Supplementary Materials of VideoScaffold: Elastic-Scale Visual Hierarchies for Streaming Video Understanding in MLLMs

This supplementary document is organized as follows:

Sec. 1 presents a comparative analysis of clustering-based and adjacent-similarity segmentation methods, highlighting their intrinsic limitations when contrasted with the prediction-driven event structures produced by EES.

Sec. 2 provides visualizations of EES-generated event hierarchies, demonstrating how prediction-guided segmentation produces temporally coherent and multi-granular event structures across diverse videos.

Sec. 3 presents qualitative case studies demonstrating how VideoScaffold forms coherent event structures and captures critical semantic transitions, enabling accurate reasoning where baseline MLLMs fail.

Sec. 4 offers ablation studies on the Elastic-Scale Event Segmentation (EES) module, analyzing the impact of hierarchical evolution and prediction execution space on constructing stable multi-level temporal representations.

Sec. 5 examines two key hyperparameters of EES—the number of event layers and the segmentation threshold ε—and assesses their influence on event granularity, temporal coherence, and long-range semantic abstraction.

Sec. 6 reports ablation experiments on the Hierarchical Event Consolidation (HEC) module, evaluating alternative aggregation strategies and essential-frame selection methods to identify how events are most effectively summarized for downstream reasoning.

1. Comparison of Event Segmentation Methods

To further evaluate the structural benefits of our event representation, we compare the *intra-event* and *inter-event* similarity produced by VideoScaffold against two representative segmentation baselines. All experiments are conducted on 100 videos (each uniformly sampled to 120 frames) from the MLVU [2] benchmark.

1.1. Baseline Segmentation Methods

K-means Token Clustering. Following Chat-UniVi [1], we adopt the DPC-KNN clustering strategy to merge spatial visual tokens extracted from a ViT encoder. Given patch tokens $\mathbf{Z} = \mathbf{z}ii = 1^L$, local token density is computed as

$$\rho_i = \exp!\left(-\frac{1}{K}\sum_{z_k \in \mathrm{KNN}(z_i,\mathbf{Z})} \|z_k - z_i\|^2\right), \tag{1}$$

and the corresponding distance index is defined as

$$\delta_i = \begin{cases} \min\limits_{j:\,\rho_j > \rho_i} \|z_j - z_i\|^2, & \text{if such } j \text{ exists,} \\ \max\limits_{j} \|z_j - z_i\|^2, & \text{otherwise.} \end{cases} \tag{2}$$

Tokens with large $\rho_i \cdot \delta_i$ values are selected as cluster centers, and all remaining tokens are assigned to their nearest center. Temporal events are then determined by grouping frames with similar cluster distributions. We report results under $k \in 8, 12, 16$ to cover both coarse and fine clustering regimes.

Adjacent-Frame Similarity Thresholding. The second baseline segments videos using cosine similarity between adjacent frame embeddings:

$$\mathrm{sim}(t, t+1) = \frac{\mathbf{v}t^\top \mathbf{v}t + 1}{\|\mathbf{v}t\|, \|\mathbf{v}t + 1\|}, \tag{3}$$

and creates an event boundary whenever

$$\mathrm{sim}(t, t+1) < \tau, \qquad \tau \in 0.60, 0.70, 0.80. \tag{4}$$

Small thresholds collapse multiple scenes into one event, while large thresholds lead to fragmented, unstable segmentation.

1.2. Intra-Event and Inter-Event Similarity Metrics

Given an event segment S, its intra-event similarity is defined as

$$\mathrm{Intra}(S) = \frac{1}{|S|(|S|-1)} \sum_{\substack{i \neq j \\ \mathbf{v}_i, \mathbf{v}_j \in S}} \cos(\mathbf{v}_i, \mathbf{v}_j), \tag{5}$$

which measures the semantic coherence within an event.

