TOWARDS HUMAN-LIKE EVENT BOUNDARY DETECTION IN UNSTRUCTURED VIDEOS THROUGH SCENE-ACTION TRANSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans segment continuous experience into episodes by detecting perceptual novelties and retrospectively consolidating them into coherent memories. Inspired by this cognitive process, we introduce a two-level, backward-only event segmentation framework designed for cognitive agents that must structure continuous sensory input into episodic memory. At Level 1, an error-driven novelty detector with a semi-supervised adaptive thresholding module identifies candidate transitions robust to noise, viewpoint shifts, and repeated micro-actions. At Level 2, an uncertainty-driven consolidation mechanism retrospectively validates and merges boundaries using multimodal cues (scene graphs, captions, audio), producing stable, semantically grounded episodes without relying on future frames. Unlike prior GEBD approaches that depend on motion cut-points or heavy task-specific supervision, our method leverages sparse labels only for threshold calibration, making it label-efficient, cognitively grounded, and broadly applicable. Experiments on ADL-GEBD and Ego4D show state-of-the-art performance, with our semi-supervised model surpassing heavily supervised baselines. This work introduces episodic segmentation for cognitive agents, bridging human memory theory with scalable machine perception.

1 INTRODUCTION

Humans naturally parse continuous experience into events, a process known as event segmentation (Nguyen et al., 2025). Boundaries are perceived when perceptual features (e.g., motion, sound) or conceptual features (e.g., goals, intentions) change, forming a hierarchical structure of fine- and coarse-grained episodes. These boundaries are not arbitrary: they scaffold episodic memory, enabling people to recall past experiences, learn new skills, and anticipate future outcomes. Figure 1 illustrates our dual-level framework: error-driven novelty detection is retrospectively consolidated into semantically coherent episodes, mirroring how human episodic memory stabilizes experience.

Inspired by these findings, we ask: *How can an artificial agent segment its continuous sensory stream into meaningful episodes suitable for episodic memory?* Unlike offline video analysis, an embodied agent must structure experience in real time based on places, participants, and task-level transitions, rather than superficial discontinuities. For example, in Activities of Daily Living (ADL), relevant transitions include entering a new room, shifting from preparing to cooking, or the arrival of a new person—precisely the type of semantic boundaries that support long-horizon memory and reasoning.

A cognitively capable agent must therefore *segment memory in a human-like way*, forming episodes that are stable, interpretable, and grounded in semantics rather than transient motion. Such episodic segmentation enables agents to compress continuous experience, support causal reasoning, and align long-term memory with dialogue and task planning. In contrast, existing Generic Event Boundary Detection (GEBD) methods often rely on motion-driven cut points, which fragment continuous streams and degrade memory stability.

We propose a cognitively grounded representation learning framework for event segmentation. Our central idea is that boundaries should emerge not from immediate motion changes, but from the retrospective stability of representations over time. To achieve this, we introduce a *backward-looking temporal windowing mechanism* that compares the present to the recent past, avoiding reliance on

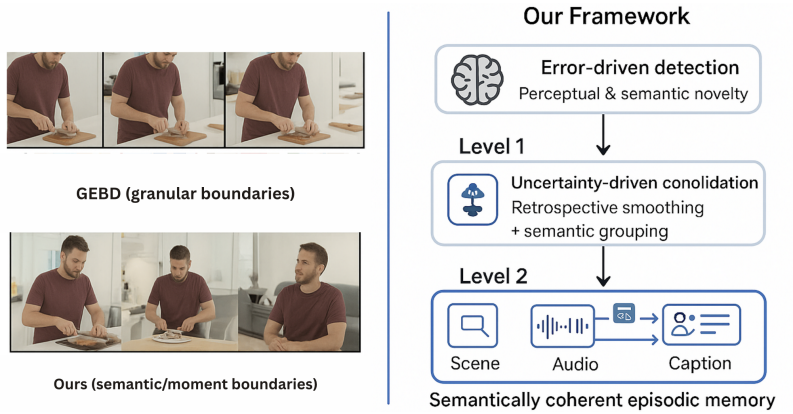

Figure 1: Our cognitively inspired hierarchical event segmentation framework. Level 1 detects fine-grained perceptual novelties (*error-driven updates*), while Level 2 retrospectively consolidates boundaries into stable, semantically coherent episodes (*uncertainty-driven updates*). Motion-driven GEBD methods often fragment actions into multiple cuts—e.g., each knife jitter while preparing a sandwich is marked as a separate boundary. In contrast, our approach groups such micro-changes into a single meaningful action, such as *spreading butter on bread*, yielding coherent episodes aligned with human perception.

unavailable future frames. At a second level, we retrospectively consolidate candidate boundaries using scene graphs, audio cues, and caption semantics, ensuring that episodes reflect stable shifts in meaning rather than transient visual changes. In addition, we introduce a *semi-supervised adaptive thresholding module* that learns to calibrate novelty sensitivity from retrospective statistics, improving robustness to noise, jitter, and viewpoint shifts.

Key Contributions

- **Cognitively grounded paradigm for cognitive agents:** We propose a dual-level, backward-only event segmentation framework inspired by human episodic memory, enabling artificial agents to structure continuous sensory streams into interpretable episodes.
- **Semi-supervised adaptive threshold detection:** A label-efficient thresholding mechanism that dynamically calibrates sensitivity from retrospective statistics, improving robustness to noise, jitter, and viewpoint changes.
- **Multimodal integration:** Our approach consolidates boundaries using semantic (captions, scene graphs), perceptual (DINOv2, SSIM, LPIPS), and linguistic (dialogue-aware) cues in a unified retrospective validator, without reliance on dense frame-level labels or task-specific fine-tuning.
- **Strong empirical validation:** State-of-the-art results on ADL-GEBD and Ego4D, where our semi-supervised framework outperforms both motion-driven GEBD methods and large supervised models, demonstrating scalability and human-aligned segmentation.

