Beyond Highlight Detection: Most-Replayed Driven Multimodal Analysis of Korean YouTube Videos for Highlight Editing Guidance

Chanhee Lee*†, Jinho Jang*, Sungjun Ha, Jinwoong Jung, Mina Jung‡

Applied Artificial Intelligence, Sungkyunkwan University

Team: You-Ha

{leechanhye, jangjinho65, hhssjj0521, wjdwlsdndpp, minajung}@skku.edu

* Equal contribution † Team Leader ‡ Corresponding Author

June 14, 2026

Abstract

This project proposes a framework for practical video editing guidance. Existing highlight detection and video summarization models remain limited for real-world editing, where creators need actionable guidance rather than simple cut-based automation. We construct **KoSum**, a Korean YouTube benchmark for video summarization, consisting of recent videos from 2024 to 2026 collected through a structured protocol across 14 fine-grained content categories (e.g., entertainment and cooking) and annotated with triple-modality signals (e.g., visual, audio, and text) and Most-Replated signals. Using a triple-modality model, we analyze highlight regions through modality attention patterns and feature-level statistics, capturing editing-related multimodal cues (e.g., shot transitions, motion dynamics, subtitle density, and audio novelty). Based on these analyses, we propose a **data-analysis-conditioned prompting framework** that augments user prompts with category-specific and statistically grounded editing cues. The proposed framework outperforms the baseline by **95 points** in total score and **3.17 points** in average score, demonstrating strong effectiveness in generating practical and structured editing guidelines. Our framework connects viewer engagement signals, multimodal analysis, and practical editing support for real-world creators. The project page is available at https://iontail.github.io/kosum/.

1 Introduction

As video creation has become more accessible, video content has rapidly spread through platforms such as YouTube and TikTok. At the same time, users increasingly prefer short form content such as Shorts and Reels, which reflects a growing demand for compact videos that deliver only essential or engaging information. To address this demand, highlight detection and video summarization have gained significant attention. Highlight detection aims to identify segments that attract viewers' attention, while video summarization focuses on selecting the core parts of a video by removing redundant content that does not contribute to understanding.

With the success of Transformer [45], many approaches [28, 32, 2, 25, 10, 22] have been proposed for these tasks using attention mechanisms and, more recently, large language models (LLMs). Although these methods achieve strong predictive performance, they still have limitations when applied to real world video editing. In practice, editing is not simply a matter of selecting important segments and cutting out the rest. Creators need actionable editing guidance that explains what kinds of visual, audio, and textual cues make certain moments engaging. Therefore, beyond automated cut based summarization, it is important to provide editing guidelines that can help creators design videos that better attract viewer attention.

Existing studies have several limitations from this perspective. First, most approaches mainly focus on improving model architecture and benchmark performance, while paying less attention to why certain segments are selected as highlights or summaries. As a result, there is still limited analysis of which video

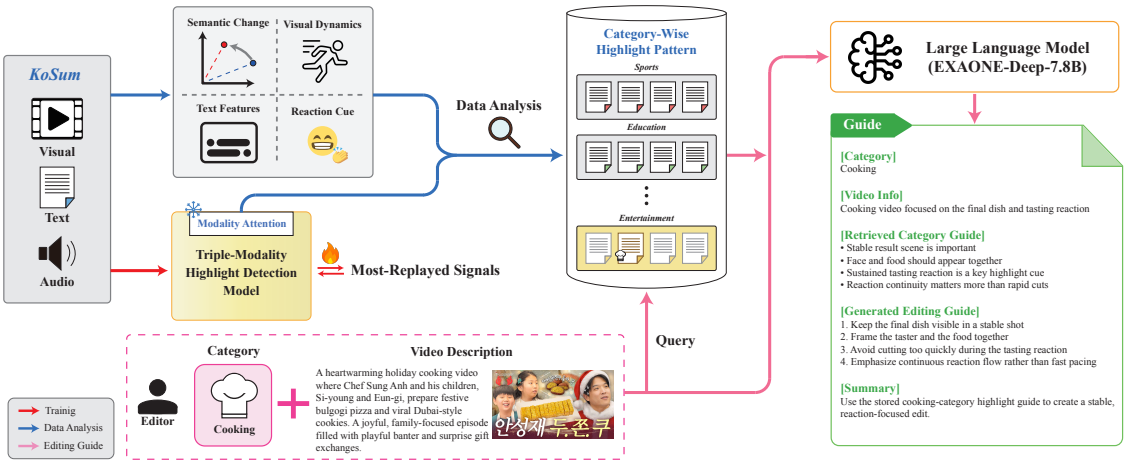


Figure 1: Overall pipeline of our framework. We construct KoSum using YouTube videos and Most Replayed signals, train a triple modality model to predict weak importance scores, and analyze highlight regions using modality attention weights and separately extracted modality features. The resulting category specific analysis is then used in our data analysis conditioned prompting framework to generate structured video editing guidelines with an LLM.

components distinguish important regions from non-important regions. Second, only a few studies [22, 10] leverage triple modality information, namely visual, audio, and textual signals, in a unified framework, which limits a comprehensive understanding of multimodal interactions. Third, existing benchmark datasets are outdated and do not fully reflect recent content consumption trends, especially in Korean target videos. For example, TVSum [40] was last updated in 2019, and datasets such as Mr.HiSum [42] and MoSu [22] are based on YouTube 8M [1], which was constructed from videos uploaded before 2014. Therefore, existing benchmarks have limitations in capturing recent Korean YouTube content and category specific editing patterns. In addition, some videos in YouTube 8M are no longer available, making it difficult to access the original content. The dataset comparison is presented in Table 1. Fourth, category-aware modeling remains underexplored. Editing strategies can vary substantially across content categories such as entertainment, cooking, sports, education, drama, and vlogs. However, most existing approaches rely on query based [28, 32, 41, 26] or caption conditioned models [13, 35, 22, 2, 25] that are content aware but not explicitly category-aware. This makes it difficult to provide tailored editing guidance for different types of videos.

To address these limitations, we use Most-Replayed statistics as a weak supervision signal for estimating temporally localized viewer interest. Rather than treating repeated viewing behavior as a definitive ground truth of video importance, we use it as an indirect signal that reflects which segments are more likely to attract viewer attention. This enables us to identify candidate highlight regions and further analyze their multimodal characteristics across visual, audio, and textual signals.

Although the videos in our dataset are already edited to some extent, distinct highlight regions still emerge as clear peaks in Most Replayed signals. This suggests that even edited videos contain moments that receive substantially higher viewer attention than surrounding regions. Moreover, we restrict video length to 7 to 30 minutes, which provides a suitable balance between structural completeness and manageable temporal complexity. This range allows us to analyze highlight patterns and editing strategies across videos that are long enough to contain meaningful narrative or informational structure, while still being suitable for detailed temporal analysis.

In this work, we shift the perspective of video analysis from viewers to creators. We construct **KoSum** (**Ko**rean YouTube Multimodal Video **Sum**marization), a Korean YouTube benchmark for video summarization and highlight detection. KoSum consists of recent videos from 2024 to 2026, collected through a structured protocol across 14 fine-grained content categories (e.g., entertainment and cooking). Each video is annotated with triple modality signals (e.g., visual, audio, and text) and importance scores derived from Most-Replayed statistics. Following standard benchmark construction protocols [42, 22], KoSum is designed to reflect recent video consumption trends and support category aware analysis of viewer engagement.

Using KoSum, we conduct analysis based on data rather than focusing only on predictive modeling. We first train a model that uses visual, audio, and text signals, and examine its attention weights to understand which

Table 1: Comparison of Video Summarization & Highlight Detection Datasets. KoSum is constructed using recent Most-Replayed signals from 2024 to 2026 and supports multimodal analysis across visual, audio, and text information for Korean content. MP represents the Most-Replayed signals and ≥24 indicates that the data was posted in 2024 or later.

Datasets	Visual	Text	Audio	Caption	Video Lang.	Text Lang.	MP	≥24
SumMe [11]	✓	✗	✗	✗	English	-	✗	✗
TVSum [40]	✓	✓	✗	✗	English	English	✗	✗
MMSum [35]	✓	✓	✗	✓	English	English	✗	✗
Mr.HiSum [42]	✓	✗	✗	✗	English	-	✓	✗
MoSu [22]	✓	✓	✓	✓	English	English	✓	✗
KoSum (ours)	✓	✓	✓	✓	Korean	Korean	✓	✓

modality contributes more strongly to highlight prediction across different categories. Separately, we extract editing features for each modality from the original videos and compare their statistical patterns between highlight regions and non highlight regions. Based on these analyses, we propose a **prompting framework** conditioned on data analysis, which augments user prompts with category specific and statistically grounded editing cues. In our LLM-based evaluation, the proposed framework demonstrates strong performance in generating practical and structured editing guidelines compared with user prompt only generation. The full pipeline is illustrated in Figure 1.
Our contributions are summarized as follows:

- We construct **KoSum**, a new Korean YouTube benchmark for video summarization and highlight detection. The dataset is built from recent videos collected between 2024 and 2026 and uses Most Replayed signals to derive importance scores that reflect current viewer engagement patterns.
- We provide analysis of highlight regions using visual, audio, and text signals. Rather than focusing only on predictive performance, we examine modality attention weights from a triple modality model and separately analyze feature statistics extracted from each modality to understand the characteristics of engaging video segments.
- We propose a **data analysis conditioned prompting framework** for practical video editing guidance. By combining user prompts with category specific and statistically grounded editing cues, the framework generates structured editing suggestions that better support real world creators.

The remainder of this report is organized as follows. Section A reviews related work. Section 2 describes the KoSum dataset construction. Section 3 presents the preprocessing and importance score construction. Section 4 describes the highlight detection framework and experimental results. Section 5 introduces modality specific feature extraction. Section 6 explains the statistical analysis pipeline. Section 7 presents category specific analysis results. Section 8 evaluates the generated editing guidelines. Finally, we discuss limitations and conclude the report.

2 Dataset

2.1 Data Collection and Filtering Criteria

To construct a representative and high quality dataset, we collect widely viewed videos for each content category based on view count. We focus on videos uploaded between 2024 and 2026 to capture recent content consumption trends, and require each video to have at least 50,000 views following prior work [21, 42]. We restrict video length to between 7 and 30 minutes to exclude extremely short or long videos while preserving sufficient temporal structure for highlight analysis. Within this range, we mostly collect videos longer than 8 minutes, which can contain mid roll advertisements, better reflect typical platform behavior, and help capture editing patterns that emerge in longer video structures.
To ensure data quality and consistency, we apply additional filtering criteria. First, we include only videos with available Most Replayed values, as these signals are essential for estimating viewer interest. We further select videos with clearly distinguishable Most Replayed peaks, where at least four prominent highlight regions are observed. Second, we collect only videos with Korean subtitles, including both manually created and automatically generated subtitles, to enable text based feature extraction. Third, we select videos with

Table 2: Detailed statistics of KoSum. Transcript density indicates the average ratio of video duration with valid text.

Statistic Category	Value
Duration Statistics	
Avg. Duration	1019.3 sec
Std. Duration	339.0 sec
Min Duration	384.0 sec
Max Duration	1800.0 sec
Textual Statistics	
Total # of Tokens	2.310M
Avg. # of Tokens/Video	3299.6
Transcript Density	88.18%
Audio Statistics	
Audio Availability	100%

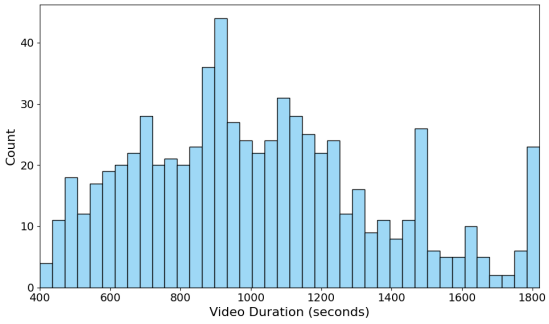


Figure 2: Video duration distribution in the KoSum dataset.

a high proportion of Korean viewers, which is inferred from titles and comment distributions to maintain a consistent linguistic and cultural context. Finally, we exclude videos that compile past clips or previously uploaded content, as well as promotional or advertisement oriented videos such as those containing paid promotion labels. These criteria help ensure that KoSum reflects natural viewing behavior and realistic highlight patterns in recent Korean YouTube videos.

2.2 Dataset Statistics

KoSum consists of 700 videos collected from 14 subcategories, with 50 videos per subcategory. Each video is reviewed by multiple annotators and checked against predefined filtering criteria to verify its suitability for analysis. Kosum provides three main types of variables for each video: modality features, category information, and importance scores derived from Most-Replayed signals. For the data analysis, we further use 10 groups of editing related features, including face features, curvature, visual consistency, motion dynamics, shot transition, subtitle density, text region density, reaction cue, speech dynamics, and audio novelty. KoSum modality statistics and the duration distribution of the dataset are presented in Table 2 and Figure 2, respectively.

3 Data Description

We follow the benchmark construction protocol of [22] to ensure consistency in dataset organization and evaluation settings. For each video, we construct temporally aligned multimodal sequences at one second intervals. Specifically, a video is represented as three modality sequences:

$$\mathbf{V} = \{v_1, v_2, \ldots, v_T\}, \quad \mathbf{L} = \{l_1, l_2, \ldots, l_T\}, \quad \mathbf{A} = \{a_1, a_2, \ldots, a_T\},$$

where $\mathbf{V}$, $\mathbf{L}$, and $\mathbf{A}$ denote visual, textual, and audio features, respectively, and T is the total number of temporal segments.

Visual features are extracted from video frames using a CLIP [32] visual encoder, textual features are extracted from aligned subtitles using a multilingual RoBERTa [29] encoder, and audio features are extracted using an Audio Spectrogram Transformer (AST) [9]. All modalities are aligned to the same one second timeline so that each timestamp has corresponding visual, text, and audio representations. Detailed encoder settings and the temporal text alignment procedure are provided in Appendix D.

3.1 Ground Truth

The ground truth labels are derived from Most-Replayed signals. The values are normalized so that the maximum value within each video is 1. This allows the model to regress values in the range from 0 to 1, effectively predicting a probability like importance score. The labels are defined at one second intervals to

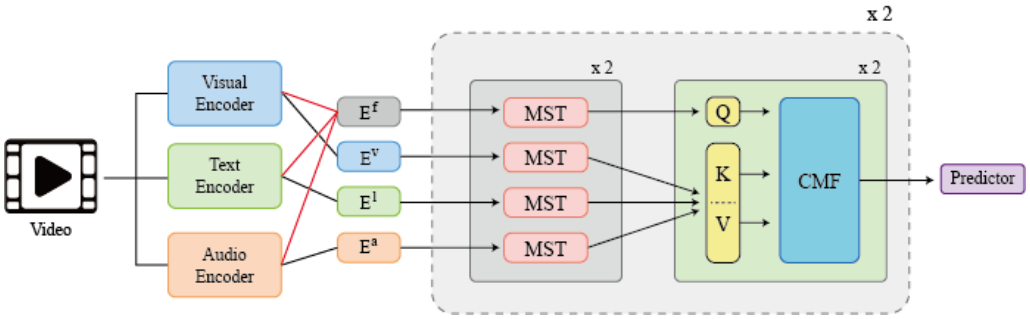


Figure 3: Simplified overview of TripleSumm [22]. The Multi-scale Temporal Block (MST) captures temporal dependencies at multiple scales by modeling both local and long-range context within each modality, while the Cross-modal Fusion Block (CMF) integrates information across visual, text, and audio modalities by dynamically attending to the most informative modality at each time step.

match the temporal resolution of the input features. Since the signal is provided as 100 discretized values, we interpolate it to align with the temporal length of the video. Furthermore, to mitigate the common artifact of anomalous spikes in the first ten seconds, we set the Most-Replayed values in this interval to zero. While Kim et al. [22] consider a shorter temporal window of 5 seconds, we extend this window to 10 seconds to account for the tendency of editors to place condensed highlight segments at the beginning of the video. A more detailed explanation is provided in Appendix B. These ground truth labels are also used in the analysis stage as core statistics in Sec. 7.

For video summarization, we further divide each video into segments using change points obtained from Kernel Temporal Segmentation (KTS) [34, 46]. The segmentation is applied to the visual feature sequence $\mathbf{V}$, where segments are determined based on visual similarity. For highlight detection, we adopt a simpler strategy and divide the video into uniform segments of 5 seconds, following a common practice [42].

3.2 Ethical and Legal Considerations

All videos used in KoSum are collected from publicly accessible YouTube content for research purposes only. We do not redistribute raw video files and store only the information necessary for analysis, such as extracted features and metadata. In addition, we exclude videos containing explicit advertisements or promotional intent to reduce potential bias in viewing behavior analysis.

4 Highlight Detection

Highlight detection aims to identify the most interesting parts of a video. The model predicts an importance score at each time step t. For evaluation, we report mAP50 and mAP15, where mAP denotes Mean Average Precision. For mAP$\mathcal{R}$, the top $\mathcal{R}\%$ segments are labeled as highlights. The score of each segment is computed by averaging the predicted scores within the segment interval.

We adopt TripleSumm [22] as the baseline model for highlight detection. Since our primary goal is feature level analysis rather than architectural improvement, we use the original model without additional modifications. We omit detailed architectural descriptions and refer readers to Kim et al. [22] for full details. A simplified overview of the architecture is shown in Figure 3.

