Infusing Theories into Deep Learning for Human-Aligned Interpretable Boundary Anticipation

Enjie Jiang †, Shohei Hidaka †

† Japan Advanced Institute of Science and Technology
s2510059@jaist.ac.jp

Abstract

Humans segment continuous action into events. Event Segmentation Theory (EST) attributes boundaries to prediction error, and recent computational models (e.g., On-GEBD/ESTimator, ICCV 2025) detect boundaries only after a change has occurred. Yet humans also anticipate boundaries before the physical transition. We ask whether a goal/anticipatory signal (GD) coexists with the error signal (ED) in the segmentation of observed human motion. Using a skeleton-graph model on real continuous motion capture (BABEL/AMASS; 982 sequences; subject-held-out), we find (i) a GD signal that rises before human-annotated boundaries, peaking about six frames earlier and positive in all 17 held-out subjects (95% CI of the pre-boundary rise excludes zero), and (ii) that GD predicts boundaries from body motion better than the error signal (AUROC 0.74 vs. 0.65) and far better than a fair timing baseline (AUPRC 0.16, near chance). The same anticipatory signal grows stronger, and dissociates cleanly from error, once a goal channel (the hands) or an observer's gaze is available. These preliminary results indicate that action event segmentation is not error-driven alone: an anticipatory component is present that current error-only models miss.

Keywords: Event Segmentation, Boundary Anticipation, Goal-driven, Skeleton Graph, Deep Learning

1. Introduction

In continuous activity, humans consistently perceive boundaries between meaningful action units such as reaching, grasping, and placing (Zacks et al., 2007; Kurby & Zacks, 2008). A dominant account, Event Segmentation Theory (EST), holds that observers perceive a boundary when a transient increase in prediction error signals that ongoing activity no longer matches their internal model.

Most computational boundary detectors instantiate this error-driven view. The most recent, ESTimator (Jung et al., 2025), predicts the next frame only to measure its prediction error and declares a boundary when that error spikes. This scheme has a structural limitation: because it responds to an error that has already occurred, it can only confirm a boundary at or after the physical transition. It cannot express the pre-change anticipation that people show when they sense an action is about to end, and that is reflected in anticipatory pre-boundary neural activity (Nguyen et al., 2025).

To capture pre-change anticipation, we hypothesize a second, goal-driven process (GD) operating alongside the error-driven one (ED). Rather than waiting for a discrete error, GD estimates how far the current action has progressed toward its goal, so anticipation can rise before the transition; ED and GD are expected to be temporally dissociable, with GD leading and ED lagging relative to the boundary. This paper provides **preliminary empirical evidence** for a GD signal in a computational model of action segmentation, and tests whether GD carries boundary-relevant information beyond ED.

2. Method

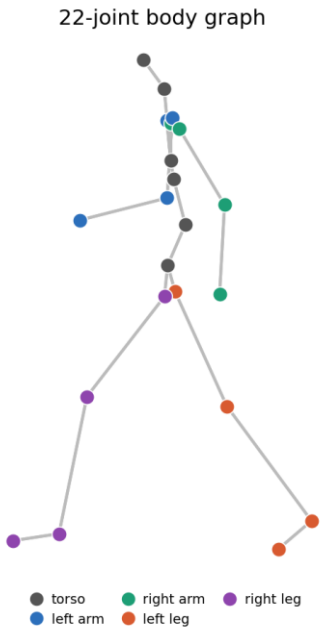

Figure 1 Input representation: each pose is a 22-joint body graph (nodes: joints; edges: bones), coloured by body part. Motion is a sequence of such graphs.

We represent each pose as a 22-joint body graph (Figure 1) and encode motion with a spatial graph-attention network (GAT; Veličković et al., 2018) over a skeleton graph (cf. ST-GCN; Yan et al., 2018), followed by a temporal recurrent encoder (GRU). From this shared backbone we read, for each sliding window, a GD (anticipation) signal and an ED (error / representation-change) signal. ED measures how far the current representation departs from a short-term prediction and is expected to peak at the transition; GD is read from the pre-transition motion and is expected to rise before it.