2 Related Work

Generic Event Boundary Detection (GEBD). Generic Event Boundary Detection (GEBD) Mike Zheng Shou & Feiszli. (8075) aims to localize perceptual transitions in video without predefined labels. Early methods formulated GEBD as frame-level binary classification Mike Zheng Shou & Feiszli. (8075); Jiaqi Tang & Wang. (3355); Dexiang Hong & Zhang. (2107), but these models often over-segment due to their reliance on superficial appearance or motion changes. More recent work incorporated contrastive learning Hyolim Kang & Kim. (2106), compact encodings Congcong Li & Zhang. (1396), and transformer-based architectures Sourabh Vasant Gothe & Kashyap. (2023); Congcong Li & Wen. (2206), often coupled with optical flow Rui Qian & Cui. (2112). While effective at detecting local visual novelty, such approaches tend to produce fragmented segmentations that struggle with higher-level semantics such as goals or dialogue continuity. Unsupervised variants (e.g., PySceneDetect Castellano., PredictAbility Mike Zheng Shou & Feiszli. (8075), CoSeg Xiao Wang & Luo. (2109)) exploit

Aspect	GEBD Focus	Our Task Focus
Granularity	Micro changes and frame-level transitions (often motion- or appearance-driven)	Coarse, semantically coherent episodes (scene + action + dialogue continuity)
Output Style	Fragmented boundaries highlighting perceptual shifts and uncertainties	Stable, consolidated episodes preserving narrative and semantic flow
Strengths	Sensitive to subtle changes; effective at detecting ambiguous or uncertain regions	Captures long-horizon coherence; supports reasoning, memory, and downstream tasks
Limitations for Our Use Case	Over-fragmentation $\rightarrow$ splits continuous dialogue, micro-actions, or camera jitter into many segments	Possible under-segmentation if overly coarse, but maintains meaningful episodic units
Cognitive Alignment	Perceptual novelty and local frame changes	Episodic memory structure (what, when, where), retrospective consolidation

Table 1: Comparison of GEBD objectives versus our episodic segmentation task. GEBD emphasizes sensitivity to fragmented shifts, while our task prioritizes stable, semantically grounded episodes.

reconstruction losses or pixel variations, while hybrids such as UBoCo Hyolim Kang & Kim. (2007) combine multiple objectives. Despite these advances, most GEBD approaches remain focused on micro-level granularity. As summarized in Table 1, this sensitivity makes GEBD well-suited for perceptual novelty detection, but misaligned with the stability required for episodic segmentation.

Motion and Visual Correspondence Learning. Motion cues have long been central to video understanding, from classical optical flow Lucas & Kanade. (1981); Farnebäck. (2003) to modern motion-aware architectures Heeseung Kwon & Cho. (2020); Jiaqi Tang & Wang. (3355); Ayush K Rai & O'Connor. (2728). These techniques are effective for dense action localization but often generate visually reactive segmentations that neglect semantic continuity. Our approach diverges by avoiding explicit motion cues, instead leveraging semantically aligned representations with adaptive thresholds that flexibly capture both fine and coarse boundaries—crucial in egocentric or dialogue-heavy videos where appearance shifts may not correspond to meaningful transitions.

Egocentric Video and Multimodal Understanding. The Ego4D benchmark Grauman et al. (2022) has driven progress in egocentric video research, emphasizing tasks such as episodic memory and natural language query (NLQ). Current solutions typically adopt proposal-based Mo et al. (2022) or transformer-based Lei et al. (2021) pipelines, built on pretrained vision–language encoders like CLIP Radford et al. (2021b;a), VideoMAE Tong et al., or InternVideo Chen et al. (2022b). While effective for fine-grained retrieval, these systems are optimized for short-term alignment and often fail to capture higher-order transitions, such as shifts across environments or narrative stages.

Summary. In contrast to GEBD (Table 1), which emphasizes sensitivity to micro changes, our work prioritizes stable, semantically coherent episodes. By integrating semantic representations with a learnable boundary threshold, our approach captures both fine and coarse transitions without over-fragmentation. This enables structured episodic understanding, which is particularly beneficial for applications in robotics, surveillance, and assistive systems where long-horizon coherence and memory alignment are essential.

3 Approach

Episodic memory encodes not only *what* happened, but also *when* and *where* it occurred (Tulving, 2002). For autonomous agents, this requires transforming continuous sensory streams—such as egocentric video—into stable, semantically coherent episodes. We propose a cognitively inspired two-level framework: (i) **adaptive boundary detection**, which selects candidate transitions based on retrospective statistics within a short backward window, and (ii) **retrospective consolidation**, which validates and merges boundaries into coherent episodes using semantic, perceptual, and dialogue cues. Both mechanisms are strictly *backward-facing*, reflecting the episodic memory constraint that only past context is available at decision time.