For two different event segments S_a and S_b, the inter-event similarity is

$$\mathrm{Inter}(S_a, S_b) = \frac{1}{|S_a||S_b|} \sum_{\mathbf{v}_i \in S_a} \sum_{\mathbf{v}_j \in S_b} \cos(\mathbf{v}_i, \mathbf{v}_j), \tag{6}$$

where lower values indicate better semantic separation.

We further compute the stability ratio

$$\Gamma = \frac{\mathrm{Intra}}{\mathrm{Inter}}, \tag{7}$$

with larger Γ indicating stronger event discrimination.

1.3. Evaluation Protocol

We conduct all evaluations on a set of 100 videos from the MLVU [2] benchmark, each uniformly sampled to 120 frames to ensure consistent temporal coverage. All methods employ the EVA-CLIP vision encoder for feature extraction, allowing fair comparison across segmentation strategies. For our approach, VideoScaffold, we adopt a 3-layer EES configuration with a segmentation threshold of $\varepsilon = 0.4$.

Baseline implementations include multiple clustering sizes k for the Chat-UniVi k-means pipeline, as well as a range of adjacent-frame similarity thresholds τ for threshold-based segmentation. For each video, we compute the average intra-event similarity, inter-event similarity, and the resulting stability ratio Γ. This protocol provides a controlled and comprehensive assessment of how different segmentation paradigms organize temporal structure and preserve semantic continuity.

1.4. Benchmark Results

Table 1. Comparison of intra-event similarity, inter-event similarity, and stability ratio Γ on 100 MLVU videos (120 frames each). VideoScaffold achieves the strongest event cohesion and separation.

Method	**Intra Sim.** $\uparrow$	**Inter Sim.** $\downarrow$	$\mathbf{\Gamma}$ $(\uparrow)$
K-Means (k=8)	0.55	0.41	1.34
K-Means (k=12)	0.58	0.39	1.49
K-Means (k=16)	0.59	0.38	1.55
Similarity (τ=0.6)	0.63	0.37	1.70
Similarity (τ=0.7)	0.66	0.35	1.89
Similarity (τ=0.8)	0.68	0.34	2.00
VideoScaffold (Ours)	**0.71**	**0.32**	**2.22**

Intra-event similarity. Across all 100 evaluation videos, VideoScaffold achieves the highest intra-event similarity, indicating that the frames grouped within each event remain highly consistent in visual semantics. This advantage stems from the prediction-guided boundary selection in EES, which aligns event segmentation with true scene transitions rather than heuristic distance thresholds.

Inter-event similarity. VideoScaffold also obtains the lowest inter-event similarity, demonstrating that its segmented events are well separated and semantically distinct. By contrast, clustering with small k frequently merges heterogeneous scenes, while threshold-based segmentation with aggressive cutoffs introduces noisy and unstable boundaries.

Stability ratio Γ**.** VideoScaffold further yields the highest stability ratio Γ, confirming its ability to jointly maximize within-event cohesion and between-event separation. Baselines based on clustering or adjacent similarity show substantial variance across videos, whereas VideoScaffold maintains stable performance across diverse scenes and temporal dynamics.

2. Visualization of Elastic-Scale Event Structures

Figure 1 provides qualitative results of the proposed Elastic-Scale Event Segmentation (EES) module across a diverse set of videos from the MLVU benchmark. Each example visualizes the full three-layer hierarchical event structure inferred from 60–120 frames, illustrating how EES adaptively adjusts its segmentation granularity as visual content evolves over time.

At the bottom layer (Level 1), EES produces fine-grained segments that respond sensitively to local temporal fluctuations. As the hierarchy ascends, Level 2 consolidates short-term segments into semantically coherent mid-level events, effectively smoothing local noise while preserving meaningful boundaries. At the highest layer (Level 3), the model captures stable, abstract event units that summarize major transitions across long temporal spans.

EES robustly adapts to diverse temporal dynamics. Videos with rapid scene changes exhibit denser boundary placement in lower layers, while slower-paced or more stable content results in longer, smoother segments. Crucially, the hierarchical structures remain well aligned across layers, demonstrating consistent multi-scale abstraction and strong temporal coherence.

