 15$) — the embedding-vs-raw gap is the diagnostic point. The reconstruction objective evidently does not organise the embedding space so that within-event frames are more cosine-similar than across-event frames, which is precisely the property the geometry-shaping objectives of the recent literature (BoCo, VICReg) train for. The forward direction of Chapter 6 therefore survives this test in sharpened form: the $T \times T$ readout must be *learned*, and the first thing it must learn is the similarity geometry itself.

5.5.5 Per-verb breakdown rules out class-imbalance artefacts

A natural concern is that MCC@15 on Assembly verb GT might be carried by the most common verb — `attach`, 40.2% of the test boundaries. The per-verb breakdown rules this out: detection rates are roughly uniform across the seven non-trivial verbs (`unscrew` and `demonstrate` have only 6 test instances each), and `screw` — not `attach` — is the verb with the highest per-verb recall in every Phase 1 score file. The per-verb F1 spread is tightest for EST grid=4 (0.068) and moderate for RV-PD (0.130); the wider spreads for

BL (0.500) and RV-PL k=3 (0.436) are driven by the two rare verbs. Macro recall tracks micro MCC and every verb is detected at a comparable rate, so the global MCC@15 is a faithful summary of multi-verb performance, not an artefact of the test-set distribution.

5.6 Statistical analysis and alternate framings

The point-estimate MCC@15 numbers reported above are mean values across the 54 evaluable test videos. This section adds (i) bootstrap 95% confidence intervals on the per-video distribution, (ii) paired Wilcoxon signed-rank tests with Bonferroni correction across all six paper-scoped model pairs, and (iii) two alternate metric framings (micro-MCC pooled across videos; macro-MCC averaged across verbs) to expose how robust the ranking is to the choice of aggregation. The headline v2.1 per-video bootstrap CIs (10,000 resamples, seed 12345; `docs/experiments/statistical_analysis_v2_1/`) are, at MCC@15: RV-PD 0.405 $[0.369, 0.440]$, EST grid= 4 0.390 $[0.359, 0.422]$, RV-PL k=3 0.228 $[0.175, 0.283]$, BL 0.228 $[0.177, 0.278]$. The two reconstruction encoders are separated from the RV-PL/BL pair by a clear margin, while RV-PL and BL overlap almost completely. The perception-free random floors of Section 5.5.1 lie well below the two encoder intervals at every budget: the encoder-matched random floor ($\approx$ 0.28, D1/D2 at n=22) sits below the lower bound of both reconstruction encoders' MCC@15 CIs, and the calibrated floor (0.18–0.19) further still. The Wilcoxon, micro/macro and per-verb tables below are computed on the original-extraction score files (audited in Appendix B); their conclusions transfer to the v2.1 headline for RV-PD, EST and BL, with the single exception of RV-PL, whose v2.1 collapse changes its position.[3]

Bootstrap confidence intervals. Table 5.9 reports 95% CIs from 10,000 bootstrap resamples (seed = 12345) of the per-video values for each of the four paper-scoped models.

Table 5.9: Mean precision, recall, F1, and MCC at tolerance 15f with bootstrap 95% confidence intervals. The CI widths quantify the per-video variance underlying each headline number.

Model	Precision@15	Recall@15	F1@15	MCC@15
BL	0.204 [0.154, 0.257]	0.316 [0.238, 0.394]	0.215 [0.168, 0.262]	0.228 [0.178, 0.279]
EST	0.167 [0.145, 0.192]	0.886 [0.844, 0.925]	0.270 [0.239, 0.303]	0.362 [0.333, 0.390]
RV-PD	0.217 [0.186, 0.250]	0.853 [0.796, 0.906]	0.329 [0.290, 0.369]	0.406 [0.369, 0.443]
RV-PL	0.223 [0.188, 0.260]	0.597 [0.517, 0.675]	0.295 [0.255, 0.335]	0.336 [0.294, 0.376]

The CIs are tight enough that the rank ordering is stable, but the overlap between

[3]The original-extraction MCC@15 means underlying Table 5.9, Table 5.10 and the per-verb breakdown are EST grid= 4 0.362, RV-PD 0.406, RV-PL k=3 0.336. On that generation RV-PL joins the ViT cluster and separates significantly from BL; its v2.1 retrain collapses to 0.228 (Table 5.1), tying BL, so under the headline generation RV-PL is no longer separable from BL — consistent with the v2.1 bootstrap CIs above ($[0.175, 0.283]$ versus BL $[0.177, 0.278]$). The RV-PD, EST and BL conclusions are unaffected.

EST (0.362 [0.333, 0.390]) and RV-PD (0.406 [0.369, 0.443]) is substantial. The next paragraph confirms this with a paired test.

Pairwise significance. Table 5.10 reports the Wilcoxon signed-rank statistic on per-video MCC@15 for the six pairwise model comparisons, with Bonferroni-corrected p-values (multiplied by 6).

Table 5.10: Paired Wilcoxon signed-rank tests on per-video MCC@15 across the 54-video test split, with Bonferroni correction ($k = 6$).

Model A	Model B	n	W	p_{raw}	p_{bonf}	sig. ($p < 0.05$)
BL	EST	54	244.0	1.769e-05	0.0001061	✓
BL	RV-PD	54	142.0	3.833e-07	2.3e-06	✓
BL	RV-PL	54	307.0	0.0002988	0.001793	✓
EST	RV-PD	54	388.0	0.006122	0.03673	✓
EST	RV-PL	54	617.0	0.2799	1	–
RV-PD	RV-PL	54	451.0	0.01208	0.07246	–

All three pairwise comparisons against BL survive Bonferroni correction. EST versus RV-PD also survives ($p_{\text{bonf}} = 0.037$). The other two pairs — EST versus RV-PL ($p_{\text{bonf}} = 1.0$) and RV-PD versus RV-PL ($p_{\text{bonf}} = 0.072$) — do not. Reading the ranking as "RV-PD > EST > RV-PL > BL" is correct on the original-extraction mean, on which the three ViT models form a statistically indistinguishable top cluster on per-video MCC at $p < 0.05$ except for the EST–RV-PD comparison, and only BL is significantly worse than each ViT model individually. Two of these conclusions carry to the v2.1 headline: RV-PD and EST stay significantly above BL, and the RV-PD–EST pair stays the only separable comparison among the encoders. RV-PL's does not — its v2.1 retrain collapses to 0.228 (Table 5.1), tying BL, so on the headline generation RV-PL drops out of the ViT cluster and joins BL at the bottom (its v2.1 CI $[0.175, 0.283]$ overlaps BL's $[0.177, 0.278]$ almost entirely).