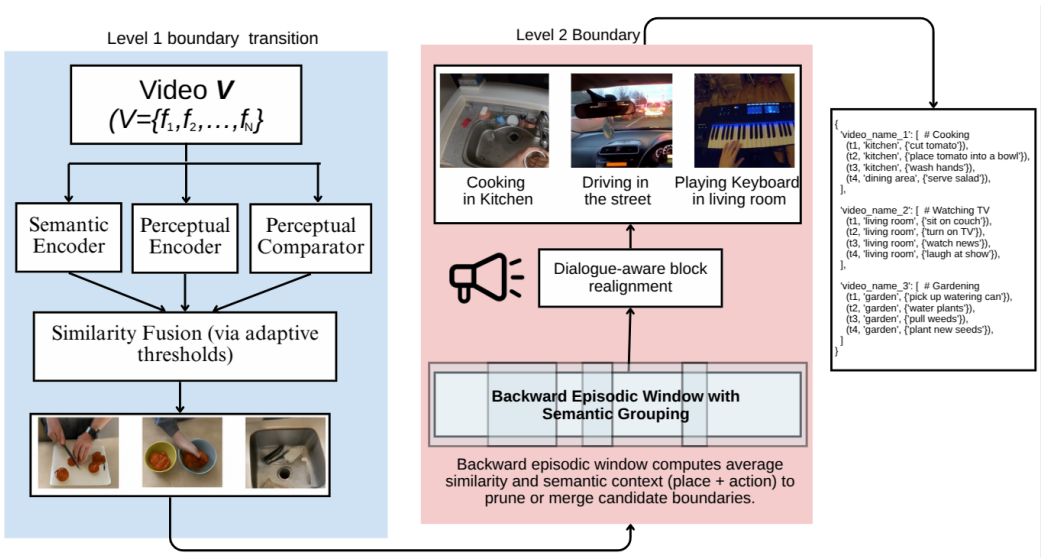

Figure 2: Overview of our cognitively inspired two-level episodic segmentation framework. **Level 1** (left, blue) identifies candidate boundaries through *adaptive thresholding*, comparing incoming frames against retrospective statistics within a fixed backward window across semantic, perceptual, and comparative encoders. **Level 2** (right, red) retrospectively validates and consolidates these candidates via *multimodal integration*, using semantic grouping (place + action), perceptual similarity, and dialogue alignment in a unified validator. Both stages are inherently *backward-facing*, operating only on past context to transform continuous sensory streams into stable, semantically coherent episodes of *what*, *when*, and *where*.

3.1 Level 1: Adaptive Boundary Detection

The first stage proposes candidate boundaries by comparing each incoming frame with a fixed window of the k most recent frames. Frames $F = \{I_1, I_2, \ldots, I_N\}$ are uniformly sampled, and a subset $K \subset F$ of keyframes is selected when local similarity drops below an adaptive threshold. Each frame I_i is encoded via three parallel streams: a **Semantic Encoder** (objects, relations, and high-level concepts), a **Perceptual Encoder** (appearance and spatial structure), and a **Comparator** (patchwise dissimilarity). The fused similarity between frames I_j and I_i is defined as:

$$\text{Sim}(I_j, I_i) = \sum_{m=1}^{M} \alpha_m \cdot S_m(I_j, I_i), \quad \sum_{m=1}^{M} \alpha_m = 1, \tag{1}$$

where S_m denotes a modality-specific similarity metric and α_m its learnable weight.

Backward-Facing Similarity. For each frame I_i, similarity is computed against the k most recent keyframes. This backward-only evaluation avoids spurious boundaries from transient viewpoint shifts (e.g., head turns in egocentric video).

3.1.1 Semi-Supervised Adaptive Thresholding

To determine whether a candidate frame I_i marks a boundary, we compute statistics over the backward window:

$$(\mu, \sigma^2, s_{\text{last}}) = \left(\frac{1}{k} \sum_{j=1}^{k} \text{Sim}(I_{i-j}, I_i),\ \text{Var}_{j=1..k}[\text{Sim}(I_{i-j}, I_i)],\ \text{Sim}(I_{i-1}, I_i) \right), \tag{2}$$

where μ and σ^2 capture similarity stability, and s_{last} measures immediate continuity. These features are passed to a neural module g_θ that predicts an adaptive threshold:

$$\tau(I_i) = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \sigma\big(g_\theta(\mu, \sigma^2, s_{\text{last}})\big), \tag{3}$$

with $\tau(I_i) \in [\tau_{\min}, \tau_{\max}]$.

The boundary probability is then defined as:

$$p(I_i) = \sigma\big(\alpha \cdot (\tau(I_i) - \mu) + \beta\big), \tag{4}$$

where α controls decision sharpness and β is a learnable bias term. A frame I_i is selected as a candidate boundary whenever $p(I_i)$ exceeds a fixed decision threshold.

Training Objective. The module g_θ is trained with a semi-supervised objective:

$$\mathcal{L} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{5}$$

Here, $\mathcal{L}_{\text{sup}}$ is a binary cross-entropy loss on sparsely labeled boundaries; $\mathcal{L}_{\text{self}} = \mathbb{E}[p(I_i)]$ prevents degenerate solutions that reject all boundaries; and $\mathcal{L}_{\text{reg}} = \mathbb{E}[\|g_\theta(\cdot)\|_2^2]$ enforces smoothness in threshold dynamics. Additional details are as given in A.1

3.2 Level 2: Retrospective Boundary Consolidation

While Level 1 captures candidate boundaries, it may still produce spurious splits due to minor appearance changes or repeated micro-actions. Level 2 retrospectively validates and merges boundaries using semantic, perceptual, and linguistic evidence.

Dynamic Consolidated Window (DCW). We check each candidate boundary I_{k_i} by looking backward into a short window

$$W_i = \{I_{k_i - w}, \ldots, I_{k_i - 1}\}.$$

We compute the average similarity:

$$\text{AvgSim}(I_{k_i}) = \frac{1}{|W_i|} \sum_{j \in W_i} \text{Sim}(I_j, I_{k_i}). \tag{6}$$

A candidate boundary I_{k_i} is pruned if its average similarity to past frames exceeds a threshold, i.e., $\text{AvgSim}(I_{k_i}) > \theta$. Furthermore, we evaluate semantic consistency within the window by comparing scene and action features. If the current boundary exhibits the same scene and activity as the preceding frames, it is merged with the earlier segment.

This *Dynamic Consolidated Window* thus functions as a backward-looking validation mechanism: boundaries are only retained when there is a meaningful change in scene or action, preventing spurious segmentation. Additional implementation details are provided in Appendix A.2.

Dialogue-Aware Alignment. Finally, boundaries are aligned with dialogue structure. If a visual boundary falls within an active utterance, it is deferred until the dialogue ends. This ensures that episodes preserve both perceptual and conversational coherence. Additional details are as given in Appendix A.3

4 Experimental Details

Datasets. Our focus is on detecting *semantically meaningful moments*, such as place changes or shifts in activity context, rather than short-term motion fluctuations. Accordingly, we evaluate on two datasets designed for naturalistic, narrative-driven segmentation: ADL-GEBD Shou et al. (2021); Ho-Le et al. (2025) and Ego4D Grauman et al. (2022).