These visualizations further validate the key properties of EES: elastic granularity, stable multi-level representations, and the ability to capture both fine-grained details and long-range event structure in streaming video. They also explain why EES achieves stronger intra-event cohesion and clearer inter-event separation compared to clustering- or similarity-based segmentation baselines.

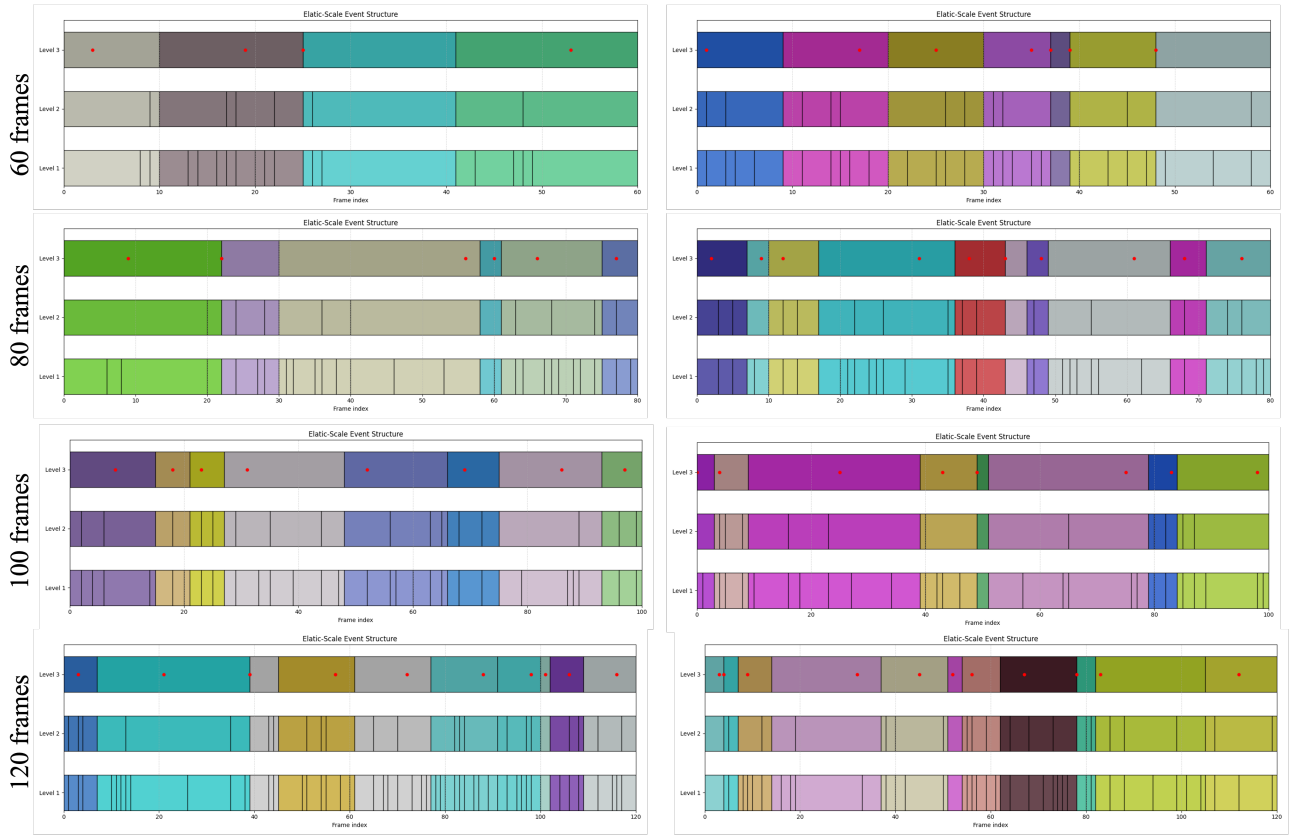


Figure 1. Additional visualizations of the Elastic-Scale Event Segmentation (EES) across diverse MLVU videos. Each example shows three hierarchical levels of events inferred from 60–120 streaming frames. Red markers indicate essential frames selected via prediction-error peaks. The results illustrate how EES adapts event granularity to content dynamics, producing coherent multi-scale structures with stable and semantically aligned boundaries.