4.1 Experiment Settings

We conduct all experiments on a single RTX4090 GPU. Following [20], we perform 5-fold cross-validation under a strict train/validation/test (TVT) split with a ratio of 8:1:1. We use the AdamW [30] optimizer with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-2}, along with a cosine learning rate scheduler with a warmup ratio of 0.1. The model is trained for 100 epochs with a batch size of 4. Our implementation is based on the TripleSumm architecture, where visual, text, and audio features are projected into a shared 768-dimensional space, and the model consists of 2 layers for each module with 4 attention heads and a dropout rate of 0.1.

Table 3: Ablation study on weight initialization strategies based on TripleSumm [22]. Higher values indicate better performance for all metrics.

Model Variants	$\tau \uparrow$	$\rho \uparrow$	mAP50 $\uparrow$	mAP15 $\uparrow$
TripleSumm + (Ramdom)	0.172	0.251	63.49	30.14
TripleSumm + (Mosu)	**0.232**	**0.332**	**66.37**	**32.22**

4.2 Metrics

Kendall's τ [18], and Spearman's ρ [50] are used to evaluate video summarization, while mAP50 and mAP15 are used for highlight detection. The only difference between these tasks lies in their target objectives, whereas the model predicts the same score, namely the Most-Replayed signal. As highlight scenes tend to be included in summary videos more frequently than others [42], we additionally report video summarization metrics. Specifically, we report rank-based metrics, τ and ρ, to assess the consistency between predicted rankings and ground truth, which better reflects the objective of video summarization, following recent practice [20, 22, 13, 33, 39, 43].

4.3 Performance

As shown in Table 3, initializing the model with weights pretrained on MoSu [22] consistently yields better performance than random initialization across all metrics. This improvement is attributed to the large-scale multimodal pre-training, which provides a strong prior for capturing general highlight patterns. The model achieves higher scores in both highlight detection (mAP50, mAP15) and video summarization (τ, ρ), indicating its effectiveness in ranking and localizing important segments.

5 Methodology

In this section, we present the selected features and explain the rationale behind our choices. Following the triple modality framework, we categorize the features into three sections: visual, text, and audio. In particular, visual features are further divided into two components: semantic change and visual dynamics, which capture rich visual information in videos. Although there exists an inherent imbalance across modalities, we expect that these modality specific features provide insight into how each modality independently influences repeated viewing behavior. By default, features are extracted at 1 fps, while features with higher temporal dynamics are extracted at a higher frame rate to better capture rapid changes in video structure. Together with these modality specific features, we use modality attention weights from the last layer for data analysis, as this representation more clearly captures discriminative modality patterns between highlight and non highlight regions. The attention comparison across different layers is provided in Appendix E.

5.1 Semantic Change (Visual)

Curvature We adopt curvature as a basic feature to capture how sharply the embedding trajectory turns over time. Following prior work [38, 14], curvature captures how sharply the token embeddings turn when viewed as a sequence in $\mathbb{R}^d$. We compute curvature at time t using adjacent visual embeddings. Specifically, we compute curvature as

$$c_t = \arccos\left(\frac{(E_t^v)^\top E_{t-1}^v}{\|E_t^v\|\,\|E_{t-1}^v\|}\right),$$

where E_t^v denotes the visual embedding extracted by the visual encoder.

Face Features We use face features to capture how facial composition relates to highlight segments. We detect faces using a YOLOv8 model [44] trained for face detection[1]. For each frame, we compute three features: face count, face area, and face center dispersion. These correspond to the number of detected faces, the size of each face region, and how widely the centers of detected faces are distributed within the frame. Based on these signals, we analyze the composition of faces in terms of the number of participants, their spatial arrangement, and the level of crowding. We expect highlight segments to contain a larger number of faces, as multiple participants often appear during important events. In addition, close-up shots

[1] arnabdhar/YOLOv8-Face-Detection

are frequently used in dramatic moments, leading to larger face regions. At the same time, faces tend to be more widely distributed across the frame, as the scene is filled with people, resulting in higher spatial dispersion compared to non-highlight segments.

5.2 Visual Dynamics (Visual)

Visual Consistency We hypothesize that highlight segments often maintain visual consistency, with limited semantic changes such as shifts in location, topic, or background. To capture such changes, we apply Kernel Temporal Segmentation (KTS) to identify semantic boundary timestamps. By comparing these boundaries with highlight segments, we expect that semantic changes are less frequent in most-replayed regions, indicating that visual continuity may be a key characteristic of highlight segments.

Motion Dynamics We use optical flow to capture motion dynamics. Optical flow estimates the magnitude of motion by measuring how each point moves between consecutive frames under assumptions such as brightness constancy, temporal persistence, and spatial coherence, which generally hold at sufficiently high sampling rates. To ensure computational efficiency, we adopt a memory efficient deep learning based optical flow model instead of classical method which needs recurrent computation. We use MEMFOF [5][2] for optical flow estimation. After estimation, we also compute average motion dynamics by summing the absolute optical flow components over spatial dimensions. Specifically, we define

$$\texttt{MD}_t = \sum_{i \in H} \sum_{j \in W} \left(\left| u_{i,j}^t \right| + \left| v_{i,j}^t \right| \right),$$

where $u_{i,j}^t$ and $v_{i,j}^t$ denote the horizontal and vertical flow components at spatial location (i, j) at time t, and H and W represent the height and width of the image, respectively.

Shot Transition We hypothesize that highlight segments contain more frequent shot transitions, as rapid visual changes can increase rhythm and viewer attention. To examine this, we detect shot transition timestamps using visual change cues. We sample each video at fixed FPS (e.g., 4). For adjacent sampled frames, we compute color histogram differences and edge change ratios. Each score is independently normalized, combined through a weighted sum. We then identify local maxima above a pre-defined percentile.

5.3 Text-based Features (Text)

Subtitle Density We use subtitle density to quantify how densely textual information appears over time. It reflects how much information or reaction is compressed within a short time interval, indicating the intensity of textual augmentation. Based on the aligned captions in *Temporal Text Alignment* (Appendix D.2), we compute the number of whitespace-separated tokens within a fixed temporal window and normalize it by the window length to obtain token density (tokens/s). We hypothesize that highlight segments exhibit higher subtitle density, as they require rapid delivery of key information or reactions. In contrast, non-highlight segments tend to have sparse or uniformly spaced subtitles, whereas highlight segments show frequent and concentrated subtitle occurrences within a short time window.

Text Region Density While subtitle density captures text provided as metadata or automatically generated captions, we focus on text information that is directly embedded in the video frames. We use text region density to quantify the presence and prominence of visual text, such as subtitles, emphasized phrases, and graphic text effects, as a key textual feature. This allows us to capture how textual content is visually delivered within the scene beyond external subtitle signals.
To extract this feature, we apply EasyOCR to perform frame level text detection. EasyOCR leverages CRAFT [4] for text region detection and a CRNN [37] based recognizer, allowing us to obtain bounding box coordinates of detected text. Based on these regions, we compute the number of text boxes, total text area, and average box size to measure text region density. We hypothesize that highlight segments exhibit higher text region density, as they often include more frequent and visually prominent textual cues, while non-highlight segments tend to contain fewer or less emphasized text elements.

[2] egorchistov/optical-flow-MEMFOF-Tartan-T-TSKH

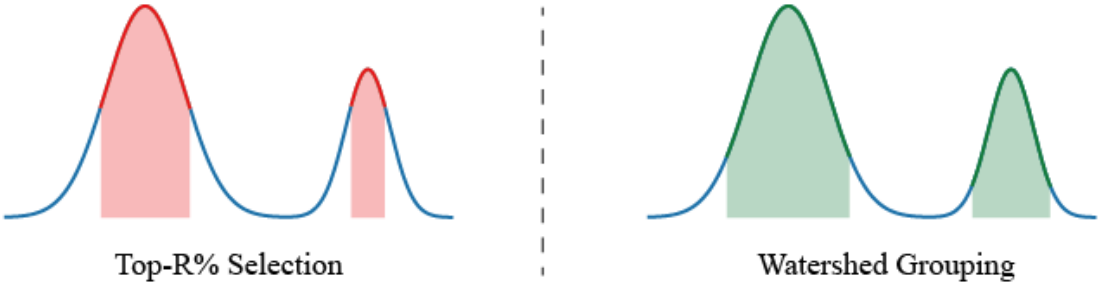


Figure 4: Comparison of highlight segmentation strategies. Left: top-$R\%$ selection. Right: watershed grouping.

5.4 Reaction Cue (Audio)

Speech Dynamics We use speech dynamics to capture emotional arousal reflected in vocal behavior. It reflects how much a speaker's delivery intensifies within a given moment, indicating heightened engagement or excitement. When a speaker becomes excited or strongly engaged, speech rate tends to increase and pitch often rises or fluctuates more sharply relative to its surrounding context. We measure speaking rate as the number of syllable nuclei detected per second within voiced regions, and pitch deviation as the degree to which the local pitch departs from its smoothed baseline. The final speech dynamics score is a weighted combination of these two components, computed per second using LibROSA [31]. We hypothesize that highlight segments exhibit higher speech dynamics, as they tend to coincide with moments of emotional peak or rapid information delivery. In contrast, non-highlight segments tend to show more stable speaking patterns with moderate speed and gradual pitch variation, reflecting calmer and less eventful portions of the content.

Audio Novelty We use audio novelty to quantify how abruptly the acoustic properties of the audio change over time. It reflects the degree of acoustic discontinuity at each moment, indicating transitions that are likely to draw viewer attention. Abrupt shifts in sound characteristics, such as sudden changes from calm dialogue to intense background music or the emergence of unexpected sound effects, serve as strong auditory cues for engagement. We measure audio novelty through two complementary signals: spectral flux, which captures how rapidly the frequency distribution evolves, and MFCC change, which captures shifts in the overall timbral structure of the audio. The final audio novelty score is a weighted combination of these two components, computed per second using LibROSA. We hypothesize that highlight segments exhibit higher audio novelty, as they tend to coincide with acoustically dynamic moments. In contrast, non-highlight segments tend to maintain consistent and predictable audio characteristics, reflecting portions of the content where the sound environment remains stable.

Laughter and Applause We use laughter and applause to capture the presence and intensity of collective audience reactions as direct indicators of shared viewer engagement. Unlike other audio features that track continuous acoustic variation, this signal reflects discrete social responses that arise specifically at moments the audience finds compelling or emotionally resonant. We estimate per-second reaction probabilities using an Audio Spectrogram Transformer (AST) classifier [9][3] pretrained on AudioSet [8], and separately track laughter and applause as distinct reaction types to preserve the specificity of each social cue. We hypothesize that highlight segments exhibit elevated reaction probabilities, as they are more likely to elicit strong and synchronized responses from the audience. In contrast, non-highlight segments are less likely to trigger such collective reactions, resulting in consistently low laughter and applause scores throughout those intervals.

5.5 Highlight Segmentation

For analysis, highlight regions should be defined as temporally coherent peaks rather than isolated high score frames. As shown in Figure 4, a naive top-$R\%$ strategy selects frames according to a fixed budget, which can cut through a highlight peak near the budget boundary or select only a subset of a broader peak. As a result, it may fragment continuous highlight regions and distort the natural peak structure of the Most-Replayed signal. While budget constrained selection is useful for generating summaries, our goal is to

[3] MIT/ast-finetuned-audioset-10-10-0.4593

identify meaningful highlight peaks for data analysis. Since the integrated mass and temporal width of highlights vary across videos, enforcing a fixed budget is not suitable for this purpose.
To address these issues, we use watershed grouping [12] for highlight segmentation. The Most Replayed signal is treated as a one dimensional landscape, and regions around local maxima are grouped into contiguous highlight segments. This allows the method to capture entire highlight peaks with adaptive widths, rather than selecting only the high score subset imposed by a fixed budget. However, directly fixing the number of modes can still be unstable. If k is too small, nearby highlight modes may be missed. If k is too large, low score modes may be selected as highlights. Therefore, we adopt **adaptive watershed grouping**, which uses thresholding with a large k to capture reasonable candidate modes while suppressing low score regions. Details of the hyperparameter selection scheme are provided in Appendix H.

6 Analysis Preparation

6.1 Overall Data Analysis Pipeline

In this section, we analyze the extracted modality specific features described in Sec. 5. Since raw feature values are difficult to interpret directly, we first convert them into descriptive statistics and inferential statistics by comparing highlight and nonhighlight regions. We then use inferential statistics to identify significant variables for each category, focusing on values such as mean delta, positive ratio, Wilcoxon test results, rank-biserial correlation, and q-value, where q-value denotes the FDR-corrected p-value.
For the selected significant variables, we combine their descriptive statistics and inferential statistics into structured analysis records. These records are passed through API calling to support more objective and consistent data interpretation. The resulting analysis files are saved in JSON format and organized into structured category level analysis cards. These cards summarize modality attention patterns, statistically significant feature patterns, and their editing implications, and are later used as data grounded context for the LLM based editing guideline generation. generation.

6.2 Descriptive Statistics

Time Index Sets For each video, we define four time index sets based on the adaptive watershed grouping based highlight segmentation. Let $\mathcal{H}$ denote the set of highlight timestamps and $\mathcal{N} = \{0, \dots, T-1\} \setminus \mathcal{H}$ denote the set of nonhighlight timestamps. We also define $\mathcal{B} \subset \mathcal{H}$ as the boundary region of highlight segments, where $\mathcal{B}$ contains timestamps within 2 seconds from either the start or end boundary of each highlight segment. In other words, for each highlight segment, the first 2 seconds and the last 2 seconds are included in $\mathcal{B}$. Finally, we define $\mathcal{C}$ as the local context around highlights, where $\mathcal{C}$ contains nonhighlight timestamps within 5 seconds of each highlight segment. These sets allow us to compare highlight regions with both the entire nonhighlight region and the local surrounding context.

Continuous Feature Metrics For a continuous feature sequence $s = (s_1, \dots, s_T)$, we summarize its behavior using level, change, boundary effect, local contrast, ratio, and effect value. The level measures the average feature strength within a region, defined as $\bar{s}_A = \frac{1}{|A|} \sum_{t \in A} s_t$ for a time set A. We compute $\mathrm{level}_H = \bar{s}_H$ and $\mathrm{level}_N = \bar{s}_N$ to compare the average feature magnitude between highlight and nonhighlight regions. The change measures short term variation, defined as $\delta_t = |s_t - s_{t-1}|$, and we compare $\mathrm{change}_H = \bar{\delta}_H$ and $\mathrm{change}_N = \bar{\delta}_N$. The boundary effect measures whether a feature becomes stronger near highlight boundaries, defined as $\bar{s}_{H \setminus B} - \bar{s}_B$. The local contrast compares highlights with their nearby context, defined as $\bar{s}_H - \bar{s}_C$. The ratio measures relative strength, defined as $\mathrm{level}_H / \mathrm{level}_N$, and the effect value measures absolute difference, defined as $\mathrm{level}_H - \mathrm{level}_N$. These metrics are used to examine whether each feature becomes stronger, more dynamic, or more locally distinctive in highlight regions.

Level-Only Continuous Metrics For features that already represent temporal change, such as curvature, we use level-based statistics without applying an additional change metric. Since curvature itself captures how sharply the embedding trajectory changes over time, computing another temporal difference may amplify noise. Therefore, for these features, we mainly report level, boundary effect, local contrast, ratio, and effect value.

Point Event Metrics For point-based features such as shot transitions, we represent detected event timestamps as a set P. The event rate measures how frequently events occur within a region, defined as

$\text{rate}_H = |P \cap H|/|H|$ and $\text{rate}_N = |P \cap N|/|N|$. We also compute burstiness to measure whether events are concentrated in a short period. Given event gaps g, burstiness is defined as σ_g/μ_g, where σ_g and μ_g are the standard deviation and mean of the event gaps. A higher burstiness value indicates that point events are more densely concentrated in specific temporal regions rather than being evenly distributed. These metrics allow us to test whether highlight regions contain denser and more clustered editing events.

Segment Metrics For segment based features such as visual consistency, we use scene segments $\mathcal{S} = \{S_1, \ldots, S_K\}$ and compare them with highlight spans. Purity measures how well each highlight segment is contained within a single visual segment, defined as the average of $\max_k |\text{span} \cap S_k|/|\text{span}|$ over all highlight spans. Fragmentation measures how many visual segments overlap with a highlight span, normalized by the span length or number of involved segments. Boundary distance measures how close highlight timestamps are to scene boundaries, defined as $\frac{1}{|H|}\sum_{t\in H} \min_{p\in P_{\text{bnd}}} |t-p|$, where P_{bnd} is the set of scene boundary timestamps. These metrics help determine whether highlights are visually coherent or frequently divided by scene changes.

6.3 Inferential Statistics

Video-Level Paired Difference We conduct inferential analysis at the video-level to avoid treating second-level timestamps as independent samples. For each video i and each metric, we compute a paired difference $d_i = \text{stat}_H^{(i)} - \text{stat}_N^{(i)}$, where $\text{stat}_H^{(i)}$ and $\text{stat}_N^{(i)}$ are the highlight and nonhighlight statistics from the same video. This within-video comparison reduces the effect of video-specific scale differences and allows each video to contribute one observation to the category-level test.

Core Statistical Values For each category and feature metric, we report mean delta, positive ratio, the Wilcoxon signed-rank test, rank-biserial correlation, and q-value. Mean delta is defined as $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and indicates the average direction and magnitude of the difference. Positive ratio is defined as $|\{i : d_i > 0\}|/n$ and measures how consistently the feature is stronger in highlight regions across videos. The Wilcoxon signed-rank test evaluates whether the median of paired differences is significantly different from zero without assuming normality. Rank-biserial correlation, defined as $r_b = (R^+ - R^-)/(R^+ + R^-)$, provides the effect size, where R^+ and R^- are the sums of positive and negative signed ranks after excluding zero differences. Finally, the q-value is obtained by applying the Benjamini-Hochberg procedure to the Wilcoxon p-values to control false discoveries across multiple metric comparisons within each (category, feature) group.