ADL-GEBD provides over 1M densely annotated frames of household activities, where boundaries are marked with precise start–end timestamps. These short-horizon transitions capture low-level novelty, making ADL-GEBD an ideal testbed for evaluating the sensitivity of our Level 1 (error-driven) boundary detection. Ego4D, in contrast, contains long-form egocentric videos across diverse daily scenarios such as cooking, exercising, and socializing. Its annotations include moment-level queries with explicit temporal spans, aligning closely with episodic memory and narrative grounding. By treating the *start and end timestamps* of these moment queries as boundary markers, we can also perform GEBD-style analysis within the Ego4D setting. This makes Ego4D particularly well suited for testing our Level 2 (uncertainty-driven) retrospective consolidation.

Together, ADL-GEBD and Ego4D span the spectrum from fine-grained perceptual updates to long-horizon episodic formation, providing a cognitively motivated evaluation setting.

Implementation Details. Our method operates in an semisupervised fashion by comparing each frame with the previous frame most recent keyframes using an adaptive similarity score. To balance

semantic, perceptual, and structural cues, we combine the outputs of several pretrained models. CLIP Radford et al. (2021b) (ViT-L/14@336px) encodes high-level semantic embeddings compared via cosine similarity. DINOv2 Caron et al. (2021) (ViT-B/14) provides dense patch-level features sensitive to spatial detail. LPIPS Alom et al. (2018) captures perceptual differences via learned feature embeddings, while SSIM quantifies structural similarity. EVA-CLIP-Large Sun et al. (2023) generates captions, from which we compute token-level alignment; scene graphs are then derived using a parser Wu et al. (2019).

During inference, a candidate boundary is triggered when similarity with all prior keyframes falls below a learned threshold $\tau(I_i)$. The final similarity score is computed as a weighted combination: CLIP (0.2), DINO (0.3), SSIM (0.2), LPIPS (0.2), and caption-token similarity (0.1). We average results over 5 random seeds and report mean $\pm$ standard deviation. Variance arises primarily from small differences in frame sampling due to decoding.

Training Adaptive Thresholds. Unlike prior GEBD methods that use a fixed cutoff (e.g., $\tau = 0.95$), our threshold is dynamically predicted by the module g_θ (see Section 3). To train this module, we leverage Ego4D *Moment Queries* Grauman et al. (2022), which provide human-queried temporal boundaries aligned with narrative-level shifts. Specifically, $(\mu, \sigma^2, s_{\text{last}})$ statistics from the backward window are paired with query-aligned ground-truth boundaries to supervise the adaptive threshold via a Binary Cross-Entropy loss. This anchors the threshold to meaningful episodic changes rather than arbitrary frame-level fluctuations.

Training is performed with the Adam optimizer (learning rate 10^{-3}) over 50 epochs, with mini-batches of 32 frames. Loss balancing is achieved by weighting the supervised term more strongly ($\lambda_{\text{sup}} = 5.0$) than the self-supervised regularizer ($\lambda_{\text{self}} = 1.0$), reflecting the importance of narrative-level human annotations. The decision sharpness is controlled via a scaling parameter $\alpha = 20.0$, while an L_2 penalty ($10^{-4}\|\tau_{\text{raw}}\|^2$) prevents degenerate solutions. These design choices were tuned to achieve both stability and generalization across domains, improving robustness under noise, jitter, and viewpoint changes.

Experimental Setup. Experiments were conducted on a Linux workstation (Ubuntu 20.04) with a single NVIDIA RTX 3090 GPU (24 GB VRAM) and 128 GB RAM. The pipeline processes approximately one hour of video at $\sim$1.5$\times$ real time, depending on resolution and caption generation latency. All pretrained encoders are used in inference-only mode; only the adaptive threshold module is trained.

Evaluation Metrics. We assess both boundary accuracy and temporal localization. For moment-level localization, we report mean Average Precision (mAP) and Recall@1 (R@1) Shou et al. (2021) at IoU 0.5. For boundary detection, we compute precision, recall, and F1 across tolerance windows ranging from 5% to 50% of video duration; a prediction is correct if it falls within any ground-truth tolerance window. For videos with multiple annotators, we follow Lei et al. (2021) and report the best-aligned score across references, restricting evaluation to videos with inter-rater F1 ≥ 0.3 to ensure reliability.

5 Experiments and Results

5.1 Comparison with Unsupervised Boundary Detection Methods

The dataset described in Section 4 features egocentric videos with frequent scene changes and dense frame-level annotations. To detect meaningful scene boundaries without over-segmenting, our method uses Level 1 detection with a dynamic backward-looking temporal window. Like an agent forming episodic memory, the model only uses past frames to decide if a candidate transition marks a real scene boundary. This helps filter out minor viewpoint changes while keeping boundaries corresponding to significant *place changes*, such as moving between distinct areas in a scene.

We evaluated several unsupervised boundary detection methods on this dataset ourselves, including SceneDetect Castellano., UBoCo Kang et al. (2021), FlowGEBD Gothe et al. (2024), SegSim Aouaidjia et al. (2025), and DDM Tang et al. (2022), and compared their performance to our adaptive-threshold approach.