3. Case Study

Both case studies further illustrate how VideoScaffold improves multimodal understanding by constructing coherent event structures and preserving key semantic transitions. In the VideoMME example (Fig. 2), VideoScaffold segments the video into meaningful event units and selects informative essential frames, enabling the model to capture the progression of visual concepts and infer the correct answer. Baseline methods, lacking hierarchical abstraction, either over-fragment or over-smooth the temporal sequence, leading to incorrect predictions. In the streaming scenario (Fig. 3), the challenge becomes even more pronounced due to causal constraints and long, continuous inputs. Flash-VStream and VideoLLM-online struggle to track the speaker's evolving topic, whereas VideoScaffold maintains stable event boundaries, anchors reasoning on decisive evidence frames, and anticipates the next topic accurately.

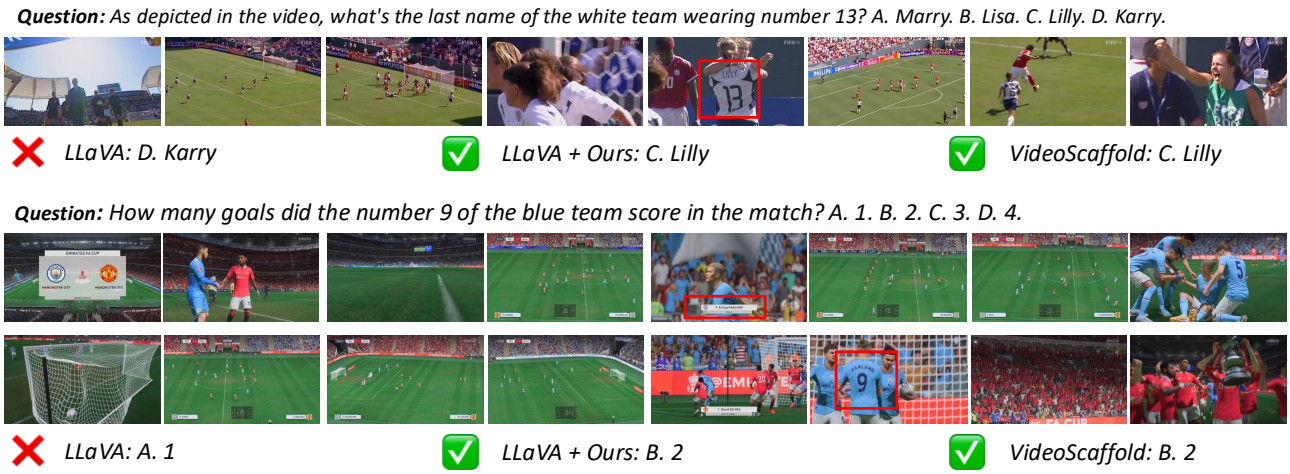


Figure 2. Case study from the VideoMME benchmark. The example demonstrates how VideoScaffold segments the video into semantically meaningful events and identifies essential frames that anchor hierarchical reasoning. By preserving critical scene transitions and suppressing redundant frames, VideoScaffold enables more accurate multimodal understanding and yields the correct answer, whereas baseline methods fail due to fragmented or noisy event structures.

4. Ablation Studies of Elastic-Scale Event Segmentation

To rigorously evaluate the proposed Elastic-Scale Event Segmentation (EES) module, we conduct extensive ablation studies examining both its hierarchical design and the execution space of its next-frame prediction mechanism. These analyses clarify how different architectural and operational choices shape segmentation behavior and influence downstream performance.