Alternate framings. On the original extraction the MCC@15 ranking RV-PD > EST > RV-PL > BL does not hold under F1: at F1@τ for any $\tau \in \{5, 10, 15\}$ EST and RV-PL swap (RV-PD > RV-PL > EST > BL), because EST's very high recall (0.886) with low precision (0.167) inflates MCC@15 relative to F1 while RV-PL's more balanced trade-off favours F1 — the reason MCC and F1 are reported together throughout. (Under the v2.1 headline this swap disappears, as RV-PL's collapse drops its F1@15 to 0.221 below EST's 0.298; the general point that MCC and F1 can reorder mid-table models is why both are tabulated.) Alternate aggregations change the headline numbers but not the rank: macro-verb MCC (restricting evaluation to one verb at a time, then averaging) is substantially higher (mean 0.593 across the paper-scoped models) than either per-video aggregation, re-validating that the per-frame noise floor inflates the per-video MCC denominator —

restricting the pool to one verb shrinks the non-boundary count and makes the detection signal more visible. The macro-verb framing is more flattering but less interpretable for cross-corpus comparison; the headline numbers stay on mean-of-per-video MCC@15.

5.7 Calibration, agreement and failure modes

Four further diagnostics complete the negative-result picture. *(i) Temporal calibration:* on the signed offset $\Delta = \text{pred} - \text{GT}$ for matched pairs within ±15f, three of the four models are systematically early by 2–3 frames (median offsets BL −3, EST grid=4 −3, RV-PL k=3 −2); only RV-PD is well-calibrated (median 0). The left-of-peak tail of the Hanning smoothing kernel ($k = 5$) in the MAD detector explains the universal slight early-bias; RV-PD's sharper symmetric peaks survive it. The bias is forgiven at ±15f but partly drives the strict-tolerance ranking flip above. *(ii) Cross-model agreement* is a weak signal: going from agreement-1 to agreement-4 across the four models doubles precision (0.115 → 0.277) but collapses recall (0.98 → 0.18), with best F1 at $k = 2$ (0.342, marginally above $k = 3$'s 0.331); even at full agreement roughly 72% of agreed frames are still no-near-GT, so the noise floor is correlated across models rather than orthogonal — formally killing the ensemble-by-intersection hypothesis. *(iii) Verb confusion:* all four models over-fire toward `attach` (its 40.2% dominance plus the nearest-GT tagging of spurious peaks inside attach segments), while the diagonals confirm `screw` is the easiest verb (RV-PD recall 0.94, EST grid=4 0.96). *(iv) Failure modes:* RV-PD and EST grid=4 have no catastrophic-failure videos (MCC≤ 0.1), their adaptive thresholds guaranteeing at least one near-GT match; BL fails catastrophically on 17 of 54 videos (31%), its pose-only score flatlining on a third of the test set; RV-PL k=3 has 3 such failures from predictive misalignment beyond the tolerance window. Over-firing is universal — between 22 and 36 of 54 videos cross the 80%-FP threshold for every ViT model, reflecting the noise floor of Section 5.5.2.

5.8 Summary of results and headline ranking

The headline ranking at MCC@15 (v2.1 generation) is RV-PD (0.405) $\gtrsim$ EST grid= 4 (0.390) > RV-PL k=3 (0.228) = BL (0.228). The two reconstruction-based encoders end within 0.015 of each other. RV-PL trails the reconstruction pair by $\approx$ 0.18 and falls level with the pose-only baseline: its predictive head is sensitive both to its training corpus' temporal cadence and to the sharpness of its input features, so under the cleaner v2.1 channels the t+3 task becomes easy again and the prediction-error signal flattens (Section 5.4.1, Section 5.2.1) — a fragility the reconstruction objectives do not show. BL is at its pose-feature ceiling and unrecoverable inside the unsupervised paradigm.

Read as a whole, the chapter establishes an empirical picture that is well-understood at every level. At inference time, 13 score-side configurations failed to lift any model past

Table 5.11: Each model on Assembly segment-C at MCC@15 across three configurations: "Raw" (Phase 1 MAD detector on the original-extraction score), "Tier 2" (architectural retrain on the original extraction, Section 5.4), and the v2.1 retrain that supplies the headline numbers of this chapter. The Raw and Tier 2 columns are original-extraction (audited in Appendix B) and are shown only for the per-model bridge. **Bold** marks each model's headline cell — the v2.1 retrain, except for BL, whose pose input is independent of the EST-feature extraction (so it has no Tier-2 retrain or v2.1 variant, and its Raw entry is the headline). For RV-PD and RV-PL the headline is *not* the row maximum: the v2.1 re-extraction left RV-PD unchanged within noise and collapsed RV-PL from its 0.336 Tier-2 value to 0.228 (Section 5.4), tying BL.

Model	Raw	Tier 2	v2.1 (headline)
BL	**0.228**	—	—
EST	0.339	0.362	**0.390**
RV-PD	0.406	0.404	**0.405**
RV-PL	0.323	0.336	**0.228**

its raw baseline. At training time, the three Tier 2 architectural retrains moved EST and RV-PL by small but mechanistic amounts. The diagnostic in Section 5.5 explains the picture from both sides: 77–79% of every model's predictions land at frames with no GT in any tolerance window; the top-K-on-score ceiling on per-frame readouts is approximately 19% regardless of the detector — the level of blind placement at the same budget; and the perception-free random chance floors (D1 uniform, D2 train-histogram) show that the encoders sit clearly above chance at every tolerance (matched floors ≈ 0.28 and calibrated floors 0.18–0.19, against an encoder band of 0.390–0.405), so the headline MCC@15 is genuine above-chance detection even though the unsupervised readout still leaves most of the available signal unrecovered. The information bottleneck sits between the encoder embedding and the per-frame scalar score, not in the detector that consumes the score.