Table 2 shows that our model achieves an average F1 score of 0.885, outperforming all baselines at every threshold. The backward-looking window and adaptive threshold help the model focus

Method	0.05	0.1	0.15	0.2	0.25	0.30	0.35	0.4	0.45	0.5	Avg
SceneDetect Castellano.	0.336	0.435	0.484	0.512	0.529	0.541	0.548	0.554	0.558	0.561	0.506
UBoCo-TSN Kang et al. (2021)	0.396	0.488	0.520	0.534	0.544	0.550	0.555	0.558	0.561	0.564	0.527
FlowGEBD Gothe et al. (2024)	0.180	0.200	0.209	0.215	0.286	0.290	0.297	0.300	0.308	0.306	0.259
SegSim Aouaidjia et al. (2025)	0.240	0.312	0.336	0.351	0.359	0.369	0.370	0.375	0.379	0.380	0.350
DDM Tang et al. (2022)	0.460	0.480	0.520	0.531	0.540	0.550	0.555	0.558	0.560	0.570	0.532
Ours	**0.71**	**0.80**	**0.81**	**0.82**	**0.824**	**0.826**	**0.828**	**0.83**	**0.836**	**0.84**	**0.885**

Table 2: Performance comparison at different relative distance thresholds. All baselines were evaluated on this dataset by us. Our adaptive-threshold method outperforms all unsupervised approaches, demonstrating strong segmentation accuracy in densely annotated videos.

on meaningful place changes while ignoring minor viewpoint fluctuations. Baselines that rely on frame-level appearance or optical flow often misinterpret jitter as boundaries.

We focus on unsupervised comparisons here even though our method is semi-supervised. Dense frame-level annotations make fully supervised methods prone to overfitting perceptual cues instead of capturing semantic boundaries. Our approach uses light supervision for calibration, preserving the spirit of unsupervised segmentation while achieving higher accuracy. In the following Ego4D experiments, we benchmark against large-scale supervised models, showing that our framework generalizes to both dense and sparse annotation settings.

5.2 Comparison with Downstream Task of Moment Queries

Ego4D contains long-form egocentric videos where understanding activities depends on narrative coherence rather than short-term cues. The Ego4D moment query dataset provides *start and end times* for annotated semantic moments. Detecting event boundaries accurately is crucial for the downstream task of *moment query detection*, where the goal is to retrieve temporally grounded video segments corresponding to a given query.

We compare against supervised vision–language grounding models—InternVideo Chen et al. (2022b), EgoVLP Lin et al. (2022), EgoVideo-V Chen et al. (2022b), and EgoVideo-MQ Chen et al. (2022a)—all trained for moment localization on Ego4D timestamps. Our method also uses moment queries but differs in consolidation: thresholds are adaptively tuned to timestamps, Level 2 refinement integrates multimodal cues, and final captions/windows are aligned with *place* and *action*, leading to more accurate event segmentation and stronger downstream localization.

#	Feature	Validation	
		Average mAP	R1@0.5
A	InternVideo + EgoVLP	27.85	46.98
B	EgoVideo-MQ	28.53	46.07
C	InternVideo + **EgoVideo-V**	31.30	50.21
D	InternVideo + **EgoVideo-MQ**	31.00	49.28
E	InternVideo + **EgoVideo-V** + **EgoVideo-MQ**	32.48	51.04
F	**Ours**	**35.2**	**57.1**

Table 3: Comparison on Ego4D validation. Baselines (A–E) are supervised vision–language grounding models trained for moment localization. Our method (**F**) adapts thresholds, applies Level 2 consolidation, and aligns place–time cues, improving both event segmentation and downstream moment query detection.

As shown in Table 3, our model achieves the highest mAP (35.2) and R1@0.5 (57.1). By consolidating multimodal signals and adapting thresholds with place–time alignment, we achieve more reliable episode boundaries, which directly benefits the downstream task of moment query localization.

5.3 Ablation Studies

Our ablations justify the core design choices of our model: adaptive thresholding, temporal windowing, semantic and perceptual reasoning, and multimodal fusion. We compare fixed hyperparameters against our learned adaptive mechanisms, showing that retrospective adaptation provides consistent improvements.

Threshold (%)	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	Avg
65	0.490	0.607	0.647	0.668	0.681	0.689	0.690	0.697	0.700	0.701	0.657
75	0.580	0.697	0.730	0.750	0.761	0.767	0.770	0.775	0.781	0.788	0.740
80	0.610	0.720	0.770	0.780	0.790	0.794	0.800	0.809	0.810	0.815	0.770
85	0.670	0.772	0.790	0.800	0.804	0.810	0.815	0.820	0.829	0.835	0.799
90	0.660	0.740	0.770	0.774	0.776	0.779	0.780	0.784	0.788	0.790	0.774
95	0.590	0.690	0.710	0.715	0.720	0.725	0.728	0.730	0.732	0.735	0.707
Ours (Adaptive)	**0.71**	**0.80**	**0.81**	**0.82**	**0.824**	**0.826**	**0.828**	**0.83**	**0.836**	**0.84**	**0.885**

Table 4: **Threshold sensitivity.** Moderate fixed thresholds (80–85%) perform best, while 95% causes over-fragmentation. Our adaptive thresholding achieves the strongest results by dynamically calibrating selectivity from retrospective evidence.

5.3.1 Threshold Selection.

We first evaluate different fixed similarity thresholds on the datasets described in Section 4. (Table 4). While moderate thresholds (80–85%) strike a balance between sensitivity and stability, extremely high thresholds (95%) over-fragment the stream, leading to degraded performance. Our *adaptive thresholding module* surpasses all fixed settings by dynamically calibrating sensitivity from retrospective statistics.

5.3.2 Context Length for Adaptive Threshold Learning.

We study how many past frames should be considered when computing the statistics $(\mu, \sigma^2, s_{\text{last}})$ that guide the adaptive threshold network. Intuitively, too short a history may make the threshold overly sensitive to transient noise, while too long a history can dilute the signal of genuine transitions.

Table 5 reports results for contexts of 2–6 frames. A three-frame context provides the best trade-off, yielding the highest overall accuracy. This suggests that three recent frames capture sufficient temporal stability for threshold calibration without introducing excess inertia from distant frames.