We use real continuous motion capture from AMASS (Mahmood et al., 2019) with frame-level BABEL action annotations (Punnakkal et al., 2021), whose transition segments provide human-annotated event boundaries (982 sequences). Subjects are split into train / validation / test with no subject overlap, and boundaries are used only for evaluation. For each window we read GD and ED at its final frame, and evaluate (a) the boundary-aligned timing of GD and (b) per-window boundary prediction (AUPRC / AUROC),

with per-subject statistics and negative controls (a fair timing baseline, shuffled and random labels). To locate where the anticipatory signal resides, we additionally analyse three settings in which a goal or attention channel is explicit: whole-body grasping with articulated hands (GRAB; Taheri et al., 2020), an observer's gaze during pick-and-place (MoGaze; Kratzer et al., 2021), and a model-free dynamical early-warning marker on AMASS (Section 3.4).

3. Results

3.1 GD rises before the boundary

The boundary-aligned GD signal increases monotonically and peaks about six frames before the annotated boundary (trend r = 0.80), then declines (Figure 2). The pre-peak rise and post-peak decay together trace a complete anticipate-then-resolve cycle centred slightly ahead of the physical transition — the temporal signature expected of an anticipatory process rather than a reactive one. This profile is not an artefact of sample size: the trend correlation rises from r = 0.38 on a 77-sequence subset to r = 0.80 on the full set, so a larger sample sharpens rather than creates the effect. The pre-boundary rise is positive in all 17 held-out subjects (aggregate rise 0.05; 95% CI 0.042–0.059; Figure 3). Although small for any single boundary (56% positive, p ≈ 0.02; Figure 4), it is highly reliable across subjects — the characteristic signature of a genuine but modest cognitive effect that surfaces only in aggregate.

3.2 GD's information is motion-derived

For per-window boundary prediction (Table 1; base rate 0.18), the GD signal read from body motion (AUPRC 0.33, AUROC 0.74) clearly exceeds the error signal ED (0.26, 0.65) and a fair phase-based timing baseline (AUPRC 0.16, near chance). The margin over the timing baseline is the decisive comparison: because that baseline is given the same temporal position within each sequence but not the motion itself, GD's advantage must originate in body motion rather than in when the window happens to occur. Nor is the signal merely deceleration: GD correlates with instantaneous speed at only r = −0.17, velocity explains just 5% of its variance, and after residualizing out speed the pre-boundary rise is essentially unchanged (+0.45; 95% CI [0.31, 0.62]) while speed alone predicts boundaries at chance. Two negative controls — shuffled and randomly assigned boundary labels — collapse to the base rate. GD therefore captures boundary-relevant information that neither timing, speed, nor error alone provides.

Table 1 Per-window boundary prediction (subject-held-out). GD exceeds ED and a fair timing baseline.

Signal	AUPRC	AUROC
GD (motion)	**0.33**	**0.74**
ED (error)	0.26	0.65
Fair timing (phase)	0.16	–
Base rate	0.18	–

3.3 GD and ED are complementary

Because GD is read from the pre-transition motion whereas ED responds to representation change at the transition, the two signals should encode different boundary evidence. Adding GD to the error signal improves per-window prediction for 13 of 17 held-out subjects (Wilcoxon p = 0.02), so GD contributes information beyond ED rather than duplicating it — consistent with a dual-signal account in which an anticipatory (GD) and a confirmatory (ED) component jointly underlie action event segmentation.