This diagnosis is the bridge to Chapter 6. The convergent direction of the recent GEBD literature — score boundaries on a 2-D temporal self-similarity manifold rather than on a 1-D per-frame scalar — is the natural way to break the noise floor identified here, because the boundary signal lives in the $T \times T$ block structure rather than in any single row or column. Translating that direction into a concrete model on the segment-C corpus is set out as future work in Chapter 6.

Chapter 6

Conclusion and Future Work

This thesis began as an exercise in operationalising Event Segmentation Theory (Zacks et al., 2007) for unsupervised Generic Event Boundary Detection. It ends as a documented *negative result* on the score-side of the unsupervised peak-detection paradigm, a per-frame-score *noise-floor diagnostic* that explains why that result is structural rather than incidental, and a *forward direction* read off the diagnostic and the convergent pattern of the recent GEBD literature. The headline segment-C numbers (RV-PD MCC@15 = 0.405, EST grid= 4 = 0.390, RV-PL k=3 = 0.228 tying BL = 0.228; Chapter 5) are not the contribution; the explanation of what they mean, and the principled path beyond them, is. This closing chapter consolidates the contributions, names the limitations, and sets out the open questions for the work that should follow.

6.1 Summary of Contributions (Revised)

The contributions are eight items: the first two are the substance of the CHItaly 2025 paper (Mirghaed et al., 2025), the third a pre-thesis development extension, and the remaining five new to this thesis.

1. **The EST-aligned feature representation** (Section 3.2): a 15-channel feature image (audited per channel in Appendix B) operationalising Event Segmentation Theory's perceptual categories of change (Zacks et al., 2007, 2009) through off-the-shelf detectors (YOLO (Jocher et al., 2022) and RAFT (Teed and Deng, 2020); earlier iterations used RTMDet (Lyu et al., 2022) and RTMPose (Jiang et al., 2023) before consolidating on YOLO, Section 3.2); the 15-channel design fixes three concrete extraction problems (dead optical flow, over-sparse keypoint splats, broadcast-scalar centroids) and is dataset-agnostic by construction.

2. **The CHItaly 2025 published model RVpd** (Mirghaed et al., 2025): a Simple-ViT with Nyström attention (Xiong et al., 2021) trained by reconstruction, with peak detection on a fused MSE-plus-velocity signal; the paper reports MCC = 0.55

on a 62-video Assembly101 subset (coarse GT, τ=15). RV-PL and BL are *not* part of the published paper.

3. **A pre-thesis experimental programme** (Section 5.1): two further models — RV-PL (a causal GRU predictor (Cho et al., 2014) on frozen RVpd embeddings) and BL (a per-video online pose-only MLP reproducing Basgol et al. (2024)) — evaluated across Assembly101 and Breakfast (Kuehne et al., 2014; Sener et al., 2022) at fine/coarse granularity, $\tau \in \{5, 10, 15\}$, with cross-dataset transfer and 9-fold cross-validation, serving as historical baselines.

4. **The segment-C corpus and unified MAD detector** (Section 3.3, Section 5.2): a re-scoped Assembly101 evaluation (250 videos × 600-frame windows, 199 verb–noun pairs collapsed to nine verbs, person-disjoint 80/20 split) with a single shared MAD detector (window = 15, $k_{\text{MAD}} = 3$) replacing the per-model recipes, giving a clean comparison substrate.

5. **A Phase A inference-side study** (Section 5.3): thirteen score-side, ensemble and MAD k-sweep configurations, of which *zero* lifted MCC@15 — the negative result that motivated the diagnostic.

6. **Tier-2 architectural improvements** (Section 5.4): the EST grid= 4 reparameterisation (a 4×4 spatial signature plus 1×1 projection replacing global average pooling) recovers a spatially localised signal that global average pooling had discarded and carries cleanly into the headline generation (MCC@15 = 0.390 after v2.1 re-extraction). The RV-PL k=3 horizon shift (predict t+3 not t+1) hardens a too-easy task and helped on the original extraction, but the same mechanism makes the head fragile: under the cleaner v2.1 channels the task becomes easy again and the headline retrain collapses to 0.228 (Section 5.4). Both effects are mechanism-confirmed.

7. **The central diagnostic finding** (Section 5.5): the per-frame score carries real boundary information — ground-truth frames average 1.75× the non-boundary score — but the unsupervised peak-detection readout cannot extract it. 77–79% of every encoder model's predicted boundaries fire with no GT within ±15 frames (the per-frame-score *noise floor*), and the top-K-on-score oracle ($K = n_{\text{gt}}$) caps at 18.8% recall on the test split (EST grid= 4) because the score's noise has a heavier tail than its signal — under one boundary in five. Perception-free random chance floors (D1 uniform, D2 train-histogram) bound the *headline* numbers from below: at every tolerance and at both the calibrated and the encoder-matched budgets the encoders sit clearly above chance — matched random floors of MCC@15 ≈ 0.28 and calibrated floors of 0.18–0.19, against an encoder band of 0.390–0.405 — so

the headline MCC@15 reflects genuine above-chance detection. The binding constraint is therefore not an absence of signal but the per-frame-scalar readout, which the noise floor and the 18.8% top-K ceiling show discards most of the boundary information that is demonstrably present.