Context Length (frames)	0.05	0.1	0.15	0.2	0.25	0.30	0.35	0.4	0.45	0.5	Avg
2	0.610	0.720	0.770	0.780	0.790	0.794	0.800	0.809	0.810	0.815	0.770
3 (Ours)	**0.698**	**0.780**	**0.810**	**0.820**	**0.850**	**0.859**	**0.863**	**0.867**	**0.870**	**0.873**	**0.829**
4	0.690	0.770	0.800	0.807	0.815	0.820	0.824	0.829	0.832	0.834	0.792
5	0.690	0.771	0.798	0.805	0.810	0.812	0.819	0.820	0.825	0.829	0.788
6	0.689	0.760	0.790	0.799	0.802	0.805	0.810	0.814	0.824	0.825	0.782

Table 5: Effect of context length on adaptive threshold learning. A 3-frame context yields the strongest performance and is adopted in our framework.

In our final design, we therefore fix the context length to three frames and use the resulting statistics $(\mu, \sigma^2, s_{\text{last}})$ as input to the adaptive threshold network. This ensures causal operation, avoids reliance on future frames, and provides a stable yet responsive signal for boundary detection.

5.3.3 Scene and Action Understanding.

To evaluate the effect of semantic reasoning, we compare place and action recognition with and without structured representations. Removing **Scene Graph + Caption** reasoning substantially degrades retrieval accuracy (Tables 6, 7). Traditional methods underperform because they rely on raw appearance features and lack semantic abstraction. Visual similarity is brittle to lighting, viewpoint, and clutter, while MMAction struggles with ambiguous egocentric activities. In contrast, **Scene Graph + Captions** capture objects, spatial context, and interactions, leading to higher accuracy and interpretability.

Component-Wise Ablation. We further ablate the Semantic Encoder, **Perceptual Encoder**, and **Comparator**. Table 8 shows that removing any component substantially reduces performance, confirming their complementary roles. The Semantic Encoder captures abstract concepts and aligns events at a high level; the Perceptual Encoder provides spatial grounding; and the Comparator directly detects fine-grained frame-to-frame changes. Removing any module causes systematic degradation: without semantics, conceptual shifts are missed; without perceptual encoding, spatial structure is lost; without comparison, fine transitions cannot be localized. Their synergy yields the most robust and generalizable segmentation.

Method	Validation	
	Avg. mAP	R1@0.5
Visual Similarity	25.16	46.18
Caption + Scene Graph	**35.20**	**57.10**

Table 6: **Place recognition.** Structured reasoning via captions and scene graphs improves retrieval over raw visual similarity.

Method	Validation	
	Avg. mAP	R1@0.5
MMAction	15.95	36.90
Caption + Scene Graph	**34.56**	**55.98**

Table 7: **Action recognition.** Structured reasoning significantly outperforms MMAction baselines.

Semantic	Perceptual	Comparator	F1@10	F1@25	F1@50
–	✓	✓	0.792	0.803	0.803
✓	✓	–	0.710	0.790	0.798
✓	–	✓	0.702	0.780	0.792
✓	–	–	0.640	0.740	0.770
–	✓	–	0.680	0.720	0.750
–	–	✓	0.580	0.650	0.670
✓	✓	✓	**0.830**	**0.836**	**0.840**

Table 8: **Component-wise ablation.** All three modules are necessary and complementary for robust segmentation.

5.4 Weight Sensitivity Analysis of Similarity Fusion

Finally, we analyze the robustness of multimodal fusion weights, which balance semantic and perceptual similarity cues. To avoid overfitting, weights are constrained to $[0.1, 0.4]$, preventing dominance by any single metric.

Configuration	CLIP Radford et al. (2021b)	DINOv2 Caron et al. (2021)	LPIPS Alom et al. (2018)	SSIM	Token Sim.	F1 Score
Selected	0.2	0.3	0.2	0.2	0.1	**0.88**
Equal Weights	0.2	0.2	0.2	0.2	0.2	0.84
High CLIP	0.4	0.1	0.2	0.2	0.1	0.79
High DINOv2	0.1	0.4	0.2	0.2	0.1	0.85
High LPIPS	0.2	0.2	0.4	0.1	0.1	0.82
No Token Sim.	0.25	0.3	0.2	0.25	0.0	0.85
High Token Sim.	0.15	0.25	0.15	0.15	0.3	0.80

Table 9: **Weight sensitivity.** Balanced weighting avoids dominance and achieves robust performance, with the selected configuration yielding the strongest results.

Equal weighting is competitive but suboptimal. Carefully differentiating weights improves performance: CLIP excels at global semantics but should not dominate, DINOv2 contributes strong spatial alignment, LPIPS and SSIM capture low-level perceptual differences, and token similarity provides auxiliary linguistic cues. Eliminating token similarity causes only minor degradation, confirming its supportive but non-essential role. Our final configuration achieves the best balance, integrating global semantics with fine-grained perceptual fidelity for robust segmentation.

6 Conclusion

We proposed a general framework for event segmentation that combines adaptive thresholding with multimodal retrospective consolidation. Our design enables causal operation, requiring only past context, and produces stable, semantically coherent episodes rather than fragmented frame-level transitions. Across ADL-GEBD and Ego4D, the framework achieves state-of-the-art performance, surpassing both unsupervised and heavily supervised baselines.

Although inspired by cognitive theories of episodic memory, our contributions are broadly applicable to machine learning: (i) adaptive thresholding as a label-efficient mechanism for robust boundary detection, and (ii) multimodal consolidation as a scalable strategy for aligning semantic, perceptual, and linguistic cues. A limitation of the current work is that it includes relatively limited analysis of dialogue-driven structure, which can be critical in conversation-heavy or instructional videos. Future work will focus on integrating more sophisticated discourse-level dialogue modeling and exploring interpretable decompositions of modality interactions. This work focuses causal segmentation for long-form video understanding, with implications for robotics, assistive AI, and embodied agents.

ACKNOWLEDGMENTS

The authors used a large language model (ChatGPT) solely to polish grammar and improve the clarity of writing. All research ideas, experiments, analyses, and conclusions are entirely the work of the authors.

REFERENCES

Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S. Awwal, and Vijayan K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches, 2018. URL https://arxiv.org/abs/1803.01164.