Interpretation Criteria We interpret each result using four complementary aspects: direction, consistency, statistical reliability, and effect size. The direction is determined by the sign of mean delta, the consistency is measured by the directional consensus ratio (positive ratio when the effect is in the highlight direction and negative ratio otherwise), the reliability is assessed by the q-value, and the effect size is measured by rank-biserial correlation. A feature is considered a strong category-specific signal when it satisfies all three of the following criteria: a consensus ratio ≥ 0.7, a q-value < 0.05, and an effect size $|r_b| \geq 0.3$ in the expected direction. Features satisfying two of the three criteria are regarded as showing a weaker tendency, while those satisfying at most one are treated as inconclusive. These criteria are used to identify which modality-specific features can support reliable editing guideline generation.

7 Analysis

Before presenting the detailed analysis results, we first summarize the meaningful modality specific features identified for each subcategory in Table 4.

Cooking In the *Cooking* category, highlight regions show a strong tendency toward visually coherent and socially reactive moments. The inferential statistics indicate that visual consistency is the most reliable visual signal, suggesting that cooking highlights are often formed within a single coherent scene rather than through rapid motion or frequent cuts. Semantic change analysis further identifies face related features as a primary signal, with highlights tending to contain more people who are also more widely distributed across the frame. In the audio modality, reaction cues are strongly associated with highlights, driven by both laughter and applause, with laughter the dominant component. Audio novelty also appears as a supporting cue, indicating that locally salient acoustic changes can help distinguish highlight moments. In contrast,

Table 4: Key features identified for each subcategory. ✓ marks a feature as key when at least one of its main-signal or component metrics meets the selection criterion ($|r_b| \geq 0.3$, consensus ≥ 0.7, and $q < 0.05$). Curv.: Curvature, FF: Face Features, VC: Visual Consistency, MD: Motion Dynamics, ST: Shot Transition, SD: Subtitle Density, TRD: Text Region Density, SpD: Speech Dynamics, AN: Audio Novelty, LA: Laughter and Applause. Detailed statistical evidence is provided in Appendix I.

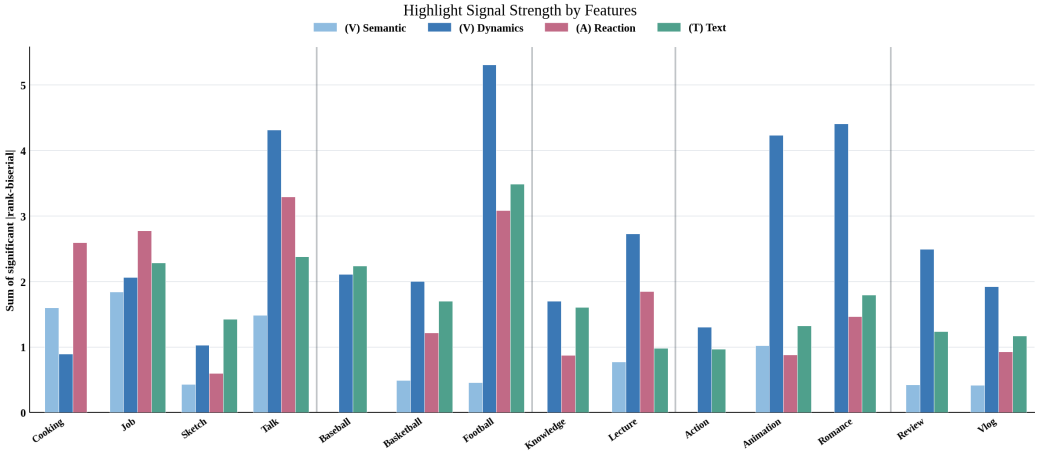
	Visual					**Text**		**Audio**		
Subcategory	Curv.	FF	VC	MD	ST	SD	TRD	SpD	AN	LA
Cooking		✓	✓						✓	✓
Job Experience	✓	✓	✓	✓		✓	✓	✓		✓
Sketch Comedy	✓	✓				✓	✓	✓		
Talk	✓	✓	✓	✓	✓	✓		✓	✓	✓
Action	✓		✓				✓			
Animation	✓		✓	✓	✓	✓	✓	✓	✓	
Romance	✓		✓	✓		✓	✓	✓	✓	
Baseball			✓			✓		✓	✓	✓
Basketball	✓		✓	✓		✓			✓	✓
Football	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Review	✓		✓	✓			✓			
Vlog	✓	✓	✓	✓			✓		✓	
Knowledge			✓		✓	✓			✓	
Lecture			✓					✓	✓	

speech dynamics and text based features do not show reliable relationships with highlights, making cooking the one category in which the textual channel is entirely uninformative. Overall, cooking highlights are characterized by coherent scenes, a distributed multi-person composition, laughter and applause centered reactions, and local acoustic changes, rather than fast cutting, heavy motion, or any textual cues.

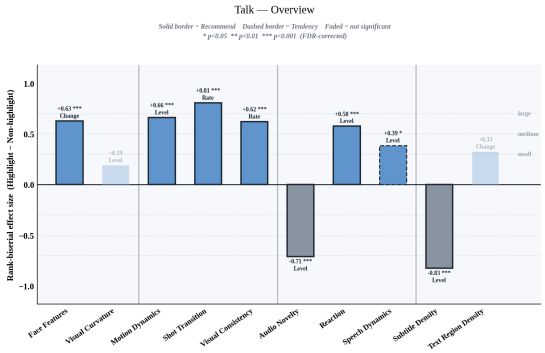
Job Experience In the *Job Experience* category, highlight regions are associated with meaningful visual changes while still maintaining scene coherence. Visual dynamics analysis shows that highlight regions stay within coherent scenes, with high purity and low fragmentation as strong effects, while motion becomes more salient relative to the local context, especially along the horizontal axis. Semantic change analysis indicates that the embedding trajectory shifts substantially, driven by adjacent-frame embedding distance, suggesting that highlights often occur when new objects, places, actions, or work related situations are introduced. In the audio modality, laughter based reactions provide the strongest signal, applause is largely absent as a cue, and pitch related speech changes serve as an additional cue. Text based analysis shows a distinctive pattern: subtitle quantity tends to decrease, subtitle changes become more frequent, and on screen text boxes appear less often. Overall, job experience highlights emphasize semantic visual changes and coherent task scenes, while using concise and fast changing textual support.

Sketch Comedy In the *Sketch Comedy* category, highlight regions are mainly characterized by reduced speech and a lighter textual load. On the visual side, the reliable key signals are the adjacent-frame semantic change (the embedding-change component of curvature) and the spatial spread of faces, indicating that sketch highlights coincide with localized shifts in visual content and in face arrangement; scene coherence contributes as a more moderate cue, whereas strong motion and frequent shot transitions are not reliable signals. In the audio modality, reaction cues such as laughter and applause are not selected as key features, while speech dynamics decreases around highlights. Text based features show that subtitle load is the dominant signal: subtitle density is reliably reduced, and alongside this reduction the subtitle turnover increases, indicating that sketch highlights pair fewer captions with faster caption changes; a reliable contraction of on screen text area accompanies this, while the overall number of text boxes changes only marginally. Overall, sketch comedy highlights are better explained by localized semantic and facial change, decreased speech activity, and a reduced yet more rapidly changing textual load, supported by moderate scene coherence, rather than by audience reactions, heavy motion, or frequent cuts.

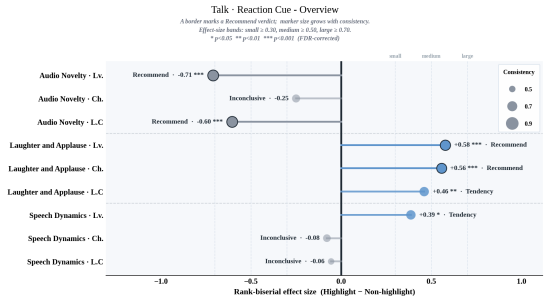
Talk In the *Talk* category, highlight regions exhibit the strongest multimodal signature. Visual dynamics shows that highlights are associated with increased motion, more frequent shot transitions, and denser



(a) Overall highlight signal strength across subcategories and feature groups.



(b) Talk category feature-level highlight pattern. Rank-biserial effect sizes summarize highlight versus non-highlight differences.



(c) Reaction Cue feature breakdown for Talk, including audio novelty, laughter and applause, and speech dynamics.

Figure 5: Summary of highlight-related feature signals. The top panel compares feature-group strength across subcategories, while the bottom panels show detailed patterns for the Talk subcategory

scene-boundary activity, with the shot-transition rate standing out as the dominant visual signal; this indicates that talk highlights are constructed through dynamic visual layering rather than scene coherence. Semantic change analysis shows that the face composition changes meaningfully, while the cosine-distance change component provides strong evidence of large adjacent-frame embedding changes, the dominant semantic signal, indicating that the visual content shifts rapidly from frame to frame. This suggests that participant turnover, camera switches between speakers, and rapid shifts in visual focus are important. In the audio modality, reaction cues increase, with both laughter and applause active, and the speaker's pitch rises markedly, while audio novelty decreases, indicating that the background acoustic environment becomes more stable and quiet. Text based analysis shows a strong bidirectional pattern within the subtitles: subtitle quantity drops sharply, but subtitle change increases, whereas on-screen text regions remain irrelevant. Overall, talk highlights are characterized by dynamic visual pacing driven by rapid cutting, changing face composition with large frame-to-frame embedding shifts, concentrated speech and reaction cues, a quieter acoustic background, and short rapidly changing subtitles.

Action In the *Action* category, highlight regions are primarily characterized by visual scene cohesion, with semantic and on-screen text variation acting as secondary signals. The dominant cue is visual consistency: scene purity is by far the strongest signal, indicating that highlights are cleanly contained within a single coherent scene, while an elevated scene-boundary rate shows that this cohesive scene is also newly begun at the highlight onset. Semantically, the highlights involve pronounced changes in adjacent-frame visual content, as the embedding distance between neighboring frames widens to introduce new objects or events, although the trajectory itself does not turn. Text based features contribute an asymmetric pattern: subtitle density decreases while the moment-to-moment variation of on-screen text regions increases, even as the average number of text boxes stays unchanged. In contrast, face features, motion dynamics, shot transitions,

and audio reaction cues do not form reliable key signals. Overall, action highlights are better explained by cohesive visual scene structure, adjacent-frame semantic change, and asymmetric text variation than by raw motion intensity, rapid cutting, or reaction sounds.

Animation In the *Animation* category, highlight regions show strong activation across the visual, semantic, text, and audio channels at once. Visual consistency, motion dynamics, and shot transitions indicate that animation highlights pair a coherent scene structure with active visual movement and frequent cuts, the rapid cuts remaining bound within a single cohesive scene rather than fragmenting it. Semantic curvature is a genuine signal, with highlights coinciding with pronounced semantic transitions in the visual content, whereas the number of on screen faces is not determining and tends if anything to thin out. Text based features add further cues, as subtitle density eases off while on screen text regions become more variable. In the audio modality the pattern is one of contrast: audio novelty quiets down while speech grows more prominent against its context through a rising vocal pitch, and reaction cues such as laughter or applause are absent as a highlight cue. Overall, animation highlights are characterized by coherent yet actively moving and frequently cut visuals, pronounced semantic transition, eased subtitles with more variable on screen text, and a quieter acoustic background against rising speech rather than reaction sounds.

Romance In the *Romance* category, highlight regions are defined by visual channels turning on while textual and audio channels are emptied. The strongest and most consistent visual cue is the scene-boundary rate, with highlights concentrated at the onset of new scenes, accompanied by reduced fragmentation, so that transitions coincide with internally coherent scene structures. Rising motion dynamics and shot transitions act as supporting cues, indicating that highlights carry controlled, restrained visual dynamism layered onto otherwise static content. Face features contribute only weakly, through larger face size in close-up framing rather than person count, so they are best read as a secondary component effect. Text based features, including subtitle density and on-screen text region density, are key cues in the opposite sense: both are markedly suppressed, as language information is cleared from the screen. In the audio modality, audio novelty and speech dynamics likewise decrease, indicating that highlights settle into quieter, slower-spoken moments, with only a faint laughter response accompanying them while applause is absent. Overall, romance highlights are shaped by scene-boundary transitions with coherent, restrained visual dynamics and close-up framing, set against the suppression of on-screen text and the calming of audio and speech.

Baseball In the *Baseball* category, highlight regions are explained by a relatively small set of features. Visual consistency indicates that important moments remain within a single coherent scene with few scene splits, while motion dynamics provides a secondary cue, with more side-to-side movement within an otherwise fixed frame. Subtitle density also appears as a key text based signal, where caption quantity decreases during highlights, indicating a reduced textual load. In the audio modality, the commentary pitch rises markedly as the strongest acoustic signal, reflecting more excited play by play, while audio novelty is reliable in the opposite direction, as its acoustic timbre becomes more stable, and laughter adds a reliable local cue. In contrast, face features, semantic curvature, shot transitions, and text region density do not provide reliable key signals. Overall, baseball highlights are mainly characterized by coherent game scenes with fewer subtitles and more excited, higher pitched commentary, rather than by frequent cuts or semantic shifts.

Basketball In the *Basketball* category, highlight regions are defined by an asymmetric structure in which the visual channel alone is actively engaged while every other channel settles into a calmer, suppressed state. The decisive visual signal is overall motion level rather than its short-term variability: highlights capture stretches of high court-wide movement held within a single, coherent scene that resists fragmentation. In the semantic channel, by contrast, the embedding trajectory becomes locally smoother, bending less than its surroundings, so highlights coincide with steadier, less abrupt semantic transitions rather than sharp turning points. Subtitle density acts as a text cue only through a localized reduction in caption volume, while on-screen text regions and speech dynamics carry no reliable signal. The acoustic channel is bifurcated: the overall sound level is lower against a quieter background, with only momentary short-term change rising. Reaction cues contribute a localized laughter signal, with a reliable rise in laughter local contrast at highlights even though the combined reaction level itself is not significant. Overall, basketball highlights are characterized by visually active, high-motion gameplay within a coherent scene, accompanied by a smoother semantic flow and largely subdued audio and text channels apart from this localized laughter response.

Football In the *Football* category, highlight regions show one of the broadest multimodal signatures, organized around fast, fragmented transitions rather than a single sustained scene. Face features, visual consistency, motion dynamics, and shot transitions are all meaningful, with the semantic signal being the temporal change in face and person composition over time rather than any turning of the embedding trajectory. Football is in fact the uniquely fragmented, multi-scene type, combining high fragmentation with frequent shot transitions. Text based features are also important but bifurcated: subtitle quantity drops the most of any category while subtitle change rises, and on-screen text region density decreases. The audio pattern is likewise bifurcated, as audio novelty drops sharply while speech pitch rises, and reaction cues are not selected as key features. Overall, football highlights are formed by lowering the audiovisual information load while raising the rate of change, rather than through explicit reaction sounds.

Review In the *Review* category, highlight regions are mainly associated with information transition and visual organization. Scene related features suggest that highlights occur where new scenes begin to cluster, with the highlighted span kept as one clean, unfragmented scene rather than broken into pieces. Motion dynamics, in contrast, tends to be calmer than usual, indicating that highlight moments are quieter, controlled transitions rather than visually active ones. Semantic features show a smooth hand off rather than an abrupt cut: the trajectory curvature flattens while the frame to frame semantic change accumulates, so the content updates gradually. Text region density is the strongest signal, reflecting the role of on screen information such as product names, ratings, or key points, which appear and turn over more frequently at highlights. In contrast, subtitle density, person and face cues, and audio features do not form reliable highlight signals. Overall, review highlights are characterized by visual information updates, controlled transitions over a coherent scene, and on screen text cues.

Vlog In the *Vlog* category, highlight regions are characterized by scene transitions, person related changes, and social reaction cues. Scene-boundary density emerges as the dominant visual signal, indicating that vlog highlights align with the onset of new scenes rather than with raw shot frequency, where the shot-transition rate is only a weak tendency. Face features reveal that highlights coincide with changes in the number of people on screen, as participants enter or exit, rather than with static facial composition; the underlying semantic signal is a rapid adjacent-frame shift in content, indicating that the visual content changes quickly at highlights. Text region density is meaningful through increased turnover, showing that on screen text boxes are swapped more frequently rather than becoming more numerous. In the audio modality, laughter reactions drive highlight formation while applause is uninformative, and vocal pitch contributes only as a secondary cue, with conversational speaking rate remaining flat. Overall, vlog highlights are shaped by scene transitions, shifts in the number of people on screen, faster turnover of on screen text, and laughter-centered audio cues.

Knowledge In the *Knowledge* category, highlight regions are characterized by stable visual presentation and selective audio emphasis. Visual consistency and shot transitions indicate that important moments occur within a single coherent explanatory scene, unbroken by external cutaways but punctuated by frequent cuts spaced evenly rather than clustered at scene boundaries. Subtitle behavior is a key text based signal: captions become steadier and less fluctuating while, if anything, slightly denser, suggesting that a stable and sustained narration supports the delivery of important information, while on-screen text acts only as a weaker auxiliary cue. Audio novelty is also meaningful, indicating a general rise in acoustic change accompanying explanation. In contrast, face features, semantic curvature, motion dynamics, reaction cues, and speech dynamics do not serve as reliable key features. Overall, knowledge highlights are defined by coherent visual structure, subtitle stabilization, and audio novelty.