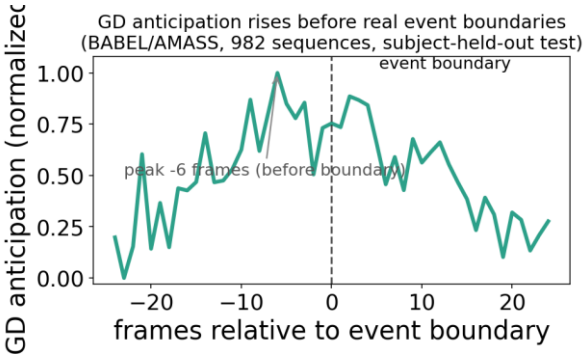

Figure 2 On real BABEL/AMASS motion, the boundary-aligned GD signal rises before human-annotated boundaries and peaks about six frames before t = 0 (subject-held-out test).

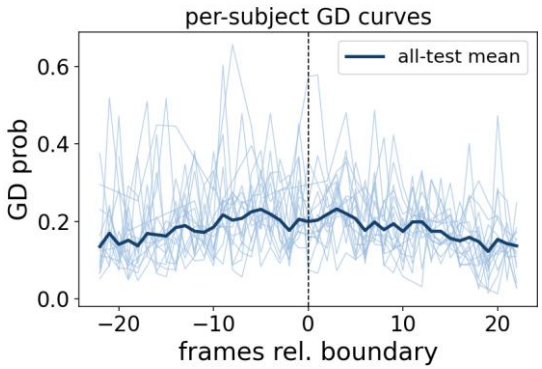

Figure 3 Per-subject robustness: all 17 held-out subjects' boundary-aligned GD curves (faint) and the test-set mean (bold) rise toward the boundary.

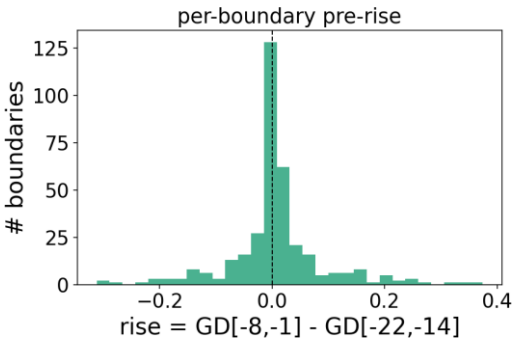

Figure 4 Per-boundary pre-rise distribution: the rise is positive for 56% of individual boundaries — small per event but reliable across subjects.

3.4 The signal is clearest in the goal and attention

channels

In bare body motion the anticipatory effect, though reliable, is modest — expected if a skeleton carries little information about the actor's goal. Three further analyses, on independent data, show that the same anticipatory signal becomes strong, and its dissociation from ED becomes clean, once a goal or attention channel is available (Table 2). When the hands are modelled explicitly (GRAB whole-body grasping), grasp pre-shaping already carries the upcoming grasp: a supervised read-out from the body reaches AUPRC 0.291 ± 0.035 (AUROC 0.728), well above a kinematic baseline (0.146), and adding the hands raises it to 0.391. When anticipation is read from an observer rather than the actor (MoGaze pick-and-place with eye tracking; Figure 5), the observer's gaze commits to the next goal about 82 frames before the boundary — roughly fifty frames ahead of the same read-out from body motion, which commits only at −32 (per-subject gaze-over-motion lead 95% CI [44, 61] frames; goal decoding 0.79 vs. 0.02 for shuffled goals; placebo clean). The attention channel thus realises the GD/ED dissociation that body motion only hints at. Finally, a model-free dynamical marker (lag-1 autocorrelation of velocity, a critical-slowing early-warning signal) rises before boundaries on AMASS in all four window sizes tested, surviving speed control (residual +0.030; 95% CI [0.009, 0.048]) and a clean placebo (Figure 6).

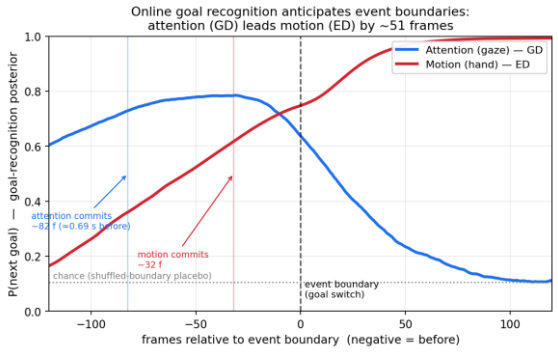

Figure 5 MoGaze pick-and-place (observer eye-tracking): the online goal-recognition posterior read from gaze (blue, GD) commits to the next goal about 82 frames before the boundary — roughly 51 frames ahead of the same read-out from body motion (red, ED), which commits at −32.