Table 2. Ablation on Elastic-Scale Event Segmentation.

Elastic	Space	Token	ActivityNet	MSVD	MSRVTT	LV-Bench
×	*Latent*	*Patch*	44.3	71.0	56.8	27.3
✓	*Pixel*	-	45.5	70.4	56.1	28.2
✓	*Latent*	*Cls*	45.2	70.2	55.5	27.8
✓	***Latent***	***Patch***	**48.9**	**72.5**	**58.4**	**31.5**

Effect of Removing the Elastic Hierarchy. We begin by assessing the role of hierarchical abstraction by disabling multi-level evolution and constraining EES to operate at a single fine-grained level. This variant removes the ability to progressively summarize temporal structure, effectively reducing the system to a uniform segmentation scheme.

As shown in Table 2, short-video benchmarks such as MSVD-QA and MSRVTT-QA exhibit only minor performance changes under this configuration due to their limited temporal complexity. However, the degradation becomes pronounced on long-form datasets, including LV-Bench and ActivityNet-QA. Without elastic hierarchical evolution, the model produces excessively fragmented segments, leading to redundant visual tokens, disrupted temporal coherence, and impaired long-range

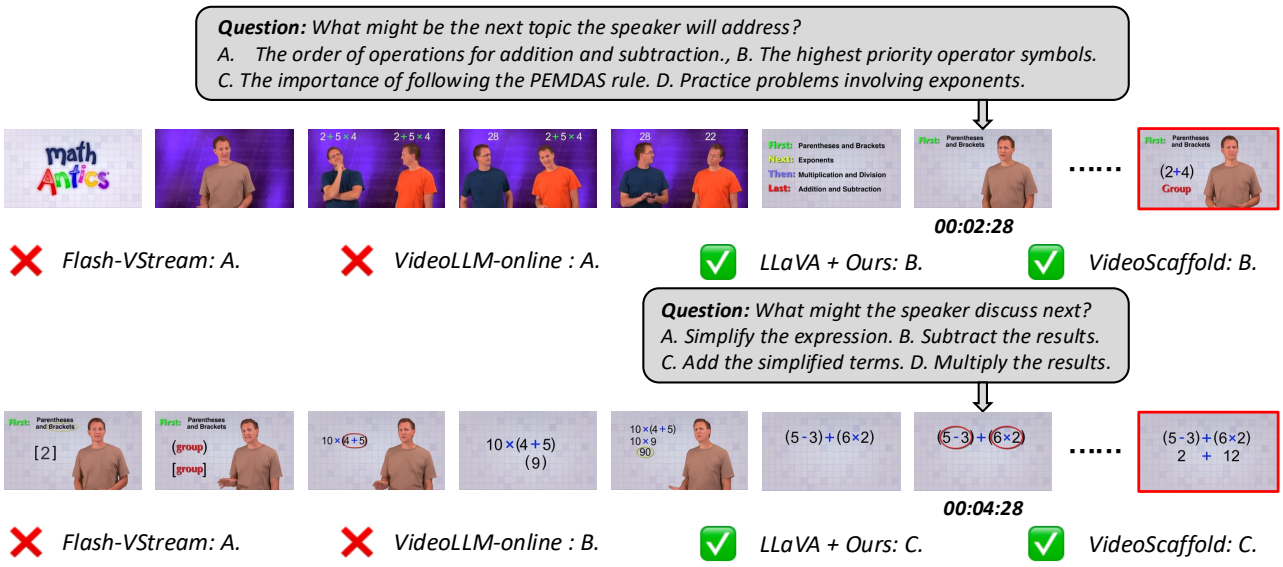


Figure 3. Case study on StreamingBench illustrating VideoScaffold's advantages in real-time video understanding. Given long streaming inputs, baseline streaming MLLMs fail to anticipate the speaker's next topic due to fragmented or noisy temporal representations. By contrast, VideoScaffold identifies the correct semantic progression, and consistently produces the correct answer. Highlighted frames (red boxes) show key evidence segments captured by our elastic-scale event representation, which enables accurate next-topic prediction even in extended instructional videos.

reasoning. These observations highlight that the hierarchical structure of EES is essential for capturing multi-scale temporal patterns and preserving semantic continuity in extended video sequences.