Lecture In the *Lecture* category, highlight regions are mainly explained by visual structure, semantic smoothness, and speech related changes. Visual consistency emerges as the dominant signal, indicating that lecture highlights sit within a single cohesive scene that is at the same time demarcated as a new, self contained instructional unit. Semantic curvature indicates that the trajectory becomes smoother rather than sharper, so visual content transitions gradually instead of through abrupt turns. Speech dynamics appears as the key audio signal, but in the direction of a slowing or quieting of the speaker's delivery rather than an intensification. Subtitle density and reaction cues act only as weak auxiliary cues, and on screen text regions remain irrelevant. Overall, lecture highlights are characterized by cohesive yet clearly bounded visual organization, smooth semantic transition, and subsiding speech rather than dense text or reaction sounds.

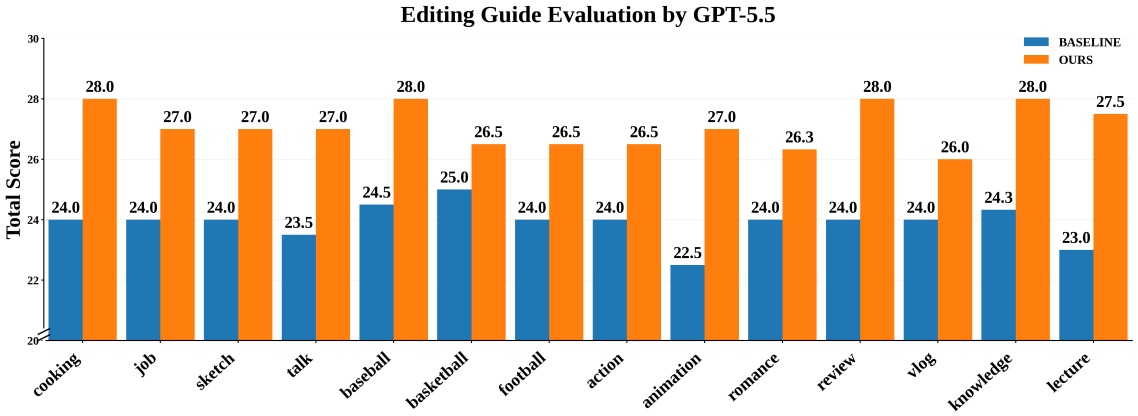


Figure 6: Average GPT-assisted evaluation scores of editing guidelines generated by our prompting framework across subcategories.

8 Results and Final Deliverables

8.1 Data Analysis Structure

To use the analyzed patterns effectively during guideline generation, we organize the KoSum analysis results into a structured category level analysis card before providing them to the LLM. Instead of inserting raw statistical results or unstructured descriptions, this format summarizes the key modality and feature patterns for each content category, allowing the LLM to generate guidelines grounded in observed data.

Each analysis card consists of four components. The *category profile* summarizes the target content category and its general editing characteristics. The *modality contribution* describes the relative importance of visual, audio, and text signals based on attention weights from the trained triple modality model. The *significant feature patterns* summarize statistically significant differences between highlight and non highlight regions using modality specific features such as face features, motion dynamics, shot transitions, subtitle density, speech dynamics, and audio novelty. Finally, the *editing implication* translates these patterns into practical editing cues for creators.

The structured analysis is combined with the user prompt, which includes the target category, video description, and desired editing direction. This helps the LLM focus on data grounded editing factors rather than generic suggestions, producing category aware and practically applicable editing guidelines.

8.2 Evaluation Protocol

We evaluate the generated editing guidelines to verify whether the proposed prompting framework provides more useful guidance than a general user prompt alone. The baseline prompt contains only the target category, a brief video description, and the desired editing direction as a user input. In contrast, our prompt additionally includes category specific analysis obtained from KoSum, such as modality attention patterns and statistical feature patterns observed in highlight regions.

For evaluation, we compare the editing guidelines generated by the baseline prompt and our prompting framework using a GPT-assisted evaluator. Each guideline is evaluated according to seven predefined criteria: editing-direction understanding, scene-selection clarity, editing-structure specificity, editing-element detail, video-content relevance, actionability, and sample specificity.

Editing-direction understanding measures whether the generated guideline correctly reflects the editing goal requested by the user. Scene-selection clarity evaluates whether the guideline provides clear criteria for selecting highlight scenes. Editing-structure specificity measures whether it describes a concrete temporal flow, such as introduction, development, climax, and conclusion. Editing-element detail evaluates whether practical elements such as subtitles, cuts, background music, sound effects, and transitions are described in sufficient detail. Video-content relevance measures whether the guideline reflects the specific scenes and events provided in the video description. Actionability evaluates whether an editor can directly apply the suggested actions on the editing timeline. Finally, sample specificity measures whether the guideline avoids generic advice and provides recommendations tailored to the given video sample.

Each criterion is scored on a five point scale. A score of 1 indicates that the guideline is largely irrelevant or merely formal, whereas a score of 3 indicates that the general direction is appropriate but remains partly

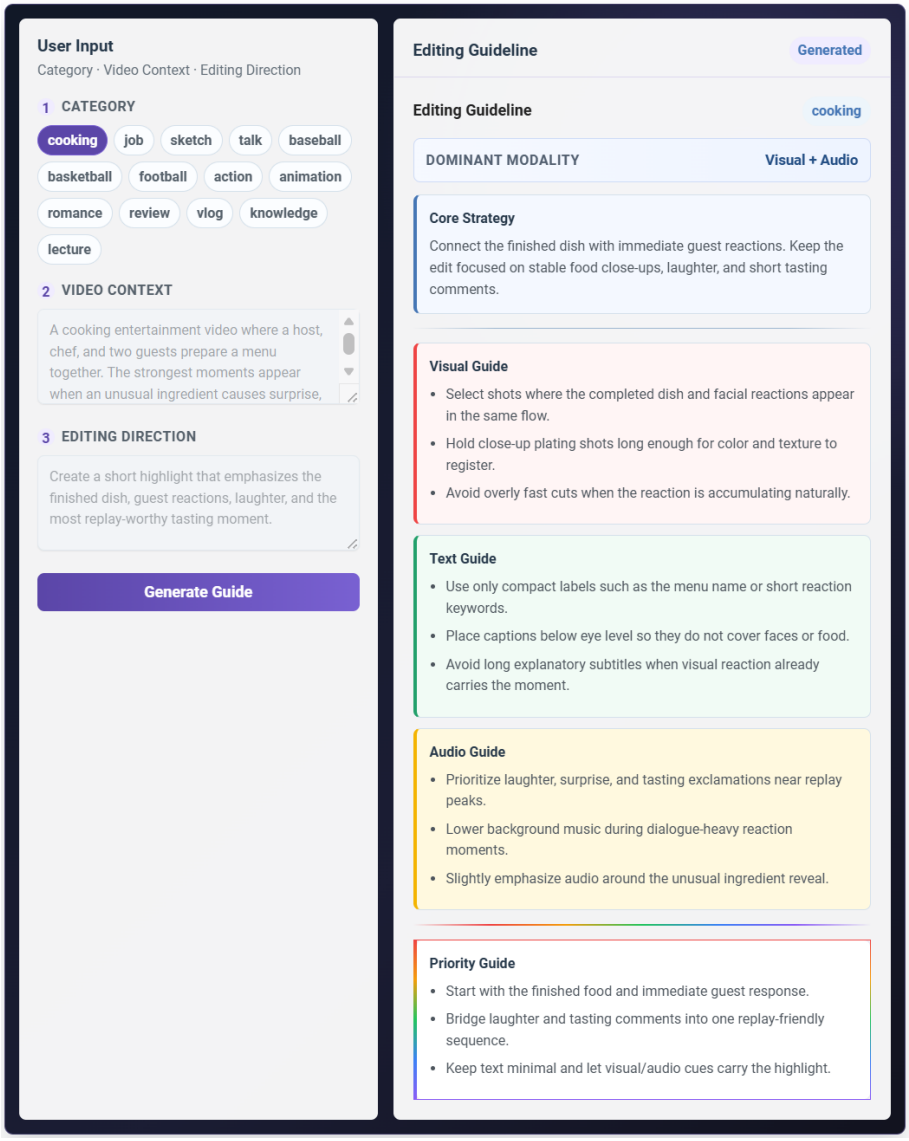


Figure 7: Demo interface of the proposed video editing guidance system.

generic or insufficiently actionable. A score of 5 indicates that the guideline is highly specific to the given sample, directly actionable by an editor, and contains few missing elements. The maximum total score for each guideline is therefore 35 points. To construct evaluation inputs, we randomly generate 30 pairs of input categories and video descriptions using an LLM. Each video description includes the overall content of the video and the desired editing direction. We then use the same inputs for both the baseline prompt and our prompting framework. The total guideline quality score is computed as the sum of the seven criterion scores, resulting in a maximum score of 35 points per guideline. Finally, we compare the overall scores between the baseline and our prompting framework to measure whether the proposed framework improves the quality of generated editing guidelines.

As shown in Figure 6, we adopt a GPT-assisted evaluation protocol based on seven predefined criteria. The evaluation is conducted on a total of 30 test samples, covering 14 subcategories with 2-3 samples per subcategory. Under this evaluation setting, the proposed **data analysis conditioned prompting framework** achieves equal or higher average scores than the user-prompt-only baseline across all seven criteria. In particular, the largest improvements are observed in sample specificity, editing-structure specificity, editing-element detail, and actionability. This indicates that the proposed framework generates editing guides that are less generic, better organized in temporal flow, and more detailed in practical editing elements such as subtitles, cuts, BGM, sound effects, and transitions.

9 Limitations

KoSum currently has limited scalability because the filtering stage relies heavily on manual inspection. After collecting candidate videos, annotators manually reviewed whether each sample satisfied the filtering criteria, which made it difficult to expand the dataset beyond the current scale. As a result, each subcategory contains about 50 videos from a limited range of creators, which may reduce the generalizability of the analyzed patterns and editing guidelines. Although we identified a more systematic protocol using the characteristics of Most-Replayed values, it was not fully applied due to time constraints. Incorporating this protocol in future work would improve scalability and enable broader analysis across more videos, categories, and creators.

There may also be a gap between the extracted modality features and the actual composition of the original videos. For example, shot transition detection may not perfectly align with real transition points when boundaries are visually ambiguous or subtle. Such feature misalignment can affect feature-level statistics and may lead to editing guidelines that do not fully reflect the original video structure. However, using multiple complementary features across visual, text, and audio modalities can partially reduce the influence of errors from any single feature.

Another limitation is that the guideline generation process relies on category-level analysis rather than fully video-specific analysis. In our prompting framework, the user provides the target category, video description, and desired editing direction, while the statistically grounded analysis is retrieved at the category level. This allows the guidelines to reflect broad editing patterns for each content type, but may be less flexible for videos with unusual narrative structures, specific editing goals, or atypical multimodal patterns. Future work can address this limitation by incorporating video-specific feature extraction, editor preferences, and iterative refinement mechanisms.

10 Conclusion and Recommendations

This project investigated how highlight regions can be analyzed and used for practical video editing guidance. We constructed **KoSum**, a Korean YouTube Video Summarization benchmark based on recent videos from 2024 to 2026. With KoSum, we analyzed highlight regions from two perspectives: modality attention weights from a triple modality model and statistical patterns of modality specific features extracted from the original videos. The results show that highlight patterns differ across content categories, suggesting that viewer attention is shaped by category specific combinations of visual, textual, and audio cues.

The main outcome of this project is a **data-analysis-conditioned prompting framework** for generating practical video editing guidelines. Instead of relying only on a general user prompt, the proposed framework augments the prompt with statistically grounded editing cues derived from KoSum. This allows an LLM to generate more concrete and actionable guidelines for creators, such as how to adjust visual dynamics, subtitle density, audio emphasis, or pacing strategies according to the target content category. In this sense, KoSum serves not only as a benchmark for video summarization and highlight detection, but also as an analysis resource for creator oriented editing support.

For practical application, the proposed framework can be extended into an editing assistant that provides category-aware recommendations during planning or post-production. Beyond current text-based LLMs, the analysis-conditioned cues derived from KoSum can also serve as useful conditioning signals for future multimodal agents or dedicated video editing models that directly understand and manipulate video content. We expect that KoSum, its extracted multimodal features, and the proposed analysis protocol can support future studies on highlight detection, video summarization, multimodal understanding, and practical editing assistance.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2, 21

[2] Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Dernoncourt, and Joon Son Chung. Scaling up video summarization pretraining with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8332–8341, 2024. 1, 2, 21

[3] Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, et al. Exaone deep: Reasoning enhanced language models. *arXiv preprint arXiv:2503.12524*, 2025. 35

[4] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019. 7

[5] Vladislav Bargatin, Egor Chistov, Alexander Yakovenko, and Dmitriy Vatolin. Memfof: High-resolution training for memory-efficient multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8187–8196, 2025. 7

[6] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009. 25

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 23

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 8

[9] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 4, 8, 24

[10] Yaowei Guo, Jiazheng Xing, Xiaojun Hou, Shuo Xin, Juntao Jiang, Demetri Terzopoulos, Chenfanfu Jiang, and Yong Liu. Cfsum: A transformer-based multi-modal video summarization framework with coarse-fine fusion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1, 2, 21

[11] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. 3, 21

[12] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987. 9

[13] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878, 2023. 2, 6, 21

[14] Eghbal Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36:43918–43930, 2023. 6

[15] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717, 2019. 21

[16] Jong-Woo Jun. Effects of uses and gratifications of youtube and content orientation on youtube uses. *Journal of Outdoor Advertising Research*, 18(2):5–21, 2021. 21

[17] MinJeong Kang, Eun-Ju Jeong, and Hae-Yoon Cho. The immersion factors and characteristics of youtube channels for generation z. *The Journal of the Korea Contents Association*, 20(2):150–161, 2020. 21

[18] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. 6

[19] Ji Won Kim and Cheol Park. Differences of consumer responses according to characteristics of shopping curation contents in short-form video. *Journal of Product Research*, 43(4):1–11, August 2025. 21

[20] Kwanseok Kim, Jaehoon Hahm, Sumin Kim, Jinhwan Sul, Byunghak Kim, and Joonseok Lee. Summdiff: Generative modeling of video summarization with diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15096–15106, 2025. 5, 6

[21] Minsun Kim, Dawon Lee, and Junyong Noh. Generating highlight videos of a user-specified length using most replayed data. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2025. 3

[22] Sumin Kim, Hyemin Jeong, Mingu Kang, Yejin Kim, Yoori Oh, and Joonseok Lee. Triplesumm: Adaptive triple-modality fusion for video summarization. *arXiv preprint arXiv:2603.01169*, 2026. 1, 2, 3, 4, 5, 6, 21, 23, 24

[23] Shamit Lal, Shivam Duggal, and Indu Sreedevi. Online video summarization: Predicting future to better summarize present. In *2019 IEEE Winter Conference on applications of computer vision (WACV)*, pages 471–480. IEEE, 2019. 21

[24] Kang-You Lee and Dong-Kyoo Sung. Factors influencing on the flow and satisfaction of youtube users. *The Journal of the Korea Contents Association*, 18(12):660–675, 2018. 21

[25] Min Jung Lee, Dayoung Gong, and Minsu Cho. Video summarization with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18981–18991, 2025. 1, 2, 21

[26] 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. 2

[27] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5584–5593, 2023. 21

[28] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3042–3051, 2022. 1, 2, 21

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4, 23

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[31] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015. 8

[32] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021. 1, 2, 4, 21

[33] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7596–7604, 2019. 6

[34] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014. 5

[35] Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, et al. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921, 2024. 2, 3

[36] 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*, pages 8748–8763. PmLR, 2021. 23

[37] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. 7

[38] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025. 6

[39] Jaewon Son, Jaehun Park, and Kwangsu Kim. Csta: Cnn-based spatiotemporal attention for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18856, 2024. 6

[40] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 2, 3, 21

[41] Tomoya Sugihara, Shuntaro Masuda, Ling Xiao, and Toshihiko Yamasaki. Language-guided self-supervised video summarization using text semantic matching considering the diversity of the video. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1, 2024. 2

[42] Jinhwan Sul, Jihoon Han, and Joonseok Lee. Mr. hisum: A large-scale dataset for video highlight detection and summarization. *Advances in Neural Information Processing Systems*, 36:40542–40555, 2023. 2, 3, 5, 6, 21

[43] Hacene Terbouche, Maryan Morel, Mariano Rodriguez, and Alice Othmani. Multi-annotation attention model for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2023. 6, 21

[44] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pages 1–6. IEEE, 2024. 6

[45] 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*, 30, 2017. 1

[46] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016. 5, 21

[47] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871, 2017. 21

[48] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7405–7414, 2018. 21

[49] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Tth-rnn: Tensor-train hierarchical recurrent neural network for video summarization. *IEEE Transactions on Industrial Electronics*, 68(4):3629–3637, 2020. 21

[50] Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999. 6

Appendix

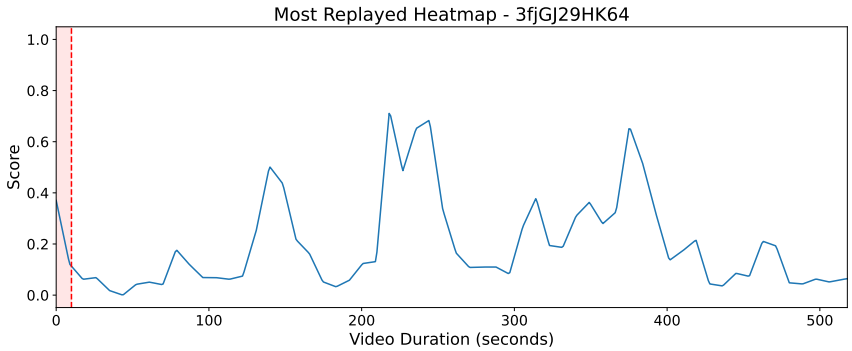
A Related Work

Video Summarization and Highlight Detection Early approaches to video summarization relied on recurrent model such as RNN or LSTM to capture temporal dependencies [46, 47, 48, 49, 23, 15]. However, these models suffer from limited capability in modeling long range dependencies and often struggle with long videos. To address this limitation, Transformer architectures have been widely adopted, enabling direct modeling of global temporal relationships [32, 2, 43, 13, 25].
Recent works further extend these approaches by incorporating multimodal information. For example, CFSum [10] leverages visual, audio, and textual features at different levels of granularity to capture complementary multimodal information. Despite these advances, most existing methods primarily focus on predictive performance and do not analyze why certain segments are selected as highlights. As a result, they provide limited interpretability and are less suitable for deriving practical editing guidelines.