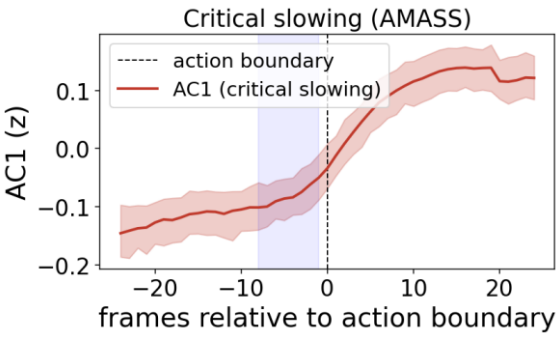

Figure 6 AMASS continuous motion: the lag-1 autocorrelation of velocity (AC1, a critical-slowing early-warning signal) rises before action boundaries (boundary-aligned mean with subject-bootstrap CI band), surviving speed control and a clean placebo.

Table 2 Converging evidence for an anticipatory (GD) signal across channels and datasets.

Channel / dataset	Anticipatory result
Body motion — BABEL	GD AUPRC 0.33 > ED 0.26 > timing 0.16; peak −6 fr; 17/17 subj
+ hands — GRAB	grasp pre-shaping 0.29 → 0.39 AUPRC (> kinematic 0.15)
Attention — MoGaze	gaze leads motion ~51 fr to the goal (0.79 vs 0.02)
Dynamics — AMASS	critical slowing (AC1) rises pre-boundary, 4/4 windows

4. Discussion and Future Work

Across four independent analyses, an anticipatory signal reliably precedes human-annotated boundaries and carries boundary information that an error-only account lacks, challenging the error-only orthodoxy embodied by current computational detectors. The signal is modest when read from bare body motion but becomes strong, and its dissociation from ED becomes clean, once the actor's goal is represented (hands) or the observer's attention is measured (gaze) — suggesting that anticipation lives in a goal/attention channel that pure skeleton kinematics under-represent. **Limitations.** The GD read-out uses goal-directed supervision, so the GD–ED comparison is not supervision-matched (we frame GD as boundary-predictive, not causal); the GD-before-ED dissociation is clean in gaze but only directional in body motion; and we do not yet align to human perceptual segmentation. **Ongoing work** formalizes GD as online goal inference (goal-completion and goal-posterior entropy) within an anticipatory goal-event model, and will validate the model's GD timing against independently collected human event-segmentation data.

5. Acknowledgements

This study is supported by KAKENHI Grant-in-Aid for Scientific Research (A) JP26H02527 and Grant-in-Aid for Scientific Research (B) JP23H0369.

References

Jung, H., et al. (2025). Online generic event boundary detection. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Kratzer, P., Bihlmaier, S., Midlagajni, N. B., Prakash, R., Toussaint, M., & Mainprice, J. (2021). MoGaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. IEEE Robotics and Automation Letters, 6(2), 367–373.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. Trends in Cognitive Sciences, 12(2), 72–79.

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., & Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Nguyen, T. T., Etzel, J. A., Bezdek, M. A., & Zacks, J. M. (2025). Multiple event segmentation mechanisms in the human brain. bioRxiv.

Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., & Black, M. J. (2021). BABEL: Bodies, action and

behavior with English labels. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
Taheri, O., Ghorbani, N., Black, M. J., & Tzionas, D. (2020). GRAB: A dataset of whole-body human grasping of objects. Proceedings of the European Conference on Computer Vision (ECCV).
Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations (ICLR).
Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. Psychological Bulletin, 133(2), 273–293.