Kamel Aouaidjia, Wenhao Zhang, Aofan Li, and Chongsheng Zhang. Improving action segmentation via explicit similarity measurement, 2025. URL https://arxiv.org/abs/2502.10713.

Julia Dietlmeier Kevin McGuinness Alan F Smeaton Ayush K Rai, Tarun Krishna and Noel E O'Connor. Motion aware self-supervision for generic event boundary detection. *In* Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2728–2739, 2023.*, 2728.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Brandon Castellano. Pyscenedetect: Intelligent scene cut detection and video splitting tool, 2018.

Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, Zhiyu Zhao, Junting Pan, Yifei Huang, Zun Wang, Jiashuo Yu, Yinan He, Hongjie Zhang, Tong Lu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo-ego4d: A pack of champion solutions to ego4d challenges, 2022a.

Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022b.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers, 2022c. URL https://arxiv.org/abs/2212.09058.

Dexiang Hong Yufei Wang Libo Zhang Tiejian Luo Congcong Li, Xinyao Wang and Longyin Wen. Structured context transformer for generic event boundary detection. arXiv preprint arXiv:2206.02985*, 2022.*, 2206.

Longyin Wen Dexiang Hong Tiejian Luo Congcong Li, Xinyao Wang and Libo Zhang. End-to-end compressed video representation learning for generic event boundary detection. *In* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13967–13976, 2022.*, 1396.

Longyin Wen Xinyao Wang Dexiang Hong, Congcong Li and Libo Zhang. Generic event boundary detection challenge at cvpr 2021 technical report: Cascaded temporal attention network (castanet). arXiv preprint arXiv:2107.00239*, 2021.*, 2107.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. *In* Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003.*, 2003.

Sourabh Vasant Gothe, Vibhav Agarwal, Sourav Ghosh, Jayesh Rajkumar Vachhani, Pranay Kashyap, and Barath Raj Kandur. What's in the flow? exploiting temporal motion cues for unsupervised generic event boundary detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6926–6935. IEEE, January 2024. doi: 10.1109/wacv57701.2024.00679. URL http://dx.doi.org/10.1109/WACV57701.2024.00679.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

Suha Kwak Heeseung Kwon, Manjin Kim and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. *In* Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 345–362. Springer, 2020.*, 2020.

Minh-Quan Ho-Le, Duy-Khang Ho, Van-Tu Ninh, Cathal Gurrin, and Minh-Triet Tran. Lsc-adl: An activity of daily living (adl)-annotated lifelog dataset generated via semi-automatic clustering, 2025. URL https://arxiv.org/abs/2504.02060.

Kyungmin Kim Taehyun Kim Hyolim Kang, Jinwoo Kim and Seon Joo Kim. Winning the cvpr'2021 kinetics-gebd challenge: Contrastive learning approach. arXiv preprint arXiv:2106.11549*, 2021.*, 2106.

Taehyun Kim Hyolim Kang, Jinwoo 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*, pages 20073–20082, 2022.*, 2007.

Chen Qian Wayne Wu Jiaqi Tang, Zhaoyang Liu and Limin Wang. Progressive attention on multi-level dense difference maps for generic event boundary detection. *In* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3355–3364, 2022.*, 3355.

Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco : Unsupervised boundary contrastive learning for generic event boundary detection, 2021. URL https://arxiv.org/abs/2111.14799.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. *In* IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.*, 1981.

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

Sicheng Mo, Fangzhou Mu, and Yin Li. A simple transformer-based model for ego4d natural language queries challenge. *arXiv preprint arXiv:2211.08704*, 2022.

Tan Nguyen, Jo Etzel, Matthew Bezdek, and Jeffrey Zacks. Multiple event segmentation mechanisms in the human brain, 07 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Liangzhe Yuan Boqing Gong Ting Liu Matthew Brown Serge Belongie Ming-Hsuan Yang Hartwig Adam Rui Qian, Yeqing Li and Yin Cui. Exploring temporal granularity in self-supervised video representation learning. arXiv preprint arXiv:2112.04480*, 2021.*, 2112.

Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation, 2021. URL https://arxiv.org/abs/2101.10511.

Rishabh Khurana Sourabh Vasant Gothe, Jayesh Rajkumar Vachhani and Pranay Kashyap. Self-similarity is all you need for fast and light-weight generic event boundary detection. *In* ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.*, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023.

Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, and Limin Wang. Progressive attention on multi-level dense difference maps for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3355–3364, June 2022.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*.

Endel Tulving. Episodic memory: From mind to brain. *Annual review of psychology*, 53:1–25, 02 2002. doi: 10.1146/annurev.psych.53.100901.135114.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Univse: Robust visual semantic embeddings via structured semantic representations, 2019. URL https://arxiv.org/abs/1904.05521.

Tao Mei Xiao Wang, Jingen Liu and Jiebo Luo. Coseg: Cognitively inspired unsupervised generic event segmentation. arXiv preprint arXiv:2109.15170*, 2021.*, 2109.

A Appendix

A.1 Theoretical Properties of the Adaptive Threshold Mechanism

At each time step t, we compute a summary vector

$$s_t = (\mu_t, \sigma_t^2, s_{\text{last},t}),$$

where μ_t is the mean similarity to recent keyframes, σ_t^2 is the variance of similarities, and $s_{\text{last},t}$ is the similarity to the immediately 3 previous frames. The adaptive threshold network, parameterized by θ, predicts

$$\tau_\theta(s_t) \in [\tau_{\min}, \tau_{\max}],$$

and the probability of a boundary at time t is

$$p_\theta(t) = \sigma\big(\alpha(\tau_\theta(s_t) - \mu_t) + \beta\big),$$

where $\sigma(\cdot)$ is the logistic function, α controls sharpness, and β is a learnable bias.