Effect of Prediction Execution Space. We further examine how the operating space of the next-frame prediction mechanism affects segmentation sensitivity and stability. Four variants are evaluated: pixel-space prediction, latent-space prediction using class tokens, latent-space prediction using patch tokens, and the non-hierarchical baseline.

As shown in Table 2, prediction in pixel space performs the worst, as it is highly sensitive to noise and low-level frame fluctuations, leading to unstable and unreliable boundary signals. In contrast, latent-space prediction produces consistently stronger results, indicating that high-level encoder features provide a more semantically grounded and robust basis for detecting temporal changes. Among the latent-space variants, prediction using patch tokens achieves the best performance. Patch tokens retain fine-grained spatial structure, enabling the model to accurately capture subtle semantic transitions across frames. Class tokens, by comparison, compress spatial information into a single global representation, weakening boundary discrimination and resulting in inferior segmentation outcomes.

Overall, these findings validate our design choice of performing prediction in the latent space with patch tokens, and underscore the complementary roles of semantic feature modeling and hierarchical abstraction in achieving reliable elastic-scale event segmentation.

5. Ablation Studies of Event Layers and Threshold

To gain a deeper understanding of the behavior of the proposed Elastic-Scale Event Segmentation (EES) module, we conduct comprehensive ablation studies on two essential hyperparameters: (1) the number of event layers, which determines the depth of semantic abstraction, and (2) the segmentation threshold ε, which controls the temporal granularity of detected boundaries. These factors fundamentally shape the hierarchical segmentation structure and directly influence EES's ability to adapt to videos with varying duration and content complexity.

Number of Event Layers. We begin by analyzing how hierarchical depth affects event abstraction quality. The number of layers is varied from two to four, and performance is evaluated across both short-video and long-video benchmarks. As shown in Table 3, increasing the hierarchy from one to three layers consistently boosts performance, confirming that progressive abstraction enhances temporal modeling: lower layers capture rapid frame-to-frame variations, while higher layers encode broader semantic transitions and longer-range structure. However, extending the hierarchy to four layers yields

only marginal gains while introducing additional computation and latency. The diminishing improvement suggests that overly deep hierarchies may introduce redundant or excessively smoothed representations that no longer benefit downstream reasoning. Overall, a three-layer hierarchy provides the most effective balance between abstraction capacity, representational compactness, and computational overhead.

Segmentation Threshold ε. We further examine how the boundary threshold regulates segmentation dynamics. Because ε controls the sensitivity of prediction-error–based boundary detection, it determines the temporal resolution of event segmentation. As shown in Table 3, smaller thresholds (e.g., 0.3) trigger boundaries too frequently, generating over-fragmented event structures that emphasize local variations while disrupting temporal coherence. Such short segments reduce cross-frame reasoning stability and increase aggregation costs. In contrast, larger thresholds (e.g., 0.5–0.6) suppress boundary formation and merge multiple scenes into coarse segments, blurring semantic transitions and weakening discriminability. The best performance occurs near $\varepsilon = 0.4$, where segmentation preserves meaningful scene changes while maintaining stable and coherent temporal grouping. This moderate setting provides a balanced compromise between fine-grained sensitivity and high-level structural stability.

Together, these ablations confirm that the interplay between event-layer depth and threshold selection is critical for producing compact, expressive, and temporally consistent event hierarchies. A three-layer structure combined with a moderate threshold yields the most robust and generalizable configuration across diverse video durations and content types, enabling EES to operate effectively in both short- and long-form streaming scenarios.

Table 3. Ablation on event layers and segmentation threshold.