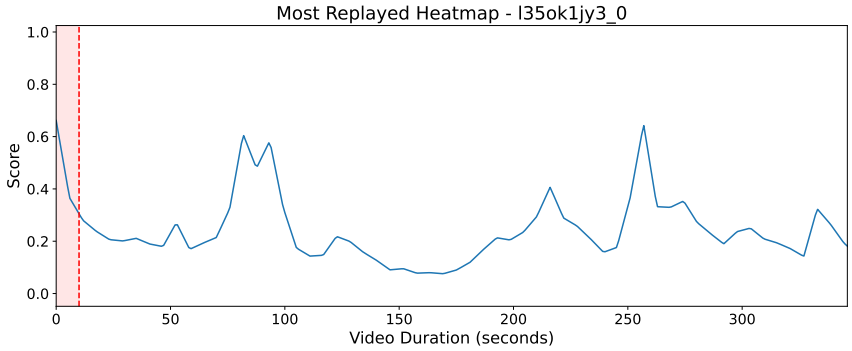
Multimodal Modeling Multimodal learning has been actively studied to improve video understanding. Many prior works [27, 2, 32, 13, 28] adopt bimodal settings, where visual features serve as the primary modality and additional signals such as text or audio are incorporated. While these approaches provide useful complementary information, they often fail to capture the full interaction among visual, audio, and textual modalities.
In practice, different modalities contribute differently depending on the content type. For example, visual dynamics are critical in sports videos, while audio cues are more important in music videos, and textual information plays a key role in dialogue driven content such as podcasts. Motivated by this observation, we adopt a triple modality framework [22] to analyze modality specific contributions and their roles in highlight formation.

YouTube Content Analysis Several studies [24, 17, 16, 19] analyze user behavior on video platforms using small scale surveys, video level metrics such as view count, watch time, and click through rate, or visual presentation factors such as thumbnails. These approaches provide useful insights into user preference, overall video performance, and pre watching engagement. However, they are limited in capturing temporal variations of viewer interest and do not directly indicate which segments attract repeated viewing.
In contrast, our work uses Most-Replayed signals as weak supervision for estimating temporally localized viewer interest. These signals provide segment level information about repeated viewing behavior, allowing us to identify candidate highlight regions without dense manual annotation. Based on these regions, we further analyze multimodal factors associated with viewer attention across diverse content categories.

Benchmark Dataset Existing benchmark datasets for video summarization have several limitations, as summarized in Table 1. Early standard datasets such as SumMe [11] and TVSum [40] are small in scale and rely on annotations from eqaul or fewer than 20 human raters per video. Due to the inherent subjectivity of summarization, such limited annotations can introduce bias, potentially resulting overfitting and degrading evaluation reliability.
To address this issue, datasets such as Mr.HiSum [42] and MoSu [22] leverage large-scale user interaction signals, particularly Most-Replayed statistics, which aggregate preferences from a large number of viewers and provide broader and more scalable supervision. However, both datasets are built upon YouTube-8M [1], which is based on videos collected prior to 2014, limiting their ability to reflect recent content trends. In addition, Mr.HiSum is limited to visual modality, and MoSu, while multimodal, focuses on English content. Furthermore, existing benchmarks do not explicitly support language diversity or category-aware analysis. In addition, many datasets are constructed from older videos, and some samples are no longer accessible due to deletion or restricted availability.
To address these limitations, we propose KoSum, a Korean YouTube benchmark constructed from recent videos between 2024 and 2026, as shown in Table 1. KoSum uses Most Replayed signals as weak supervision for importance score construction and supports analysis across visual, audio, and text modalities for practical editing guidance.



(a) Repeated viewing patterns around initial edited highlight segments.



(b) Initial engagement spike at the beginning of the video.

Figure 8: Visualization of Most-Replayed bias. Viewer engagement is typically high at the beginning of a video and decreases over time, while highlight segments may attract repeated viewing. For better visibility, the plot is truncated to the first 60% of the video.

B Most-Replayed Bias

At the beginning of a video, viewer engagement is typically high but decreases rapidly over time. This phenomenon can be attributed to several factors. Viewers often decide whether to continue watching within the first few seconds, leading to disproportionately high interactions at the early stage of the video. To accommodate this behavior, editors frequently place short and attention-grabbing highlight clips at the beginning, intentionally showcasing the most engaging moments to retain viewers.
In addition, some viewers remain briefly due to buffering or encoding delays, while others skim through the video to quickly assess its overall quality or mood. In many cases, viewers jump directly to highlight segments referenced in comments or visible in the timeline. As a result, viewing behavior is not uniformly distributed across the video, leading to a bias in Most-Replayed signals. This bias can be observed in Figure 8.

C KoSum Category Distribution

Figure 9 shows the category distribution of the KoSum dataset. KoSum consists of 14 subcategories, with 50 videos collected for each subcategory. The pie chart provides an overview of how the dataset is organized across major content categories and fine-grained subcategories.

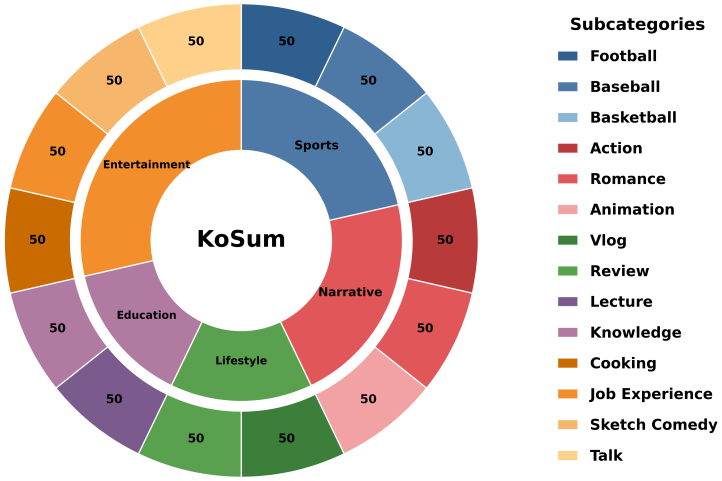


Figure 9: Category distribution of the KoSum dataset. Each subcategory contains 50 videos.

D Benchmark Modality Feature Details

D.1 Visual Features

For visual features v_i, we use a CLIP [36] visual encoder based on Vision Transformer [7][4]. To maintain temporal alignment with other modalities and reduce computational cost, we extract visual features at a rate of 1 FPS.

D.2 Text Features

For textual features l_i, we use a pretrained multilingual RoBERTa model [29][5] to obtain sentence embeddings, since KoSum primarily consists of Korean content. Unlike MoSu [22], which adopts a monolingual RoBERTa model, we use a multilingual encoder to better handle Korean text. For each timestamp t, we use the output embedding of the [CLS] token as the sentence level representation.
To extract one textual feature per second and maintain temporal alignment with other modalities, subtitle segments with different durations must be aligned to a common timeline. Since text annotations are provided with start and end timestamps, multiple subtitle segments may overlap within the same second. In such cases, we merge their contents into a single representation. An example of this temporal text alignment process is shown below.

Example: Temporal Text Alignment

```
Unaligned sequence:
[
  {start: 1.6, end: 2.5, text: "hello"},
  {start: 2.1, end: 3.5, text: "my name is"}
]

Aligned sequence:
[
    t_0: [PAD],
    t_1: "hello",
    t_2: "hello my name is",
    t_3: "my name is"
]
```

[4] openai/clip-vit-base-patch32
[5] FacebookAI/xlm-roberta-base

Table 5: Comparison of attention statistics under different highlight selection strategies.

(a) Top-$R\%$ highlight samples

Layer	JSD ($\times 10^{-4}$) ↑	E_{gap} ↓	$\mathrm{V}_{\mathcal{H}}$($\times 10^{-2}$) ↓
First	1.917	-0.015	1.665
Last	**4.230**	**-0.020**	<u>1.247</u>
All	<u>1.949</u>	<u>-0.020</u>	**0.902**

(b) Watershed highlight segments

Layer	JSD ($\times 10^{-4}$) ↑	E_{gap} ↓	$\mathrm{V}_{\mathcal{H}}$($\times 10^{-2}$) ↓
First	<u>2.002</u>	<u>-0.016</u>	1.640
Last	**2.155**	-0.010	<u>1.247</u>
All	0.669	**-0.016**	**0.902**

D.3 Audio Features

For audio features a_i, we adopt an Audio Spectrogram Transformer [9][6]. To obtain a representation at time t, we extract an audio segment from five seconds before t to five seconds after t and feed it into the model. This produces a temporally centered embedding for each timestamp, following the centered window approach [22]. As a result, all modalities have the same sequence length.

E Attention Map

Multiple layers produce attention maps, and each layer captures cross modal interactions at a different level. To choose a suitable attention representation for analysis, we compare three options: the first layer, the last layer, and the average attention across all layers. For this comparison, we divide timestamps into highlight and nonhighlight sets based on the Most Replayed scores. Following the observation that replay peaks occupy only a small portion of a video, we label the top $\mathcal{R}\%$ timestamps as highlights and set $\mathcal{R} = 10$ in this attention analysis. We intentionally use a small $\mathcal{R}$ to focus on the most prominent highlight regions, allowing the analysis to better capture strong highlight specific modality activation patterns.

Formally, let $\mathcal{T} = \{1, \ldots, T\}$ denote the full set of time indices. For the top $\mathcal{R}\%$ based attention analysis, we define the highlight and nonhighlight sets as $\mathcal{H}^{\text{top}}$ and $\mathcal{N}^{\text{top}}$, respectively. Specifically, $\mathcal{H}^{\text{top}}$ contains the timestamps with the highest $\mathcal{R}\%$ Most Replayed scores, and $\mathcal{N}^{\text{top}} = \mathcal{T} \setminus \mathcal{H}^{\text{top}}$, where $|\mathcal{H}^{\text{top}}| = \left\lceil \frac{\mathcal{R}}{100} |\mathcal{T}| \right\rceil$. For each timestamp t, let $\tilde{\mathbf{w}}_t \in \mathbb{R}^{|\mathcal{M}|}$ denote the modality attention vector normalized over the modality set $\mathcal{M} = \{\text{Visual}, \text{Text}, \text{Audio}\}$. For each attention representation, we compute the average modality attention distributions over $\mathcal{H}^{\text{top}}$ and $\mathcal{N}^{\text{top}}$ and measure their Jensen Shannon divergence (JSD), where a higher value indicates clearer separation between highlight and nonhighlight regions. We further use two auxiliary metrics to assess attention concentration and stability. The entropy gap is defined as $E_{\text{gap}} = E_{\mathcal{H}^{\text{top}}} - E_{\mathcal{N}^{\text{top}}}$, where a negative value indicates that highlight timestamps have more concentrated modality attention. Finally, $\mathrm{V}_{\mathcal{H}^{\text{top}}}$ measures the average modality wise variance of attention weights within $\mathcal{H}^{\text{top}}$, where lower values indicate more stable modality allocation across highlight timestamps.

As shown in Table 5, the last layer achieves the highest JSD under both Top $R\%$ selection and watershed segmentation, indicating the clearest separation between highlight and non highlight modality distributions. Although the layer averaged attention shows the lowest highlight variance, its JSD is substantially lower under watershed segmentation, suggesting weaker discriminative power for identifying modality differences between highlight and non highlight regions. The last layer also maintains competitive entropy gap and variance values, especially compared with the first layer. Therefore, considering JSD as the primary criterion and entropy gap and variance as auxiliary stability criteria, we use the last layer attention in Sec. 7 to analyze modality activation and identify the dominant modality in category specific highlights.

F Attention Statistics Computation

Using the notation defined in Appendix E, we compute attention statistics for a given highlight definition. Let $\mathcal{H}$ and $\mathcal{N}$ denote the highlight and nonhighlight timestamp sets, respectively. This formulation is general and can be applied regardless of how the highlight regions are selected.

For each timestamp t, attention weights are first averaged over the selected layer set and attention heads, and then normalized across modalities. Let $\mathcal{M} = \{\text{Visual}, \text{Text}, \text{Audio}\}$ denote the modality set and

[6] MIT/ast-finetuned-audioset-10-10-0.4593

$M = |\mathcal{M}|$. The normalized modality attention vector is defined as $\tilde{\mathbf{w}}_t \in \mathbb{R}^M$, where

$$\tilde{w}_t^{(m)} = \frac{w_t^{(m)}}{\sum_{k \in \mathcal{M}} w_t^{(k)} + \epsilon}, \quad m \in \mathcal{M}.$$

Jensen Shannon Divergence The average modality attention distributions for highlight and nonhighlight timestamps are computed as

$$P_{\mathcal{H}} = \frac{1}{|\mathcal{H}|} \sum_{t \in \mathcal{H}} \tilde{\mathbf{w}}_t, \quad P_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{t \in \mathcal{N}} \tilde{\mathbf{w}}_t.$$

With

$$Q = \frac{1}{2} \left(P_{\mathcal{H}} + P_{\mathcal{N}} \right),$$

the Jensen Shannon divergence is computed as

$$\mathrm{JS} \left(P_{\mathcal{H}} \parallel P_{\mathcal{N}} \right) = \frac{1}{2} \mathrm{KL} \left(P_{\mathcal{H}} \parallel Q \right) + \frac{1}{2} \mathrm{KL} \left(P_{\mathcal{N}} \parallel Q \right),$$

where

$$\mathrm{KL}(P \parallel Q) = \sum_{m \in \mathcal{M}} P^{(m)} \log \frac{P^{(m)} + \epsilon}{Q^{(m)} + \epsilon}.$$

Entropy Gap Entropy is computed for each timestamp and then averaged within each set. The highlight and nonhighlight entropy values are defined as

$$E_{\mathcal{H}} = \frac{1}{|\mathcal{H}|} \sum_{t \in \mathcal{H}} \left[- \sum_{m \in \mathcal{M}} \tilde{w}_t^{(m)} \log \left(\tilde{w}_t^{(m)} + \epsilon \right) \right],$$

$$E_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{t \in \mathcal{N}} \left[- \sum_{m \in \mathcal{M}} \tilde{w}_t^{(m)} \log \left(\tilde{w}_t^{(m)} + \epsilon \right) \right].$$

The entropy gap is defined as

$$E_{\mathrm{gap}} = E_{\mathcal{H}} - E_{\mathcal{N}}.$$

A negative value indicates that highlight timestamps have more concentrated modality attention than non-highlight timestamps.

Highlight Variance The highlight variance measures the stability of modality allocation within highlight timestamps. It is computed as the mean of modality wise variances across $\mathcal{H}$:

$$\mathrm{V}_{\mathcal{H}} = \frac{1}{M} \sum_{m \in \mathcal{M}} \mathrm{Var} \left(\left\{ \tilde{w}_t^{(m)} \right\}_{t \in \mathcal{H}} \right).$$

A lower value indicates more stable modality attention patterns across highlight timestamps.

G Feature Extraction Details

G.1 Visual Dynamics (Visual)

Shot Transition We detect shot transition points using the algorithm described in Algorithm 1. For edge extraction, we employ Canny edge detection with hysteresis thresholding [6]. In addition, we use histogram bins of $(32, 32, 32)$, resulting in $32^3 = 32{,}768$ dimensional histogram features. Histogram differences and edge change ratios are computed at 4 FPS, which is sufficient to capture most editing cuts while maintaining computational efficiency for long videos. We select transition candidates whose scores exceed a percentile based threshold (top 10%) and remove detections that are temporally too close by applying a minimum boundary gap of 1 second, preventing multiple detections of the same transition.

Algorithm 1 Shot Transition Detection

Require: Video V
Ensure: Shot-transition timestamps $\mathcal{T}$
1: Sample frames from V at a fixed FPS
2: Compute histogram differences and edge change ratios (ECR) between adjacent frames
3: Normalize each score sequence independently
4: Compute a weighted sum of the normalized scores
5: Apply Gaussian smoothing to the combined score sequence
6: Determine a threshold using a percentile of the smoothed scores
7: Select local maxima above the threshold as transition candidates
8: Remove candidates that are closer than a minimum temporal gap
9: **return** $\mathcal{T}$

G.2 Visual Dynamics (Visual)

Motion Dynamics We compute optical flow at 4 FPS, which is sufficient to satisfy the assumptions of optical flow estimation while maintaining temporal consistency. However, optical flow extraction is computationally expensive. In our experiments, processing the *Entertainment* category alone required approximately 40 hours. Therefore, instead of reducing the sampling rate, which could harm motion estimation quality, we resize videos so that the shortest side is 480 pixels to improve computational efficiency.