8. **Dual-corpus evaluation with the full cross-corpus matrix** (Section 5.2.1): Assembly segment-C and a Breakfast Coarse rebuild (268 videos, 16 participants, 48-action GT, person- and content-disjoint 200/68 split by participant, Section 4.1.2) under an identical protocol, with EST grid= 4, RV-PD and RV-PL k=3 each retrained once per corpus and evaluated on both test splits. The pure-encoder matrix is symmetric (RV-PD 0.405/0.577/0.410/0.593; EST grid= 4 0.390/0.522/0.400/0.514 across A→A/A→B/B→A/B→B, $|\Delta| \leq 0.016$); RV-PL k=3 breaks it (0.228/0.393/0.292/0.425), sitting below the encoder band everywhere (Section 5.4.1). The per-corpus ceiling tracks the test corpus, not the encoder — the structural prediction the noise-floor diagnostic makes.

6.2 Empirical Findings Synthesis

Three observations frame the thesis. First, the per-frame-scalar readout — not the encoder — is the binding constraint on this corpus: the score carries real signal (ground-truth frames average 1.75× the non-boundary score), but collapsing each frame to one scalar and peak-detecting it cannot separate that signal from a heavier-tailed noise distribution. Phase A's thirteen configurations all failed, the MAD detector is already at every model's per-video optimum, and the perception-free random floors below (the encoders clear D1/D2 at every tolerance, confirming above-chance detection) and the 18.8% top-K-on-score oracle ceiling above close the *scalar* route from both sides (Section 5.3, Section 5.5.3, Section 5.5.1). Second, the two Tier-2 gains are real but bounded by that abstraction: EST grid= 4 recovers a spatially localised signal — which hand is at which corner of the manipulated piece — that average-pooling had lost, while RV-PL k=3 re-introduces a boundary spike by hardening a too-easy prediction task, but the same mechanism makes its head fragile (under the cleaner v2.1 extraction the in-domain task becomes easy again and the signal collapses to 0.228; Table 5.1). Third, the literature has converged on a different primitive: UBoCo (Kang et al., 2022), DDM-Net (Tang et al., 2022), SC-Transformer++ (Li et al., 2022), EfficientGEBD (Lin et al., 2024), DyB-Det (Wang et al., 2024) and DiffGEBD (Wang et al., 2025) all score boundaries on a $T \times T$ similarity or difference manifold, and FlowGEBD (Zheng et al., 2024) reaches F1@0.05 = 0.71 on Kinetics-GEBD with zero training on such a manifold, whereas all five models here score per-frame. The path forward is therefore not another inference-side tweak but a structurally different scoring primitive.

6.3 Limitations

The thesis carries seven limitations that should be stated plainly.

Corpus size constraint. The segment-C corpus (193 train, 57 test videos at 600 frames) is small next to Kinetics-GEBD's (Shou et al., 2021) ∼54K videos; the cross-dataset transfer confirms generalisation and Section 5.4.1 attributes most apparent shortfalls to segment-length effects on the MCC denominator, but a richer encoder trained on more data could plausibly find more discriminative per-frame structure.

Verb-level GT abstraction. Collapsing Assembly's 199 verb–noun pairs to nine verbs (chosen for label-density and reliability, Chapter 3) sacrifices fine granularity, and a denser (verb, noun) target would likely amplify rather than attenuate the noise-floor diagnostic of Section 5.5.2.

MAD detector and MCC denominator effects. The TN-driven MCC inflates with segment length, complicating cross-corpus comparison (Section 5.2.2); the shared MAD detector at window = 15 is empirically optimal on segment-C, but its cross-corpus generalisation is a hypothesis, not a verified property.

No optical-flow channel in BL. The pose-only baseline uses only 17 COCO keypoints (34-D); a flow channel was never tested, and the gap between BL's segment-C MCC@15 = 0.228 and the unsupervised flow-only ceiling (FlowGEBD F1@0.05 = 0.71 (Zheng et al., 2024)) is consistent with that omission.

Hardware budget. All experiments ran on a single NVIDIA RTX 4090 (24 GB), which precluded larger encoders, longer schedules, multi-run ensembles, and the full $T \times T$ similarity tensor at $T = 600$ in batch sizes above four; the modelling choices reflect this envelope.

No inter-annotator agreement baseline available. Assembly101's coarse annotations have a single annotation per session, so inter-annotator variance cannot be estimated; the ±15-frame headline tolerance (3 s at 5 fps) serves as a conservative upper bound, but its calibration against human-vs-human variance remains open, and the systematic 2–3 frame early bias in three models (Section 5.7) is too small relative to this window to separate from annotator timing conventions.

Verb-level GT is materially incomplete on segment-C. A model-agreement triangulation on the test split found that of 3,077 no-near-GT predictions, 290 frames had *all four* architecturally-distinct models firing in agreement — 2.55× the ∼114 predicted by

an independent-FP null; visual inspection of 13 four-of-four clusters from 12 participants found 9 ($\approx 60\%$) to be clearly missed verb transitions. The headline 77–79% noise floor (Section 5.5.2) is therefore an upper bound on the true FP rate (lower bound $\approx$ 65–70%); the per-frame-score conclusion still holds but is joined by a secondary GT-design effect the thesis cannot disentangle without multi-annotator overlap data.

6.4 Future Work

The diagnosis and the convergent literature pattern point at the same forward direction. The noise-floor (22% precision ceiling), top-K-on-score (18.8% recall ceiling) and cross-corpus matrix findings (Section 5.5.2, Section 5.5.3, Section 5.2.1) all identify per-frame scoring as the wrong primitive on this corpus; every GEBD method to set a new state of the art since 2022 (Section 6.2) instead scores a two-dimensional similarity or difference manifold. On segment-C, the training-free instance of this primitive has been measured (Section 5.5.4): Foote-style checkerboard novelty on the frozen-embedding self-similarity matrix stays well below the encoder band, and even its best raw-feature variant reaches only MCC@15 = 0.331 against the encoder band's 0.390–0.405 — so here the 2-D readout must be *learned*, and the embedding-vs-raw gap shows the first thing to learn is the similarity geometry itself. The forward direction is therefore: *stop scoring frames; score the two-dimensional structure that contains the boundary structure intrinsically.* A $T \times T$ cosine self-similarity matrix has diagonal blocks for within-event sequences separated by lighter cross-event regions, and the corner where two blocks meet — the boundary — survives the per-frame noise that destroys a one-dimensional scalar.