Layers	ε	ActivityNet	MSVD	MSRVTT	LV-Bench
3	0.3	48.1	71.3	57.6	31.2
3	**0.4**	**48.9**	**72.5**	**58.4**	**31.5**
3	0.5	47.6	70.8	56.6	30.8
3	0.6	47.0	69.7	56.2	30.3
2	0.4	44.3	68.6	53.9	27.7
4	0.4	48.4	72.9	58.7	31.6

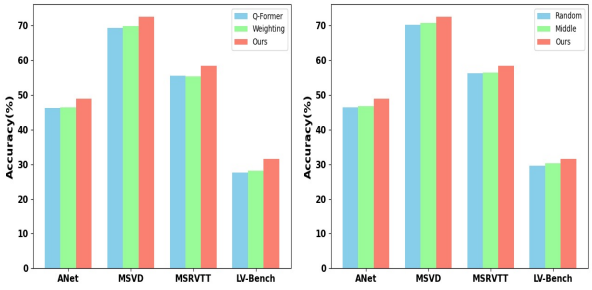
6. Ablation Studies of Hierarchical Event Consolidation

The Hierarchical Event Consolidation (HEC) module integrates multi-level event segments into compact, semantically enriched representations for downstream reasoning tasks. Its performance hinges on two core components: (1) the mechanism used to aggregate event content, and (2) the choice of the essential frame that serves as the semantic anchor for each event. Suboptimal aggregation may wash out key visual cues, while poor anchor selection can misalign the abstraction process. To fully understand these effects, we conduct ablation studies on both aspects.

Aggregation Mechanisms. To analyze how different consolidation strategies influence hierarchical abstraction, we compare three formulations within the HEC module: Q-former–based aggregation, distance-based weighting, and our proposed cross-attention mechanism. As shown in Fig. 4(a), the Q-former baseline relies on a single learnable query shared across all events. Although this approach is effective for global summarization, it fails to adapt to the heterogeneous visual content found across event segments, resulting in representations that lack event-level specificity and semantic depth. The distance-based weighting strategy aggregates features purely according to temporal closeness. However, by ignoring spatial structure and semantic relevance, it often collapses multiple visual patterns into overly coarse representations, diminishing the distinctiveness of each event.

In contrast, our cross-attention–based formulation directly leverages spatial tokens from the essential frame to guide the consolidation process. By allowing the essential token to attend selectively to salient regions within the event, this design ensures that the resulting representation emphasizes content that is truly characteristic of the event's semantics. This targeted, content-aware summarization leads to more discriminative, coherent, and faithful event embeddings, and consistently yields the highest downstream performance. These results confirm that event-conditioned attention provides a more flexible and semantically aligned mechanism for hierarchical representation construction.

Essential Element Selection. An important component of the HEC pipeline is the selection of an *essential frame*, which serves as the semantic anchor for bottom-up event aggregation. We compare three strategies for this selection: (1) random sampling, (2) middle-frame selection, and (3) our proposed maximum prediction-error criterion. As shown in Fig. 4(b), random sampling introduces substantial variance and frequently chooses frames that fail to capture meaningful semantic



(a) Essential Frame Selection **(b) Event consolidation**

Figure 4. Ablation on essential frame selection and consolidation.

transitions. Middle-frame selection provides slightly more stability but still tends to overlook the most informative moment within an event, particularly when scene changes occur near the event boundaries.

In contrast, selecting the frame with the highest prediction error—directly supplied by the EES segmentation dynamics—consistently achieves superior performance. Frames exhibiting the largest prediction error correspond to moments where the model's expectation diverges most strongly from the observed input, which typically align with structurally or semantically significant transitions. Using such frames as anchors ensures that the hierarchical aggregation process is guided by the most salient and representative cues within each event. These results demonstrate the effectiveness of coupling prediction-error signals from EES with the consolidation operations in HEC, enabling more faithful and context-aware event-level summarization.

References

[1] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1

[2] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 1, 2