H Watershed Grouping

We provide examples of the proposed watershed based highlight segmentation strategy in Figure 10. The key idea is to generate sufficient candidate modes by setting k to a large value and then remove unreliable candidates using a minimum peak threshold, enabling adaptive highlight detection without fixing the final number of detected segments. Although mAP50 in highlight detection is commonly defined using top highlight segments rather than point wise scores, we approximate this setting by analyzing the top 50 percentile Most-Replayed points, which results in an average value of 0.391. Based on this observation, we set the minimum peak threshold in watershed grouping to $\tau = 0.4$. Furthermore, through empirical inspection across all videos, we observe that the number of clear highlight modes rarely exceeds 12. To reduce the risk of missing valid highlight regions, we set the number of candidate modes to $k = 15$ with a small margin.

I Detailed Analysis

This section reports detailed statistical evidence for the subcategory analysis in Sec. 7. In the following tables, Component denotes the signal on which the metric is computed: a dash (—) marks the feature's overall signal, whereas a named entry (e.g., Pitch, Dispersion, Δ-emb.) marks a specific sub-channel of that feature. H and N denote the average values in highlight and nonhighlight regions, respectively. $\Delta = H - N$ denotes the effect size in terms of mean difference between the two regions. Cons. denotes the consistency ratio (the fraction of videos whose per-video effect agrees in sign with the overall effect), r_b denotes rank-biserial correlation, and q (q-value) denotes the FDR-corrected p-value. For metrics such as local contrast or boundary based statistics, the values are computed through local comparisons rather than direct comparisons between highlight and nonhighlight regions. Therefore, separate highlight and nonhighlight averages are not available for these metrics. In such cases, the corresponding entries are marked as N/A.

Cooking Table 6 shows that Cooking highlights are primarily explained by coherent visual scenes, visible faces, and reaction based audio cues. The strongest result is visual consistency, where purity increases from 0.498 in nonhighlight regions to 0.668 in highlight regions, with $\Delta = 0.170$, Cons. $= 0.86$, $r_b = 0.898$, and $q < 10^{-6}$. This means that cooking highlights are usually located inside a single coherent visual segment rather than being formed through frequent scene changes. Face features also show a reliable positive pattern. The face count increases from 1.313 to 1.439 with $r_b = 0.575$, and the local contrast reaches $r_b = 0.609$, indicating that highlight regions tend to include more visible faces than their surroundings. The face dispersion change is even slightly stronger, with $r_b = 0.617$ and $q = 2.4{\times}10^{-4}$, suggesting that highlights also coincide with faces becoming more spatially spread across the frame. Semantic motion

Table 6: Detailed statistical evidence for Cooking.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.116	0.114	0.003	0.68	0.423	0.017
Face Features	—	Level	1.439	1.313	0.126	0.80	0.575	3.9×10^{-4}
Face Features	—	Local contrast	N/A	N/A	0.138	0.72	0.609	3.0×10^{-4}
Face Features	Dispersion	Change	0.050	0.046	0.005	0.76	0.617	2.4×10^{-4}
Visual Consistency	—	Purity	0.668	0.498	0.170	0.86	0.898	$< 10^{-6}$
Subtitle Density	—	Change	1.127	1.083	0.044	0.66	0.325	0.135
Audio Novelty	—	Local contrast	N/A	N/A	0.011	0.74	0.493	0.006
Audio Novelty	MFCC-Δ	Local contrast	N/A	N/A	0.929	0.70	0.503	0.005
Laughter and Applause	—	Level	0.006	0.004	0.003	0.70	0.685	1.0×10^{-5}
Laughter and Applause	—	Local contrast	N/A	N/A	0.003	0.80	0.736	1.0×10^{-5}

Table 7: Detailed statistical evidence for Job Experience.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.146	0.139	0.007	0.80	0.699	5.0×10^{-6}
Curvature	Δ-emb.	Local contrast	N/A	N/A	0.011	0.84	0.805	$< 10^{-6}$
Face Features	—	Level	2.319	2.187	0.132	0.68	0.448	0.016
Face Features	Dispersion	Change	0.061	0.058	0.003	0.72	0.503	0.003
Visual Consistency	—	Purity	0.673	0.569	0.104	0.84	0.885	$< 10^{-6}$
Visual Consistency	—	Fragmentation	0.080	0.108	-0.028	0.76	-0.724	5.0×10^{-6}
Motion Dynamics	Horiz. u	Local contrast	N/A	N/A	0.769	0.70	0.525	0.003
Subtitle Density	—	Level	8.646	9.192	-0.546	0.74	-0.585	4.0×10^{-4}
Subtitle Density	—	Change	0.969	0.899	0.070	0.72	0.605	4.0×10^{-4}
Text Region Density	—	Level	1.773	1.850	-0.078	0.70	-0.537	0.002
Speech Dynamics	Pitch	Level	76.886	70.986	5.900	0.76	0.553	0.001
Laughter and Applause	Laughter	Change	0.002	0.001	0.001	0.82	0.857	$< 10^{-6}$

contributes only weakly: the adjacent-frame embedding change rises from 0.114 to 0.116 with $r_b = 0.423$ and $q = 0.017$, so local semantic velocity provides modest auxiliary support rather than a primary cue.
The audio results show that laughter and applause are another strong signal. The combined reaction level increases from 0.004 to 0.006, and its local contrast shows Cons. $= 0.80$ with $r_b = 0.736$ and $q = 1.0\times10^{-5}$. This supports the interpretation that laughter centered reactions are strongly associated with cooking highlights. Audio novelty is weaker than reaction cues but still meaningful in local contrast, with $\Delta = 0.011$, $r_b = 0.493$, and $q = 0.006$, and the mfcc-delta component shows a comparable local effect ($r_b = 0.503$, $q = 0.005$), suggesting that local acoustic changes can support highlight detection. In contrast, subtitle density is not reliable: although the subtitle change metric trends positive ($\Delta = 0.044$, $r_b = 0.325$), its corrected significance is weak with $q = 0.135$. Therefore, cooking highlights are best described as coherent cooking scenes with more visible and dispersed faces and audible social reactions and laughter, rather than text heavy or cut heavy segments.

Job Experience Table 7 shows that Job Experience highlights combine coherent task scenes with meaningful semantic changes. Visual Consistency is the most reliable cue: purity increases from 0.569 in non-highlight regions to 0.673 in highlight regions, with $r_b = 0.885$ and $q < 10^{-6}$, while fragmentation decreases from 0.108 to 0.080 with $r_b = -0.724$. This indicates that job highlights are not scattered across many scene boundaries, but are usually contained within a stable and coherent work scene. The semantic-change signal is equally strong. The adjacent-frame embedding change rises from 0.139 to 0.146 with $r_b = 0.699$, and its local contrast reaches Cons. $= 0.84$ and $r_b = 0.805$ with $q < 10^{-6}$. This suggests that job highlights often coincide with the appearance of a new object, action, or work-related situation. Face features add a consistent positive pattern, with the face count rising by 0.132 ($r_b = 0.448$) and the face-dispersion change reaching $r_b = 0.503$, indicating that highlight moments tend to show more or more spatially spread faces than their surroundings.
Motion Dynamics is weaker than scene coherence, but its horizontal-flow local contrast is still significant with $\Delta = 0.769$ and $q = 0.003$, showing that lateral motion becomes more salient relative to the local context around highlights. The audio evidence is led by laughter and applause: its laughter-component change reaches $r_b = 0.857$ with $q < 10^{-6}$, the strongest audio result in this subcategory. The pitch component of Speech Dynamics provides a clearer vocal signal, rising from 70.986 to 76.886 with $r_b = 0.553$. The text pattern is distinctive and bidirectional. Subtitle density level decreases from 9.192 to 8.646 ($r_b = -0.585$), yet subtitle change increases from 0.899 to 0.969 ($r_b = 0.605$), and on-screen text-box count likewise decreases from 1.850 to 1.773 ($r_b = -0.537$). Therefore, job highlights are best described as coherent task scenes marked by visible semantic change, laughter and pitch variation, and concise but rapidly changing text.

Table 8: Detailed statistical evidence for Sketch Comedy.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Local contrast	N/A	N/A	0.008	0.70	0.485	0.005
Face Features	Dispersion	Change	0.040	0.035	0.005	0.70	0.500	0.005
Visual Consistency	—	Purity	0.683	0.602	0.081	0.64	0.460	0.010
Visual Consistency	—	Fragmentation	0.090	0.125	-0.034	0.68	-0.573	0.001
Subtitle Density	—	Level	8.247	9.090	-0.843	0.80	-0.614	2.6×10^{-4}
Subtitle Density	—	Local contrast	N/A	N/A	-0.654	0.70	-0.482	0.004
Text Region Density	Box area	Level	1.791	2.139	-0.349	0.76	-0.587	5.6×10^{-4}
Speech Dynamics	—	Local contrast	N/A	N/A	-0.013	0.78	-0.598	4.1×10^{-4}
Speech Dynamics	Speak. rate	Local contrast	N/A	N/A	-0.125	0.72	-0.528	0.003
Laughter and Applause	—	Level	0.004	0.004	-0.000	0.54	-0.037	0.893

Table 9: Detailed statistical evidence for Talk.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.095	0.087	0.007	0.84	0.839	$< 10^{-6}$
Face Features	—	Change	0.541	0.460	0.080	0.74	0.630	1.6×10^{-4}
Visual Consistency	—	Boundary rate	0.060	0.047	0.014	0.70	0.624	3.2×10^{-4}
Motion Dynamics	—	Level	1.505	1.182	0.323	0.80	0.661	6.0×10^{-5}
Motion Dynamics	—	Change	1.267	1.011	0.256	0.76	0.584	3.0×10^{-4}
Shot Transition	—	Rate	0.170	0.135	0.034	0.86	0.807	$< 10^{-6}$
Subtitle Density	—	Level	9.487	10.291	-0.804	0.80	-0.826	$< 10^{-6}$
Subtitle Density	—	Change	0.755	0.680	0.075	0.80	0.754	1.0×10^{-6}
Speech Dynamics	Pitch	Level	48.654	41.734	6.920	0.82	0.749	3.0×10^{-6}
Audio Novelty	—	Level	0.659	0.674	-0.015	0.76	-0.711	9.0×10^{-6}
Laughter and Applause	—	Level	0.010	0.007	0.003	0.72	0.576	6.5×10^{-4}

Sketch Comedy Table 8 shows that Sketch Comedy highlights have a different structure from cooking and job experience, leaning on moderate scene coherence and a marked reduction in speech and text load. Visual consistency is reliable but only moderate in strength. Purity increases from 0.602 in nonhighlight regions to 0.683 in highlight regions, with $r_b = 0.460$ and $q = 0.010$, while fragmentation decreases from 0.125 to 0.090, with $r_b = -0.573$ and $q = 0.001$. This indicates that sketch highlights tend to sit inside a coherent scene rather than across many boundaries, but the effect is weaker than in cooking or job experience. Semantic change provides only auxiliary support: the main curvature signal is not reliable, yet the adjacent-frame embedding change measured by local contrast reaches $\Delta = 0.008$, Cons. $= 0.70$, $r_b = 0.485$, and $q = 0.005$, and the face-dispersion change is also significant, with $r_b = 0.500$ and $q = 0.005$, suggesting that the spatial spread of faces shifts modestly around highlights.

The most distinctive evidence is the negative speech and text pattern. Speech dynamics local contrast has $\Delta = -0.013$, Cons. $= 0.78$, $r_b = -0.598$, and $q = 4.1\times10^{-4}$, and the speaking-rate component local contrast similarly decreases with $r_b = -0.528$ and $q = 0.003$, indicating that sketch highlights tend to occur when speech slows or pauses rather than when it accelerates. Text load drops as well: subtitle level decreases from 9.090 to 8.247 with $r_b = -0.614$, its local contrast confirms the same direction with $r_b = -0.482$, and the text-box area component falls from 2.139 to 1.791 with $r_b = -0.587$. Notably, the reaction cue is essentially inert here: laughter and applause level shows almost no difference between regions, with $r_b = -0.037$ and $q = 0.893$, so laughter and applause are not reliable indicators of sketch highlights. Therefore, sketch highlights are best described as moments of reduced speech rate and reduced on-screen text, supported by moderate scene coherence rather than by audience reaction cues.

Talk Table 9 shows that Talk has the strongest and most broadly multimodal highlight signature among the entertainment subcategories. Visual dynamics is highly active. Motion level increases from 1.182 to 1.505 with $r_b = 0.661$, and motion change increases from 1.011 to 1.267 with $r_b = 0.584$. Shot transition rate also rises from 0.135 to 0.170, with Cons. $= 0.86$ and $r_b = 0.807$, one of the strongest effects in this subcategory. These results indicate that talk highlights rely on dynamic visual pacing, more motion, and more frequent cuts. The adjacent-frame semantic change reinforces this: the curvature delta component level increases from 0.087 to 0.095 with Cons. $= 0.84$ and $r_b = 0.839$, suggesting that large frame-to-frame embedding changes are strongly associated with highlights. Face features add a consistent cue, with face change increasing from 0.460 to 0.541 and $r_b = 0.630$, reflecting speaker switches, participant reactions, or camera focus changes.

Unlike cooking and job experience, visual consistency here is led by boundary rate rather than purity. The boundary rate increases from 0.047 to 0.060 with $r_b = 0.624$, meaning talk highlights are not necessarily contained within one coherent scene but instead sit near visually dynamic transitions. The audio pattern is equally distinctive. The speech pitch component increases markedly from 41.734 to 48.654 with

Table 10: Detailed statistical evidence for Action.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.142	0.136	0.006	0.76	0.555	9.0×10^{-4}
Curvature	Δ-emb.	Local contrast	N/A	N/A	0.007	0.68	0.468	0.003
Visual Consistency	—	Purity	0.629	0.477	0.152	0.82	0.802	$< 10^{-6}$
Visual Consistency	—	Boundary rate	0.055	0.048	0.007	0.70	0.507	0.004
Motion Dynamics	—	Level	6.856	6.956	-0.100	0.44	-0.071	0.985
Subtitle Density	—	Level	4.790	5.198	-0.408	0.66	-0.417	0.034
Text Region Density	—	Change	0.335	0.297	0.037	0.74	0.553	0.001
Text Region Density	Box area	Change	1.691	1.353	0.338	0.68	0.500	0.005
Audio Novelty	MFCC-Δ	Change	7.656	7.370	0.287	0.68	0.398	0.041
Laughter and Applause	Applause	Level	0.000	0.001	-0.000	0.62	-0.388	0.051

Table 11: Detailed statistical evidence for Animation.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	—	Level	1.795	1.779	0.016	0.78	0.680	2.0×10^{-5}
Curvature	Δ-emb.	Level	0.140	0.129	0.011	0.82	0.898	$< 10^{-6}$
Visual Consistency	—	Purity	0.675	0.497	0.179	0.86	0.827	$< 10^{-6}$
Visual Consistency	—	Neg. purity	0.308	0.209	0.099	0.76	0.548	0.001
Motion Dynamics	—	Level	3.726	3.202	0.524	0.70	0.500	0.003
Motion Dynamics	—	Change	3.978	3.305	0.673	0.80	0.660	6.3×10^{-5}
Shot Transition	—	Rate	0.111	0.088	0.023	0.72	0.661	2.1×10^{-5}
Subtitle Density	—	Level	5.013	5.188	-0.175	0.72	-0.393	0.048
Text Region Density	—	Change	0.272	0.238	0.034	0.74	0.525	0.003
Speech Dynamics	Pitch	Level	73.707	66.266	7.442	0.82	0.776	$< 10^{-6}$
Audio Novelty	—	Level	0.616	0.624	-0.009	0.70	-0.481	0.008
Laughter and Applause	—	Level	0.002	0.002	0.000	0.48	0.146	0.554

$r_b = 0.749$, and reaction cues rise from 0.007 to 0.010 with $r_b = 0.576$, showing stronger vocal pitch and stronger laughter and applause in highlights. In contrast, audio novelty decreases from 0.674 to 0.659 with $r_b = -0.711$, indicating that the acoustic background becomes more stable while speech and reactions become more salient. The text pattern is bidirectional: subtitle level decreases from 10.291 to 9.487 with $r_b = -0.826$, while subtitle change increases from 0.680 to 0.755 with $r_b = 0.754$. Therefore, talk highlights are best described as moments of dynamic visual pacing, speaker or focus changes, stable background audio with stronger pitch and reactions, and shorter but more rapidly changing subtitles.