Proposition 1 (Causality)**.** *The boundary probability $p_\theta(t)$ depends only on past observations $X_{\leq t}$.*

Proof. The features s_t are computed from past keyframes $X_{\leq t}$ only. Hence $p_\theta(t)$ and the boundary decision $d_t = \mathbf{1}\{p_\theta(t) > \text{threshold}\}$ are causal, with no access to future frames $X_{>t}$. □

Proposition 2 (Non-degeneracy). *Consider the training loss*

$$L(\theta) = \lambda_{sup} L_{sup}(\theta) + \lambda_{rate}\big(\mathbb{E}_t[p_\theta(t)] - \rho\big)^2 + \lambda_{reg}\|\theta\|^2,$$

where $\rho \in (0,1)$ is a target boundary frequency. Then any minimizer $\theta^\star$ satisfies

$$\big|\mathbb{E}_t[p_{\theta^\star}(t)] - \rho\big| \leq \frac{L(\theta^\star)}{\lambda_{rate}},$$

which prevents collapse to trivial all-zero or all-one predictions.

Sketch. By definition of L, we have $\lambda_{\text{rate}}(\mathbb{E}_t[p_{\theta^\star}(t)] - \rho)^2 \leq L(\theta^\star)$. Rearranging gives the bound. □

Proposition 3 (Adaptivity). *The function $\tau_\theta(s_t)$ adjusts the decision threshold according to summary statistics. When similarities are stable (low variance σ_t^2), even small deviations in μ_t can trigger boundaries; when context is noisy (high σ_t^2), the threshold adapts upward to avoid false positives.*

Intuition. Because τ_θ maps $(\mu_t, \sigma_t^2, s_{\text{last},t})$ to a bounded threshold, its output varies with contextual stability. Thus the mechanism is robust to both steady and noisy regimes. □

Together, these properties show that the adaptive threshold mechanism is causal, avoids degenerate behavior, and adapts dynamically to context.

A.2 Dynamic Consolidated Window Theoretical Justification.

The DCW acts as a local temporal coherence constraint: a boundary is valid only if it coincides with a semantic discontinuity. Formally, let $\mathcal{S}(I)$ denote semantic context (scene, place, action). Then a retained boundary must satisfy

$$\mathcal{S}(I_{k_i - 1}) \neq \mathcal{S}(I_{k_i}),$$

ensuring that splits occur only when there is a genuine semantic change. This reduces false positives caused by transient low-level variations (e.g., lighting or camera motion) and aligns with the principle that episodic segmentation in cognition occurs at context shifts rather than at every perceptual fluctuation.

A.3 Dialogue-Aware Video Segmentation

While visual discontinuities are a common cue for event boundaries, many real-world videos are **dialogue-driven**, where semantic structure is carried by speech rather than visual change. In such cases—sitcoms, interviews, or instructional tutorials—editing conventions like shot-reverse-shot introduce frequent appearance shifts that do not correspond to genuine narrative transitions. A purely visual method therefore risks fragmenting coherent dialogue into artificial segments.

A.3.1 Dialogue-Aware Refinement

Figure 3 illustrates this misalignment: visual boundaries (green) derived from frame-level changes often occur mid-utterance, while the underlying dialogue (red) remains continuous. This leads to segmentation that splits coherent discourse units, breaking narrative flow and weakening downstream applications such as summarization, question answering, or episodic memory modeling.

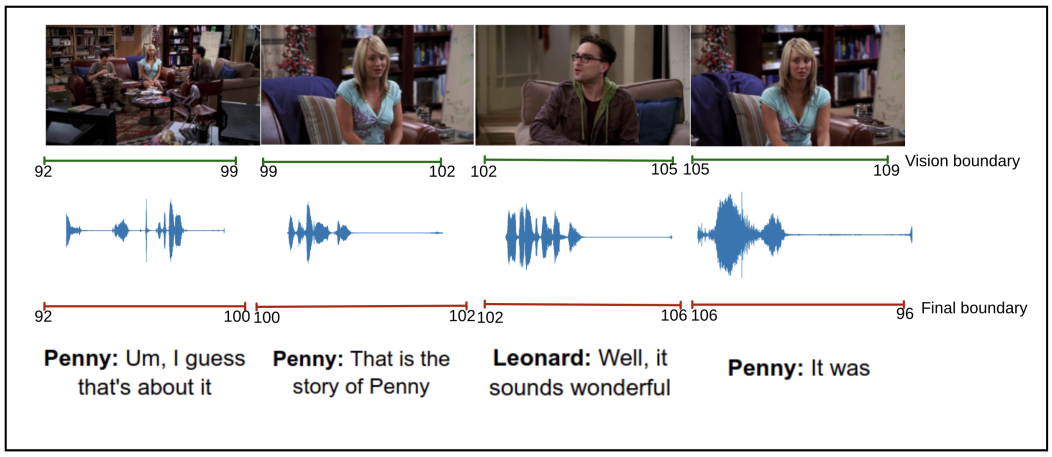

Figure 3: Visual boundaries (green) frequently misalign with dialogue structure (red), fragmenting continuous speech. Our dialogue-aware refinement defers segmentation until utterances end, ensuring audio-visual coherence in narrative-driven content.

To address this, we introduce a **dialogue-aware refinement step** that aligns event boundaries with acoustic continuity. We extract speech segments using Mel-Frequency Cepstral Coefficients (MFCCs) and higher-level prosodic embeddings such as BEAT Chen et al. (2022c). If the audio stream indicates ongoing speech across a visual boundary, segmentation is deferred until the utterance completes. This simple adjustment ensures that:

- Dialogue remains intact within a single segment, preserving discourse continuity;
- Adjacent visual segments with uninterrupted speech are merged;
- Final event boundaries reflect both visual structure and linguistic flow.

A.4 Why Dialogue-Aware Refinement Matters

This step highlights a broader principle: **multimodal event segmentation must respect linguistic as well as visual coherence**. In dialogue-heavy domains, speech—not motion—defines the natural unit of experience. By fusing acoustic and visual cues, our framework produces segments that align more closely with human perception of episodes, strengthening its utility for narrative understanding, summarization, and episodic memory grounding.