Translating this into a model that beats the segment-C baselines is left open, but the scoped questions are concrete. The objective must be collapse-resistant — preliminary SimSiam- and DINO-style trials at batch size 8 (single-GPU memory) collapsed early — so the first is which collapse-resistant geometry-shaping objective (VICReg (Bardes et al., 2022), BoCo (Kang et al., 2022), or a combination) best decorrelates within- from across-segment similarity on the EST features; how the $T \times T$ readout should be designed so the block-corner is amplified rather than washed out; whether the multi-order difference channels of EfficientGEBD (Lin et al., 2024) and DyBDet (Wang et al., 2024) suffice at the input or require an additional 2-D head; and how to integrate the result with RV-PD's encoder without losing the cross-corpus stability of Section 5.2.1.

Adjacent open directions. Three further directions are worth recording. *Scaling to larger corpora*: Kinetics-GEBD ($\approx$54K videos) (Shou et al., 2021) and TAPOS would test whether the noise-floor diagnosis is corpus-specific or universal — the diagnosis predicts the latter. *Cross-modal features*: BL's pose-only ceiling reflects too-little wrist-trajectory information to separate *attach* from *detach*; hand-only keypoints, audio, or an optical-flow

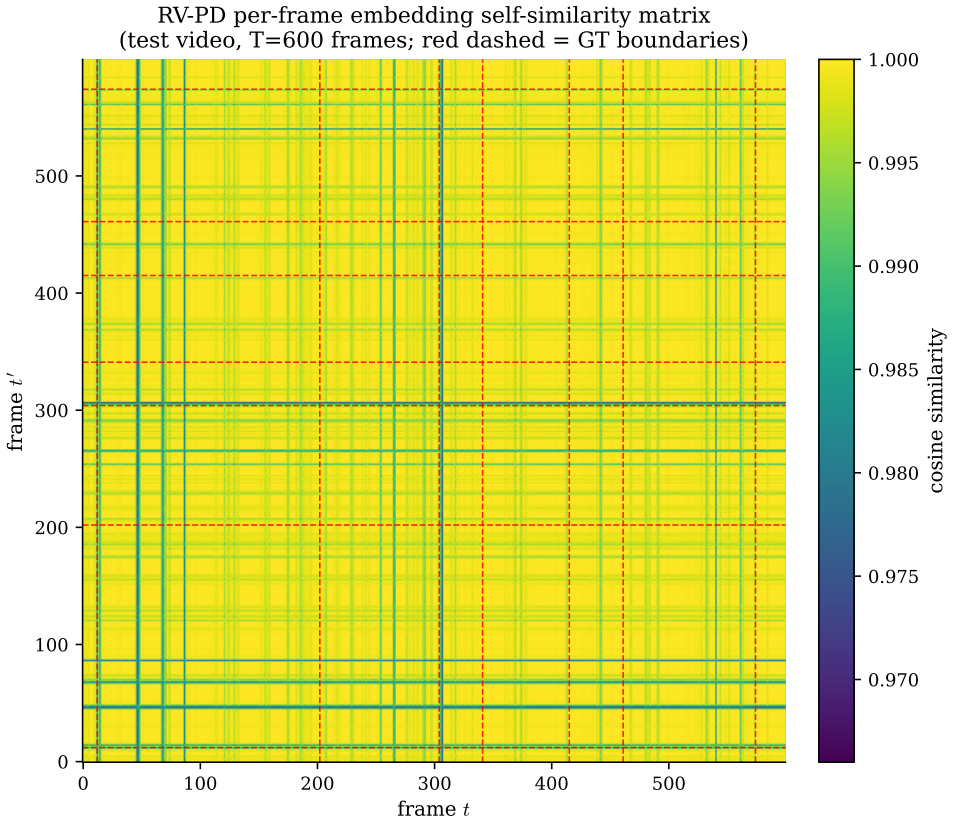

Figure 6.1: Per-frame embedding self-similarity matrix for one representative test video. Diagonal blocks correspond to within-event embedding similarity; off-diagonal corners mark verb boundaries (red dashed lines). The 2-D structure of these boundaries survives noise that destroys the 1-D per-frame score signal — the motivating observation for the future direction set out in this chapter.

channel (cf. FlowGEBD (Zheng et al., 2024)) would address it directly. *Linking back to cognitive theory*: Event Segmentation Theory posits seven categories of change (Zacks et al., 2007), of which the five observable ones are operationalised statically in Chapter 3 (causes and goals are not given dedicated channels); EST's further prediction that humans segment at nested timescales simultaneously suggests a hierarchical, multi-scale self-similarity matrix as the faithful unfinished extension.

Closing remark. The central reframing of this thesis is that an honest negative result, supported by a sharp diagnostic, can be a more valuable contribution than a marginal improvement to a headline number. The four paper-scoped models on segment-C have reached the ceiling of the per-frame-score unsupervised peak-detection paradigm; the diagnostic explains why; the recent GEBD literature suggests where the work should go next. Translating that direction into a model that beats the segment-C baselines of this thesis is the natural next step — and the empirical question it asks is the one a follow-up project should set out to answer.

Bibliography

Sathyanarayanan N. Aakur and Sudeep Sarkar. A perceptual prediction framework for self-supervised event segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1197–1206, 2019.

Yazan Abu Farha and Jurgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3579, 2019. doi: 10.1109/CVPR.2019.00369.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. V-JEPA: Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2412.10925*, 2024.

Hamit Basgol, Inci Ayhan, and Emre Ugur. Predictive event segmentation and representation with neural networks: A self-supervised model assessed by psychological experiments. *Cognitive Systems Research*, 83:101167, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

Eleonora Ceccaldi. *CEST: A Cognitive Event Based Semi-automatic Technique for Behavior Segmentation*. Phd thesis, Università degli Studi di Genova, 2021. URL https://hdl.handle.net/20.500.14242/63962. XXXIII Ciclo, Informatica e Ingegneria dei Sistemi (Computer Science). URN:NBN:IT:UNIGE-63962.

Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. ShotCoL: Shot contrastive self-supervised learning for video scene detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting (TS-TCC). In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370. Springer Berlin Heidelberg, 2003.

Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 452–455, 2000.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. McMillan, 2004.

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

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399*, 2023.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, et al. ultralytics/yolov5: v7.0 — YOLOv5 SOTA realtime instance segmentation, 2022.

Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. UBoCo: Unsupervised boundary contrastive learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–787, 2014.

Congcong Li, Xinyao Yuan, Chi-Man Pun, Mei Yu, Xinyang Guan, Zhuo Chen, et al. Structured context transformer for generic event boundary detection. In *Proceedings of the CVPR Workshops on Long-form Video Understanding*, 2022. Winner of the CVPR 2022 LOVEU GEBD Challenge.

Yuxin Lin et al. EfficientGEBD: Efficient generic event boundary detection via difference mixing. *arXiv preprint arXiv:2407.12622*, 2024.

Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. RTMDet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022.

Amir Arsalan Serajoddin Mirghaed, Nadia Benini, Gualtiero Volpe, and Giovanna Varni. An EST-based generic event boundary detector. In *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025)*. ACM, 2025. doi: 10.1145/3750069.3755958.

Ramy Mounir, Sathyanarayanan Aakur, and Sudeep Sarkar. Chapter 12 — self-supervised temporal event segmentation inspired by cognitive theories. In E.R. Davies and

Matthew A. Turk, editors, *Advanced Methods and Deep Learning in Computer Vision*, pages 405–448. Academic Press, 2022.

Izzeddin Sayed, Isma Hadji, Bashir Khanloo, et al. Temporal DINO: A self-supervised video strategy to enhance action prediction. *arXiv preprint arXiv:2308.04589*, 2023.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21064–21074, 2022. doi: 10.1109/CVPR52688.2022.02042.

Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8075–8084, 2021.

Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, and Limin Wang. DDM-Net: Progressive attention-based feature modulation for generic event boundary detection. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2022.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419. Springer, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Jiaqi Wang et al. DyBDet: Dynamic boundary detection with multi-exit subnet routing and multi-order differences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Yuhang Wang et al. DiffGEBD: Denoising diffusion for generic event boundary detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, 2021.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148, 2021.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. ASFormer: Transformer for action segmentation. In *Proceedings of the 32nd British Machine Vision Conference (BMVC)*, 2021. doi: 10.5244/C.35.49.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Jeffrey M. Zacks and Khena M. Swallow. Event segmentation. *Current Directions in Psychological Science*, 16(2):80–84, 2007.

Jeffrey M. Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127(1):3–21, 2001.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2): 273–293, 2007.

Jeffrey M. Zacks, Nicole K. Speer, and Jeremy R. Reynolds. Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, 138(2):307–327, 2009.

Yuxin Zheng et al. FlowGEBD: Unsupervised generic event boundary detection via optical-flow normalised similarity. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Appendix A

Feature Channel Analysis

This appendix characterises the 15-channel EST feature tensor produced by the pipeline of Section 3.2 and reports the model-side outcomes of training the three encoder-based paper-scoped models on those features. Together the two halves underpin the diagnostic argument of Chapter 5: the binding constraint on this corpus is not feature quality or encoder architecture but the per-frame-score abstraction shared by every paper-scoped model.

A.1 Per-channel audit summary

Static inspection recovers four structural channel classes: channels 0–2 are *continuous dense* per-pixel fields (frame-difference energy, optical flow x/y) carrying dense per-pixel spatial structure; channels 3–8 are *sparse keypoint splats* (Gaussian-RBF, $\sigma = 4\,\text{px}$, $\approx 87\,\%$ zeros); channels 9–11 are *centroid heatmaps* (a Gaussian splat at the person centroid modulated by displacement magnitude, $\approx 75\,\%$ zeros); channels 12–14 are *binary masks*. A correlation analysis gives only $k_{95} = 12$ effective degrees of freedom (strongest redundancy: keypoint velocity vs acceleration $r = +0.92$; the two centroid components $r = +0.78$), ruling out channel redundancy as the cause of the noise floor of Section 5.5.2.

The decisive measurement is per-channel boundary discriminability on the train split: the single best raw channel attains a *top-K-on-channel* recall of 0.329, only marginally above the corresponding train-split model-side top-K-on-score ceiling of ≈ 0.30 for RV-PD (the stricter canonical test-split per-model ceilings – 18.8% for EST grid= 4, 15.8% for RV-PD – are in Section 5.5.3) (best absolute-signal discriminator: keypoint velocity, AUC = 0.673; best temporal-difference: optical flow x, AUC = 0.618). No single raw channel materially exceeds this band. Hence the models' recall ceilings are set by what is present in the input, and no per-frame-score detector on this tensor – regardless of architecture or training – can beat the model-side ceiling: the strongest single result in support of the noise-floor diagnosis of Chapter 5.

A.2 Model-side outcomes on the features

Each of the three encoder-based paper-scoped models (RV-PD, RV-PL k=3, EST grid= 4) was trained from scratch on the features using its standard configuration and the same MAD detector (window = 15, $k_{\text{MAD}} = 3$). The pose-only baseline BL does not consume the 15-channel tensor and is shown for context. Evaluation is on the segment-C test split (54 videos with GT in scope), with the canonical evaluator computing precision, recall, F1 and MCC; the headline metric is MCC@15.

Table A.1: Cross-model summary at tolerance 15 frames (segment-C test split, v2.1 retrains, unified MAD detector). The two reconstruction-based encoders cluster at 0.390–0.405; the v2.1 RV-PL retrain collapses to 0.228 in-domain (0.336 on the original extraction; see Section 5.4); the pose-only BL sits at 0.228 because its input modality (8 reliable joints) caps recall at 0.316. Balanced accuracy ($\frac{1}{2}$(recall+specificity)) and event-level IoU (TP/(TP+FP+FN)) are reported for completeness.