Action Table 10 shows that Action highlights are driven by visual scene coherence combined with sharp adjacent-frame semantic change. Visual consistency is the strongest signal: purity increases from 0.477 in nonhighlight regions to 0.629 in highlight regions, with $\Delta = 0.152$, Cons. $= 0.82$, $r_b = 0.802$, and $q < 10^{-6}$. This means action highlights are usually contained within a single coherent scene rather than scattered across many segments. At the same time, boundary rate increases from 0.048 to 0.055 with $r_b = 0.507$, indicating that these coherent highlight scenes still tend to sit near a scene boundary. The adjacent-frame semantic change signal is also reliable: the curvature delta component level increases from 0.136 to 0.142 with $r_b = 0.555$, and its local contrast reaches $r_b = 0.468$ with $q = 0.003$. This suggests action highlights coincide with rapid frame-to-frame embedding changes, consistent with sudden movements or new events appearing on screen.
The remaining cues are weaker and partly negative. Subtitle density moves opposite to the highlight, with subtitle level decreasing from 5.198 to 4.790 ($r_b = -0.417$, $q = 0.034$), so highlights carry less spoken text. Text region density, in contrast, is meaningful through its change metrics: the on-screen text box count change increases from 0.297 to 0.335 with $r_b = 0.553$, and the box-area component change increases from 1.353 to 1.691 with $r_b = 0.500$, indicating that highlights coincide with growing or fluctuating on-screen text rather than a higher static text load. Notably, motion dynamics is not a reliable indicator despite the action setting: optical-flow level barely changes ($\Delta = -0.100$, $r_b = -0.071$, $q = 0.985$), showing that raw motion magnitude does not separate highlights from their surroundings. Audio novelty offers only borderline support through its mfcc-delta change component ($\Delta = 0.287$, $r_b = 0.398$, $q = 0.041$), while laughter and applause are likewise ineffective, with the applause component only borderline and negative ($r_b = -0.388$, $q = 0.051$). Therefore, action highlights are best described as coherent scenes marked by abrupt semantic change and fluctuating on-screen text, rather than by raw motion intensity or audience reactions.

Animation Table 11 shows that Animation highlights are driven by a combination of strong semantic motion and coherent scene structure rather than by social audio cues. The clearest visual signature comes from the embedding trajectory. The curvature level increases from 1.779 to 1.795 with $r_b = 0.680$, while

Table 12: Detailed statistical evidence for Romance.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Local contrast	N/A	N/A	0.008	0.72	0.520	0.002
Face Features	Face area	Change	0.021	0.018	0.003	0.64	0.451	0.015
Visual Consistency	—	Boundary rate	0.058	0.046	0.012	0.78	0.636	2.2×10^{-4}
Visual Consistency	—	Purity	0.662	0.578	0.083	0.72	0.475	0.005
Motion Dynamics	—	Level	3.392	2.927	0.465	0.68	0.539	0.002
Motion Dynamics	—	Local contrast	N/A	N/A	0.165	0.72	0.412	0.011
Shot Transition	—	Rate	0.125	0.102	0.022	0.68	0.512	0.003
Subtitle Density	—	Level	5.067	5.988	-0.921	0.78	-0.726	6.0×10^{-6}
Text Region Density	Box area	Level	0.944	1.799	-0.855	0.82	-0.710	1.2×10^{-5}
Speech Dynamics	—	Level	0.243	0.262	-0.019	0.78	-0.664	5.4×10^{-5}
Audio Novelty	—	Level	0.524	0.562	-0.038	0.84	-0.802	$< 10^{-6}$
Laughter and Applause	—	Level	0.003	0.002	0.002	0.62	0.327	0.066

the adjacent-frame semantic change is even stronger, rising from 0.129 to 0.140 with Cons. $= 0.82$ and $r_b = 0.898$, the single largest effect in this subcategory. This indicates that animation highlights occur when the visual content shifts rapidly in semantic space, that is, when new scenes, actions, or characters appear in quick succession. Scene structure reinforces this reading. Visual consistency purity increases from 0.497 to 0.675 with $r_b = 0.827$, and negative purity also increases from 0.209 to 0.308 with $r_b = 0.548$, showing that highlights sit inside a dominant, internally coherent scene even while the embedding moves quickly within it.

The remaining visual and audio evidence is consistent with fast, vivid action. Motion dynamics is reliably elevated, with the optical-flow change rising from 3.305 to 3.978 ($r_b = 0.660$) and the flow level from 3.202 to 3.726 ($r_b = 0.500$), and the shot-transition rate increases from 0.088 to 0.111 with $r_b = 0.661$, so highlights pair heavy motion with more frequent cuts. On the audio side, the pitch component of speech dynamics is the dominant cue, rising from 66.266 to 73.707 with Cons. $= 0.82$ and $r_b = 0.776$, indicating more expressive, higher-pitched voice acting in highlights, whereas audio novelty decreases from 0.624 to 0.616 with $r_b = -0.481$, suggesting a more stable acoustic background. Text behaves bidirectionally: subtitle level falls slightly from 5.188 to 5.013 ($r_b = -0.393$), while text-region density change increases from 0.238 to 0.272 ($r_b = 0.525$), pointing to fewer but more rapidly changing on-screen captions. Notably, the laughter-and-applause cue that drives several other subcategories is inert here, with $\Delta = 0.000$, $r_b = 0.146$, and $q = 0.554$. Therefore, animation highlights are best described as fast, semantically shifting action inside coherent scenes, marked by motion, frequent cuts, and expressive pitch rather than audience reactions.

Romance Table 12 shows that Romance highlights are driven by visually dynamic moments that are paired with a marked reduction in textual and acoustic load. On the visual side, the activity is broad. Motion level increases from 2.927 to 3.392 with $\Delta = 0.465$ and $r_b = 0.539$, and its local contrast is also reliable, with Cons. $= 0.72$ and $r_b = 0.412$, while shot transition rate rises from 0.102 to 0.125 with $r_b = 0.512$, indicating that romance highlights involve more on screen movement, locally salient motion, and more frequent cuts. Visual consistency contributes through both of its facets. Purity increases from 0.578 to 0.662 with $r_b = 0.475$, so highlights still sit inside coherent scenes, but boundary rate also increases from 0.046 to 0.058 with Cons. $= 0.78$ and $r_b = 0.636$, the strongest visual result. This combination suggests that highlights are anchored to coherent emotional scenes that are nonetheless punctuated by frequent segment boundaries rather than long static takes. Face features add a secondary cue through the face area component, whose change increases by 0.003 with $r_b = 0.451$, meaning that highlight regions tend to show larger or growing facial presence, consistent with close interpersonal framing.

The semantic and audio evidence reinforces this reading. The adjacent frame embedding change is meaningful in local contrast, with $\Delta = 0.008$, Cons. $= 0.72$, and $r_b = 0.520$, indicating that highlights coincide with locally elevated semantic velocity even though the main curvature signal is flat. In sharp contrast, the textual and acoustic channels are strongly suppressed. Subtitle density falls from 5.988 to 5.067 with $r_b = -0.726$, text box area drops from 1.799 to 0.944 with $r_b = -0.710$, speech dynamics declines from 0.262 to 0.243 with $r_b = -0.664$, and audio novelty decreases from 0.562 to 0.524 with $r_b = -0.802$, the single strongest effect in the subcategory. Together these point to highlights that are quieter, less talky, and visually less cluttered with text. Laughter and applause, which one might expect to flag emotional peaks, is not a reliable indicator here: reaction level rises only slightly to 0.003 and remains nonsignificant at $q = 0.066$. Therefore, romance highlights are best described as visually active and emotionally focused moments, with closer faces and more cuts, that are accompanied by reduced speech, sparser on screen text, and a calmer, more stable acoustic background rather than by laughter driven reactions.

Table 13: Detailed statistical evidence for Baseball.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	—	Level	1.866	1.860	0.006	0.66	0.357	0.055
Face Features	Dispersion	Change	0.074	0.070	0.004	0.64	0.451	0.015
Visual Consistency	—	Purity	0.731	0.599	0.132	0.82	0.887	$< 10^{-6}$
Visual Consistency	—	Fragmentation	0.082	0.103	-0.021	0.70	-0.401	0.032
Motion Dynamics	—	Change	10.577	9.895	0.682	0.64	0.454	0.014
Subtitle Density	—	Level	8.210	8.813	-0.604	0.72	-0.683	2.7×10^{-5}
Text Region Density	Box area	Change	5.144	5.636	-0.492	0.66	-0.462	0.007
Speech Dynamics	Pitch	Level	97.096	90.727	6.369	0.78	0.631	1.5×10^{-4}
Speech Dynamics	Speak. rate	Level	3.540	3.617	-0.077	0.64	-0.396	0.042
Audio Novelty	MFCC-Δ	Level	28.291	29.119	-0.828	0.70	-0.570	8.9×10^{-4}
Laughter and Applause	Laughter	Local contrast	N/A	N/A	0.000	0.74	0.655	1.7×10^{-4}

Table 14: Detailed statistical evidence for Basketball.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	—	Local contrast	N/A	N/A	-0.015	0.70	-0.495	0.004
Face Features	Dispersion	Level	0.183	0.178	0.005	0.64	0.395	0.043
Visual Consistency	—	Purity	0.704	0.610	0.094	0.72	0.603	6.0×10^{-4}
Visual Consistency	—	Fragmentation	0.090	0.109	-0.019	0.66	-0.484	0.006
Motion Dynamics	—	Level	12.617	11.613	1.005	0.72	0.475	0.009
Motion Dynamics	Horiz. u	Level	20.656	18.933	1.722	0.70	0.427	0.024
Shot Transition	—	Rate	0.096	0.087	0.008	0.64	0.332	0.082
Subtitle Density	—	Level	9.249	9.730	-0.481	0.78	-0.619	1.1×10^{-4}
Subtitle Density	—	Change	0.856	0.820	0.037	0.66	0.390	0.016
Audio Novelty	—	Change	0.069	0.066	0.003	0.72	0.445	0.008
Audio Novelty	Spec. flux	Change	0.021	0.020	0.002	0.72	0.504	0.005
Laughter and Applause	Laughter	Local contrast	N/A	N/A	0.000	0.74	0.437	0.022

Baseball Table 13 shows that Baseball highlights are anchored most strongly by scene coherence. Visual consistency is the dominant signal: purity increases from 0.599 in nonhighlight regions to 0.731 in highlight regions, with $\Delta = 0.132$, Cons. $= 0.82$, $r_b = 0.887$, and $q < 10^{-6}$, while fragmentation decreases from 0.103 to 0.082 with $r_b = -0.401$. This means that baseball highlights are usually contained within a single coherent scene, such as a sustained play or replay, rather than being formed through frequent scene boundaries. Visual dynamics provides a secondary signal. Motion change increases from 9.895 to 10.577 with $r_b = 0.454$ and $q = 0.014$, and the face-dispersion change reaches $r_b = 0.451$ with $q = 0.015$, indicating that highlights tend to coincide with more on-field motion and more spatially shifting faces. The main curvature level shows only a borderline tendency, with $\Delta = 0.006$ and $q = 0.055$, so trajectory turning is not a decisive cue here.

The audio and text evidence sharpens this picture. The pitch component is the strongest acoustic signal: pitch level rises from 90.727 to 97.096 with $r_b = 0.631$ and $q = 1.5\times10^{-4}$, reflecting more excited or elevated vocal commentary in highlights. In contrast, the mfcc-delta component of audio novelty decreases from 29.119 to 28.291 with $r_b = -0.570$, and the speaking-rate component of speech dynamics also decreases with $r_b = -0.396$, suggesting that the acoustic timbre becomes more stable and delivery does not necessarily speed up even as pitch climbs. Text features show a consistent reduction: subtitle level decreases from 8.813 to 8.210 with $r_b = -0.683$ and $q = 2.7\times10^{-5}$, and the text-box-area change of text region density decreases with $r_b = -0.462$. Reaction-based audio is weak; only the laughter local contrast is reliable, with Cons. $= 0.74$ and $r_b = 0.655$, while the combined reaction level is not significant. Therefore, baseball highlights are best described as coherent on-field scenes with elevated commentary pitch and lighter text load, rather than text-heavy or laughter-driven segments.

Basketball Table 14 shows that Basketball highlights are driven primarily by motion and visual scene coherence rather than by semantic trajectory turning. Visual consistency is the strongest cue: purity increases from 0.610 in nonhighlight regions to 0.704 in highlight regions, with $\Delta = 0.094$, Cons. $= 0.72$, $r_b = 0.603$, and $q = 6.0\times10^{-4}$, while fragmentation decreases from 0.109 to 0.090 with $r_b = -0.484$. This indicates that basketball highlights sit inside a single coherent scene, such as a sustained play or replay, rather than being assembled from many scene boundaries. Motion dynamics is the second pillar. Optical-flow magnitude increases from 11.613 to 12.617 with $r_b = 0.475$, and the dominant horizontal component rises consistently by 1.722 ($r_b = 0.427$), reflecting the fast, sweeping court action that characterizes highlight moments.

The semantic and audio evidence plays a supporting role. The main curvature signal is not the turning angle but the local contrast, which is significantly negative ($\Delta = -0.015$, Cons. $= 0.70$, $r_b = -0.495$, $q = 0.004$), indicating that highlights tend to be locally smoother in semantic content than their surroundings rather than marked by sharp embedding turns. Face dispersion shows a mild positive effect, with the level

Table 15: Detailed statistical evidence for Football.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.104	0.095	0.009	0.80	0.736	2.0×10^{-6}
Face Features	Dispersion	Change	0.049	0.043	0.006	0.70	0.457	0.013
Visual Consistency	—	Purity	0.630	0.497	0.133	0.74	0.729	1.0×10^{-5}
Visual Consistency	—	Boundary rate	0.060	0.047	0.013	0.68	0.551	8.1×10^{-4}
Motion Dynamics	—	Level	14.768	12.442	2.325	0.76	0.713	9.0×10^{-6}
Shot Transition	—	Rate	0.129	0.094	0.035	0.76	0.791	$< 10^{-6}$
Subtitle Density	—	Level	8.205	9.542	-1.337	0.90	-0.950	$< 10^{-6}$
Text Region Density	—	Level	1.495	1.722	-0.227	0.74	-0.591	5.1×10^{-4}
Speech Dynamics	Pitch	Level	69.933	62.647	7.285	0.82	0.682	2.5×10^{-5}
Speech Dynamics	Speak. rate	Change	0.886	0.949	-0.062	0.70	-0.553	0.001
Audio Novelty	—	Level	0.546	0.580	-0.034	0.90	-0.937	$< 10^{-6}$
Laughter and Applause	—	Level	0.005	0.005	-0.000	0.54	-0.020	0.909

Table 16: Detailed statistical evidence for Review.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.136	0.127	0.009	0.74	0.645	6.6×10^{-5}
Curvature	Δ-emb.	Local contrast	N/A	N/A	0.008	0.72	0.551	4.9×10^{-4}
Visual Consistency	—	Boundary rate	0.057	0.046	0.011	0.74	0.652	1.3×10^{-4}
Visual Consistency	—	Neg. purity	0.263	0.344	-0.081	0.64	-0.496	0.003
Visual Consistency	—	Local contrast	N/A	N/A	0.015	0.84	0.576	6.3×10^{-4}
Motion Dynamics	—	Level	2.725	3.018	-0.294	0.72	-0.416	0.029
Text Region Density	—	Level	2.154	1.977	0.177	0.68	0.511	0.002
Text Region Density	—	Change	0.584	0.474	0.110	0.82	0.732	6.0×10^{-6}
Text Region Density	Box area	Change	1.598	1.358	0.240	0.68	0.478	0.008
Speech Dynamics	Pitch	Level	49.357	46.605	2.752	0.60	0.387	0.050
Laughter and Applause	—	Level	0.002	0.002	0.000	0.54	0.104	0.524

rising from 0.178 to 0.183 ($r_b = 0.395$, $q = 0.043$), consistent with players spreading across the frame. Audio novelty contributes through change rather than level: the novelty change metric increases with $r_b = 0.445$ ($q = 0.008$), and its spectral-flux component is the cleaner signal at $r_b = 0.504$ ($q = 0.005$), capturing the spectral bursts of crowd and commentary. The laughter local contrast is also weakly significant ($r_b = 0.437$, $q = 0.022$). Text behaves bidirectionally: subtitle level drops from 9.730 to 9.249 ($r_b = -0.619$), while subtitle change rises slightly ($r_b = 0.390$, $q = 0.016$). Shot transition rate, which one might expect to spike at scoring cuts, only trends upward ($r_b = 0.332$, $q = 0.082$) and is not reliable. Basketball highlights are therefore best described as coherent, motion-rich plays accompanied by acoustic bursts and concise but faster-changing on-screen text.

Football Table 15 shows that Football highlights are driven by visual dynamics and pacing rather than by audio reactions. Motion dynamics is one of the strongest signals: motion level increases from 12.442 to 14.768, with $\Delta = 2.325$, Cons. $= 0.76$, and $r_b = 0.713$. Shot transition is even stronger, with the cut rate rising from 0.094 to 0.129 and $r_b = 0.791$ at $q < 10^{-6}$. These results indicate that football highlights coincide with fast on-field action and rapid camera cutting. The visual trajectory also moves quickly: the adjacent-frame embedding change increases from 0.095 to 0.104 with $r_b = 0.736$, confirming that highlights are moments of large frame-to-frame semantic change. Face features contribute only a secondary cue, where the spatial spread of face centers changes more sharply ($r_b = 0.457$), consistent with shifting crowds or players in frame.
Visual consistency presents a dual pattern. Purity increases from 0.497 to 0.630 with $r_b = 0.729$, yet boundary rate also increases from 0.047 to 0.060 with $r_b = 0.551$, suggesting that highlights sit inside coherent action segments that are nonetheless bracketed by more frequent scene boundaries. The audio profile is distinctive. Pitch rises from 62.647 to 69.933 with $r_b = 0.682$, reflecting more excited commentary, while the speaking-rate change actually decreases ($r_b = -0.553$). Audio novelty falls sharply from 0.580 to 0.546 with $r_b = -0.937$, indicating that the acoustic background becomes more stable as sustained crowd and commentary noise dominates highlights. Text load drops too: subtitle level decreases from 9.542 to 8.205 with $r_b = -0.950$, and text region density decreases from 1.722 to 1.495. Notably, laughter and applause is not a reliable indicator here, with $r_b = -0.020$ and $q = 0.909$. Therefore, football highlights are best described as fast, cut-heavy, motion-rich action segments accompanied by higher-pitched commentary and a stable acoustic background, rather than text-heavy or laughter-driven moments.