Model	P@15	R@15	F1@15	MCC@15	BalAcc@15	IoU@15
BL (pose only)	0.204	0.316	0.215	0.228	0.653	0.132
RV-PD	0.215	0.857	0.327	**0.405**	0.917	**0.205**
RV-PL k=3	0.226	0.271	0.221	0.228	0.632	0.139
EST grid= 4	0.187	0.916	0.298	0.390	**0.944**	0.183

Three observations follow. First, despite materially different model families – a single-pass SimpleViT autoencoder and a spatial-grid-equipped GRU autoencoder – the two reconstruction models differ by less than 1.5 percentage points of MCC@15, and the predictive variant (RV-PL k=3, 0.228 under v2.1; 0.336 on the original extraction) does not reach them: architecture choice does not break through the $\sim$ 0.40 ceiling on this corpus. Second, the reconstruction models land at recall in the 0.86–0.92 band but precision capped between 0.19 and 0.22 – roughly four of every five predictions fall at frames with no GT boundary within ±15 frames, the same FP-dominated noise floor of Table 5.5 viewed model-by-model (the v2.1 RV-PL is the opposite exception, under-firing at recall 0.271 because its flattened prediction-error signal rarely crosses the MAD threshold). Third, BL is structurally separate: at MCC@15 = 0.228 and recall 0.316 it operates only on eight reliable joints with no access to the dense channels the encoder-based models read, and is best read as a lower bound on pose-only boundary detection rather than a comparison-axis competitor. Together these facts pin the binding constraint on the per-frame-score abstraction – the one design choice common to all – rather than on encoder architecture or input feature quality, motivating the structural next step of Section 6.4: stop scoring frames, and start scoring on the two-dimensional temporal self-similarity manifold.

A.3 Breakfast canonical retraining: training schedule and checkpoints

This section records the training-side details of the Breakfast cells of the cross-corpus matrix reported in Section 5.2.1, so that the in-domain B→B and cross-corpus B→A results are reproducible end-to-end from the same code tree.

Substrate. Breakfast features of shape $(15, 135, 240)$ are extracted at 5 fps on the EST feature pipeline (Section 3.2) into a single source HDF5 of 268 videos spanning 16 participants. The matrix of Chapter 5 uses a person- and content-disjoint substrate, materialised uncompressed: a by-participant (hence content-disjoint) split into 200 train videos (≈145 k frames) and 68 test videos (≈39 k frames), keeping the source keys end-to-end and asserted person- and content-disjoint at build time. GT is regenerated from the official `segmentation_coarse` files with the documented 15→5 fps convention (48 coarse action labels).

Per-model retraining. All three encoder-based models are retrained from scratch on the Breakfast train HDF5 with the same drivers and recipes used on Assembly segment-C. EST grid= 4 (`est_gebd_model/train.py`; SmoothL1 loss, AdamW at 10^{-4}, Jaccard-stability early-stop) converged in eight epochs, the **ep8** checkpoint producing its B→B and B→A numbers. RV-PD (`tools/run_rvpd.py`; SimpleViT-Nyströmformer, 6 layers, 8 heads, 256-dim, AdamW at $3{\times}10^{-4}$, cosine annealing, FP16, batch 128) ran the full 50 epochs, the best checkpoint by validation loss landing at **epoch 49** (`val_loss` = 0.0098, roughly two orders of magnitude below its segment-C validation loss, consistent with Breakfast's smaller within-video variance). RV-PL k=3 (`rvpl_pipeline/train.py`; 2-layer causal GRU, hidden 256, plus prediction MLP, trained on precomputed frozen-encoder embeddings, MSE, horizon $k = 3$, 64-frame windows, AdamW at 10^{-4}, 30 epochs) selected its best checkpoint at **epoch 25**. Every Breakfast-trained checkpoint is evaluated under exactly the same protocol as the Assembly-trained one (shared MAD detector with window = 15, $k_{\text{mad}} = 3$, against the 68-video Breakfast Coarse test split via `tools/evaluate_model.py`).

Why no separate Breakfast channel audit. The EST features use the same extraction recipe on both corpora, so the per-channel statistics above carry over qualitatively; only the temporal-density profile changes (Breakfast videos are longer and slower), which the cross-corpus matrix of Section 5.2.1 reads directly.

A.4 Full cross-corpus metric grid

For completeness, Table A.2 gives the complete six-metric breakdown — precision, recall, F1, MCC, balanced accuracy and event-level IoU — at every tolerance $\tau \in \{5, 10, 15, 30\}$ frames for all twelve train/test cells of the cross-corpus matrix (A = Assembly segment-C, B = Breakfast Coarse V2.2), plus the pose baseline BL on both corpora (BL is retrained online per video, so it has no global training corpus and only its on-corpus rows A→A and B→B are defined — the transfer cells A→B and B→A do not apply). Every value is a mean-per-video score under the shared MAD detector and unified evaluator (`tools/evaluate_model.py`); balanced accuracy and IoU are derived from the same per-video TP/FP/FN/TN counts by `tools/compute_balacc_iou.py`, and the P/R/F1/MCC columns reproduce the headline tables exactly. The headline metric throughout the thesis is MCC@15 (the $\tau = 15$ row of each block).

Table A.2: Full per-metric, per-tolerance cross-corpus results. Mean-per-video Precision (P), Recall (R), F1, MCC, Balanced Accuracy (BalAcc) and event-level IoU at $\tau \in \{5, 10, 15, 30\}$ frames, for the three encoders across the 2×2 train/test matrix (A = Assembly segment-C, B = Breakfast Coarse V2.2) and the pose baseline BL. Same MAD detector, evaluator and mean-per-video aggregation as the headline tables. †BL is retrained *online* per video, so it has no global training corpus: only its on-corpus rows (A→A, B→B) are defined; the transfer cells A→B and B→A do not apply.

Model	Dir.	τ	P	R	F1	MCC	BalAcc	IoU
EST grid= 4	A→A	5	0.124	0.611	0.198	0.255	0.790	0.114
		10	0.175	0.861	0.278	0.364	0.916	0.168
		15	0.187	0.916	0.298	0.390	0.944	0.183
		30	0.200	0.968	0.316	0.415	0.969	0.197
	A→B	5	0.220	0.476	0.291	0.303	0.726	0.179
		10	0.336	0.720	0.442	0.471	0.850	0.296
		15	0.373	0.792	0.488	0.522	0.886	0.340
		30	0.433	0.914	0.565	0.608	0.948	0.417
	B→A	5	0.132	0.621	0.209	0.266	0.795	0.122
		10	0.181	0.872	0.286	0.373	0.922	0.174
		15	0.194	0.929	0.306	0.400	0.950	0.190
		30	0.202	0.968	0.320	0.418	0.970	0.200
	B→B	5	0.206	0.488	0.277	0.294	0.730	0.169
		10	0.316	0.744	0.424	0.461	0.860	0.280
		15	0.353	0.819	0.472	0.514	0.899	0.324
		30	0.411	0.932	0.546	0.595	0.956	0.399
RV-PD	A→A	5	0.144	0.574	0.218	0.265	0.775	0.128

continued on next page

Table A.2 continued

Model	Dir.	τ	P	R	F1	MCC	BalAcc	IoU
		10	0.193	0.774	0.294	0.362	0.875	0.181
		15	0.215	0.857	0.327	0.405	0.917	0.205
		30	0.232	0.919	0.351	0.436	0.949	0.224
	A→B	5	0.245	0.540	0.327	0.344	0.758	0.208
		10	0.343	0.774	0.460	0.495	0.877	0.313
		15	0.403	0.884	0.535	0.577	0.932	0.384
		30	0.436	0.953	0.578	0.625	0.967	0.427
	B→A	5	0.140	0.546	0.212	0.256	0.760	0.125
		10	0.200	0.833	0.306	0.383	0.905	0.190
		15	0.214	0.886	0.327	0.410	0.931	0.206
		30	0.229	0.943	0.351	0.439	0.960	0.224
	B→B	5	0.256	0.501	0.323	0.336	0.740	0.207
		10	0.398	0.775	0.500	0.531	0.879	0.356
		15	0.447	0.853	0.558	0.593	0.919	0.408
		30	0.490	0.923	0.610	0.649	0.954	0.465
RV-PL k=3	A→A	5	0.113	0.156	0.115	0.118	0.575	0.068
		10	0.176	0.223	0.175	0.180	0.608	0.107
		15	0.226	0.271	0.221	0.228	0.632	0.139
		30	0.311	0.377	0.303	0.317	0.686	0.204
	A→B	5	0.192	0.212	0.183	0.180	0.600	0.115
		10	0.335	0.365	0.317	0.323	0.677	0.210
		15	0.406	0.440	0.383	0.393	0.715	0.263
		30	0.528	0.572	0.498	0.516	0.782	0.361
	B→A	5	0.087	0.291	0.126	0.143	0.635	0.071
		10	0.148	0.464	0.210	0.243	0.722	0.125
		15	0.177	0.547	0.252	0.292	0.764	0.154
		30	0.245	0.704	0.326	0.382	0.843	0.209
	B→B	5	0.148	0.260	0.177	0.179	0.623	0.111
		10	0.270	0.436	0.308	0.321	0.712	0.216
		15	0.381	0.548	0.403	0.425	0.769	0.295
		30	0.487	0.654	0.490	0.522	0.822	0.367
BL (pose)†	A→A	5	0.108	0.171	0.115	0.119	0.581	0.067
		10	0.150	0.240	0.160	0.168	0.615	0.097
		15	0.204	0.316	0.215	0.228	0.653	0.132
		30	0.264	0.406	0.279	0.298	0.699	0.178

continued on next page

Table A.2 continued

Model	Dir.	τ	P	R	F1	MCC	BalAcc	IoU
	B→B	5	0.278	0.275	0.256	0.256	0.632	0.163
		10	0.424	0.422	0.393	0.399	0.707	0.268
		15	0.480	0.487	0.448	0.458	0.740	0.320
		30	0.565	0.574	0.526	0.541	0.784	0.386

Appendix B

Code and Reproducibility

Every experiment in this thesis is driven by scripts in the accompanying code repository. The narrative chapters describe each procedure in prose; this appendix gives the one-to-one map from procedure to the script that implements it, so that any reported number can be traced to the exact code that produced it. Paths are relative to the repository root. Random seeds are recorded inline where they matter (Section 5.6); the split manifests additionally record the seed and strategy tag alongside the corpus they define.

Table B.1: Procedure-to-script map for the experiments of this thesis.

Procedure	**Script (repository-relative path)**
Corpus, ground truth and splits	
Segment-C corpus construction (seed 42, Strategy-C window placement)	`tools/build_segment_corpus.py`
Segment-C 193/57 train/test split	`tools/build_segment_split.py`
Verb-level ground-truth generation (raw release → 5 fps)	`tools/generate_verb_gt.py`
Person- and content-disjoint split assertion	`tools/smoke_tests/test_person_disjoint.py`
Models	
RV-PD / SimpleViT encoder (reference implementation)	`pipeline_v2/model.py`
Pipeline and detector defaults (d_p, h_p)	`pipeline_v2/config.py`
EST GRU-autoencoder architecture	`est_gebd_model/model.py`
EST training (seed 0)	`est_gebd_model/train.py`
RV-PL predictor head (`prediction_horizon` knob)	`rvpl_pipeline/model.py`
Detection, evaluation and orchestration	
MAD / peak boundary detection (`detect_peaks` lineage)	`tools/detect_boundaries.py`
Unified evaluator (scores HDF5 → per-tolerance metrics)	`tools/evaluate_model.py`
Canonical Phase-0 evaluator (boundaries JSON → metrics)	`tools/evaluate_boundaries_json.py`
Five-model segment-C orchestration chain	`tools/chain_rvpd_rvpl.sh`
Diagnostics	
Perception-free chance-floor dummies	`tools/dummy_chance_floors.py`
Training-free TSM novelty baseline	`tools/tsm_novelty_baseline.py`