Review Table 16 shows that Review highlights are driven by semantic change, scene transitions, and on-screen text rather than by motion or audience reactions. The adjacent-frame embedding change is the strongest curvature signal: its level rises from 0.127 to 0.136 with $\Delta = 0.009$, Cons. $= 0.74$, $r_b = 0.645$,

Table 17: Detailed statistical evidence for Vlog.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	Δ-emb.	Level	0.129	0.122	0.007	0.74	0.638	8.4×10^{-5}
Face Features	—	Change	0.402	0.337	0.066	0.74	0.421	0.027
Visual Consistency	—	Boundary rate	0.060	0.046	0.013	0.84	0.823	$< 10^{-6}$
Visual Consistency	—	Neg. purity	0.217	0.296	-0.079	0.66	-0.512	0.002
Motion Dynamics	Vert. v	Change	5.238	4.761	0.477	0.70	0.432	0.021
Subtitle Density	—	Change	0.851	0.788	0.063	0.66	0.517	0.003
Text Region Density	—	Change	0.428	0.376	0.053	0.72	0.655	7.2×10^{-5}
Speech Dynamics	Pitch	Level	55.664	51.670	3.994	0.66	0.464	0.011
Audio Novelty	MFCC-Δ	Change	5.054	5.261	-0.207	0.70	-0.407	0.035
Laughter and Applause	Laughter	Change	0.001	0.001	0.000	0.66	0.525	0.002

and $q = 6.6\times10^{-5}$, while the corresponding local contrast confirms the same pattern with $r_b = 0.551$. This indicates that review highlights tend to occur where the semantic content shifts quickly, such as when a new product, scene, or topic is introduced. Visual consistency reinforces this interpretation through boundaries rather than coherence. Boundary rate increases from 0.046 to 0.057 with $r_b = 0.652$, and its local contrast is even stronger, with Cons. $= 0.84$ and $r_b = 0.576$. Consistently, negative purity decreases from 0.344 to 0.263 with $r_b = -0.496$, meaning highlights sit near scene transitions rather than deep inside one homogeneous segment.

The text channel provides the clearest editorial cue. Text region density change increases from 0.474 to 0.584 with Cons. $= 0.82$, $r_b = 0.732$, and $q = 6.0\times10^{-6}$, while the box count level rises from 1.977 to 2.154 with $r_b = 0.511$ and the box-area change grows from 1.358 to 1.598 with $r_b = 0.478$. This shows that highlights coincide with more, and more rapidly changing, on-screen text overlays, which is typical of caption-heavy review editing. By contrast, motion dynamics moves in the opposite direction: optical-flow level decreases from 3.018 to 2.725 with $r_b = -0.416$, so highlights are calmer, more static moments rather than action-driven ones. Speech dynamics contributes only weakly through the pitch component, which rises from 46.605 to 49.357 but reaches only borderline significance ($r_b = 0.387$, $q = 0.050$). Audience reactions are essentially absent: laughter and applause level shows no difference between regions ($\Delta = 0.000$, $r_b = 0.104$, $q = 0.524$). Therefore, review highlights are best described as text-rich, semantically shifting moments near scene transitions, with reduced motion and no reliance on laughter or applause.

Vlog Table 17 shows that Vlog highlights are driven by visual transition and change related cues rather than by static scene coherence. The strongest signal is visual consistency, where the boundary rate increases from 0.046 to 0.060, with Cons. $= 0.84$, $r_b = 0.823$, and $q < 10^{-6}$. Unlike cooking or job experience, this is a boundary based rather than a purity based pattern, indicating that vlog highlights cluster near scene or segment transitions instead of sitting inside one coherent scene. Consistent with this, negative purity decreases from 0.296 to 0.217 with $r_b = -0.512$, confirming that highlight regions are less likely to belong to incoherent surroundings. The change oriented reading is reinforced by the visual dynamics: the adjacent-frame embedding change (curvature Δ-emb. level) rises from 0.122 to 0.129 with $r_b = 0.638$, and face count change increases from 0.337 to 0.402 with $r_b = 0.421$. This suggests that vlog highlights occur where the semantic content shifts quickly and where the number of visible faces changes, such as when a new subject, place, or person enters the frame.

The remaining evidence describes a moderate but consistent multimodal pattern. Vertical optical-flow change increases from 4.761 to 5.238 with $r_b = 0.432$, indicating that highlights coincide with more vertical motion, while the main motion magnitude itself is not reliable. On screen text behaves dynamically rather than densely: text box count change increases from 0.376 to 0.428 with $r_b = 0.655$, and subtitle change increases from 0.788 to 0.851 with $r_b = 0.517$, so highlights feature faster turnover of text rather than simply more text. On the audio side, the vocal pitch level rises from 51.670 to 55.664 with $r_b = 0.464$, and the combined laughter and applause cue is weakly positive, with laughter change reaching $r_b = 0.525$, pointing to mild emotional or reactive emphasis. In contrast, audio novelty does not support highlights: its mfcc-delta change actually decreases from 5.261 to 5.054 with $r_b = -0.407$, meaning the acoustic timbre becomes more stable rather than more novel. Therefore, vlog highlights are best described as transition rich moments with rapid semantic and face changes, more vertical motion, fast changing text, and higher pitch, rather than as acoustically novel or text heavy segments.

Knowledge Table 18 shows that Knowledge highlights are anchored most strongly in visual coherence and acoustic novelty. The dominant signal is visual consistency, where purity rises from 0.482 in nonhighlight regions to 0.663 in highlight regions, with $\Delta = 0.181$, Cons. $= 0.86$, $r_b = 0.862$, and $q < 10^{-6}$. This indicates that knowledge highlights typically sit inside a single coherent explanatory scene rather than

Table 18: Detailed statistical evidence for Knowledge.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Visual Consistency	—	Purity	0.663	0.482	0.181	0.86	0.862	$< 10^{-6}$
Visual Consistency	—	Neg. purity	0.306	0.220	0.086	0.68	0.388	0.040
Motion Dynamics	Vert. v	Level	0.758	1.090	-0.332	0.70	-0.380	0.056
Shot Transition	—	Rate	0.129	0.119	0.010	0.70	0.454	0.009
Subtitle Density	—	Level	13.142	12.874	0.268	0.66	0.440	0.014
Subtitle Density	—	Change	0.299	0.318	-0.019	0.72	-0.418	0.014
Text Region Density	—	Change	0.374	0.330	0.044	0.54	0.347	0.048
Text Region Density	—	Local contrast	N/A	N/A	0.270	0.64	0.404	0.037
Audio Novelty	—	Level	0.705	0.697	0.008	0.72	0.525	0.003
Audio Novelty	Spec. flux	Level	0.393	0.391	0.002	0.72	0.492	0.006
Laughter and Applause	—	Level	0.001	0.001	0.000	0.64	0.230	0.472

Table 19: Detailed statistical evidence for Lecture.

Feature	Component	Metric	H	N	Δ	Cons.	r_b	q
Curvature	—	Level	1.894	1.904	-0.010	0.64	-0.413	0.021
Curvature	Δ-emb.	Level	0.090	0.086	0.004	0.64	0.481	0.005
Visual Consistency	—	Purity	0.661	0.475	0.186	0.82	0.776	$< 10^{-6}$
Visual Consistency	—	Neg. purity	0.386	0.167	0.219	0.72	0.707	1.0×10^{-5}
Visual Consistency	—	Boundary rate	0.057	0.048	0.009	0.74	0.468	0.006
Subtitle Density	—	Change	0.622	0.574	0.048	0.68	0.432	0.011
Subtitle Density	—	Local contrast	N/A	N/A	-0.339	0.68	-0.550	0.002
Speech Dynamics	—	Level	0.314	0.320	-0.006	0.72	-0.520	0.003
Speech Dynamics	Speak. rate	Level	3.232	3.330	-0.097	0.78	-0.600	3.9×10^{-4}
Audio Novelty	Spec. flux	Level	0.392	0.390	0.002	0.72	0.410	0.016
Audio Novelty	MFCC-Δ	Local contrast	N/A	N/A	-1.353	0.76	-0.598	4.1×10^{-4}
Laughter and Applause	—	Local contrast	N/A	N/A	0.007	0.68	0.467	0.011

being assembled from scattered scene boundaries. Negative purity also increases from 0.220 to 0.306 with $r_b = 0.388$ and $q = 0.040$, suggesting that the surrounding context is comparatively less coherent. Shot transition reinforces this picture from the opposite direction: the shot-cut rate increases from 0.119 to 0.129 with Cons. $= 0.70$ and $r_b = 0.454$, so highlight moments still carry slightly more frequent cuts even while staying within a coherent scene. Vertical motion, by contrast, is suppressed, with the vertical flow component dropping from 1.090 to 0.758 ($r_b = -0.380$), although its corrected significance is only borderline at $q = 0.056$.

The audio and text channels provide the complementary evidence. Audio novelty increases from 0.697 to 0.705 with $r_b = 0.525$ and $q = 0.003$, and its spectral-flux component shows the same pattern with $r_b = 0.492$ and $q = 0.006$, indicating that highlights coincide with locally distinct acoustic content. Text behaves bidirectionally and informatively: subtitle level increases from 12.874 to 13.142 ($r_b = 0.440$), while subtitle change decreases from 0.318 to 0.299 with $r_b = -0.418$ and $q = 0.014$, so highlights carry denser but more steadily delivered narration. On-screen text also intensifies, with text-region change rising from 0.330 to 0.374 ($r_b = 0.347$) and its local contrast reaching $r_b = 0.404$ with $q = 0.037$. In contrast, laughter and applause are not reliable here, with $r_b = 0.230$ and $q = 0.472$. Therefore, knowledge highlights are best described as coherent, explanation-driven segments marked by acoustically novel narration and denser on-screen text, rather than by reaction-based cues.

Lecture Table 19 shows that Lecture highlights are dominated by scene coherence and a distinctive reduction in speech effort. Visual consistency is the strongest signal: purity rises from 0.475 in nonhighlight regions to 0.661 in highlight regions, with $\Delta = 0.186$, Cons. $= 0.82$, $r_b = 0.776$, and $q < 10^{-6}$, and negative purity rises from 0.167 to 0.386 with $r_b = 0.707$. Boundary rate also increases from 0.048 to 0.057 with $r_b = 0.468$. Together these indicate that lecture highlights tend to sit inside a coherent presentation scene while still being marked by slightly more frequent segment boundaries. The semantic trajectory pattern is mixed: the main curvature turning angle decreases ($\Delta = -0.010$, $r_b = -0.413$, $q = 0.021$), so highlights follow a smoother semantic path, whereas the adjacent-frame embedding change increases from 0.086 to 0.090 ($r_b = 0.481$, $q = 0.005$), showing that local frame-to-frame semantic motion still rises.

The audio evidence is led by reduced speech rather than added reactions. Speech dynamics level decreases from 0.320 to 0.314 ($r_b = -0.520$, $q = 0.003$), and the speaking-rate component drops from 3.330 to 3.232 with Cons. $= 0.78$ and $r_b = -0.600$, indicating that highlights coincide with slower, more deliberate speech. Audio novelty is selective: its spectral-flux component is higher in highlights ($\Delta = 0.002$, Cons. $= 0.72$, $r_b = 0.410$), while the mfcc-delta local contrast is strongly negative ($\Delta = -1.353$, $r_b = -0.598$, $q = 4.1\times10^{-4}$), meaning the spectral surface is more active but the timbral background steadies. Text behaviour mirrors the speech slowdown: subtitle change increases from 0.574 to 0.622 ($r_b = 0.432$) yet subtitle local contrast is negative ($\Delta = -0.339$, $r_b = -0.550$, $q = 0.002$), so highlights carry locally

lighter but more rapidly turning captions. Laughter and applause register only weakly, appearing as a mild local-contrast effect ($\Delta = 0.007$, Cons. $= 0.68$, $r_b = 0.467$, $q = 0.011$) rather than a sustained presence, confirming that audience reactions are not a defining cue here. Lecture highlights are therefore best described as coherent presentation scenes with slowed, deliberate speech, concise but rapidly changing text, and a steadier acoustic background.

J Statistics Visualization

The statistical results show that highlight regions are not characterized by uniform feature activation. Reliable effects appear on both sides of the rank-biserial axis: some features become stronger inside highlights, while others are consistently suppressed. Figure 11 shows this asymmetry at the aggregate level. Strong positive effects correspond to amplified cues such as coherent visual structure, motion, cuts, or reactions in the categories where they apply, whereas strong negative effects correspond to reduced cues such as subtitle load or audio novelty. The concentration of many points near zero indicates that many feature–subcategory pairs do not provide stable highlight evidence.

Figure 12 shows that reliable effects are strongly structured by subcategory. Visual consistency is the most broadly recurring positive signal, suggesting that many highlights remain tied to coherent scene structure, while subtitle density often shows a reliable negative effect, indicating reduced textual load in highlight regions. Other channels, including reaction cues, audio novelty, speech dynamics, and on-screen text regions, become informative only for particular subcategories. These patterns support the main analysis claim that highlight cues are category-dependent feature combinations rather than isolated universal signals.

K LLM Prompt Instruction

We organize the category-specific analysis results into prompt-ready guide packs constructed from our feature-level analysis. The user provides three inputs: a target video subcategory, a whole-video description, and a desired editing direction. The video description summarizes the overall content, participants, major scenes, actions, and atmosphere of the target video, while the editing direction describes the intended pacing, tone, emphasis, and viewing experience.

For our method, the system retrieves the prompt-ready analysis pack corresponding to the input subcategory and inserts it into the prompt as additional analysis context. Therefore, the three fields provided by the user are kept separate from the retrieved context. The baseline condition uses only the user input and the common prompt instructions, whereas our condition additionally includes the retrieved category analysis. Both conditions use the same output format and common prompt components.

The retrieved analysis context contains prompt-ready qualitative evidence derived from modality attention patterns and feature-level statistical analysis. For each subcategory, it summarizes prioritized feature blocks, cross-feature editing implications, unsupported interpretations that should be avoided, and a policy for using modality attention as supplementary evidence.

Because raw statistical values are difficult to apply directly during editing, the framework primarily converts the analyzed tendencies into qualitative and relative editing actions. Examples include maintaining a coherent scene, increasing or reducing subtitle load, emphasizing reaction cues, controlling visual pacing, or stabilizing the acoustic background. The prompt instructs the LLM not to invent unsupported feature values or scene details. Numerical values may be used only when they are explicitly included in the provided context; otherwise, the guideline must use qualitative conditions.

The primary evidence comes from four feature-analysis groups: visual dynamics, semantic change, reaction cues, and text-based features. Modality attention is used only as supplementary context. If modality-attention tendencies conflict with the feature-level analysis, the feature-level analysis takes priority. For guideline generation, we use LGAI-EXAONE/EXAONE-Deep-7.8B [3].

L Example of Generated Editing Guideline

Figure 13 presents an example output of the editing guideline generated by the LLM.

Prompt C

[Task]
Generate structured video editing guidelines that describe how highlight regions should be edited differently from non-highlight regions.

[Input]
1. [main category]: Broad content category provided by the user.
2. [sub category]: Fine-grained content subcategory provided by the user.
3. **[main category analysis]**: Retrieved analysis for the main category, including broad modality tendencies and feature-level patterns.
4. **[sub category analysis]**: Retrieved analysis for the subcategory, including modality attention, feature tendencies, and highlight-specific patterns.
5. [video description & editing direction]: User-provided description of the video content, scene structure, editing goal, target audience, and intended viewing experience.

[Objective]
1. Use [sub category analysis] and [video description] as the primary basis for the guideline.
2. Use [main category analysis] only as supplementary context.
3. Generate relative and qualitative highlight guidelines rather than exact feature values or numerical thresholds.
4. Explain how highlight regions should differ from non-highlight regions in terms of visual, text, and audio editing.
5. Convert analyzed feature tendencies into concrete editing actions that can be directly applied on the editing timeline.

[Evidence Rules]
1. Prioritize [sub category analysis] over [main category analysis].
2. Adapt every guideline to [video description].
3. Do not output exact numerical thresholds, raw feature values, or unit-based rules unless they are explicitly requested.
4. Do not invent unsupported feature tendencies.
5. If main category and subcategory patterns differ, follow the subcategory pattern.
6. Use modality attention only as auxiliary evidence when feature-level analysis provides the main explanation.

[Guideline Style] Use relative and qualitative expressions such as more frequent, more concentrated, more dynamic, clearer, stronger, more selective, more stable, less distracting, or more emphasized.
For example, write ``highlight regions should use more frequent cut transitions than non-highlight regions'' instead of giving a fixed transition value.

[Output Format]
[Category Information]
Main Category: Write the main category.
Sub Category: Write the subcategory.

[Video Context]
Summarize the user's video description in one sentence.

[Sub Category Focus]
Summarize the main highlight tendency of the subcategory.

[Dominant Modality]
State the dominant modality or modality combination.
Mention the main category tendency only when it helps interpret the subcategory pattern.

[Relative Highlight Direction]
Describe how highlight regions should generally stand out from non-highlight regions in this video.

[Visual Guide]
Provide relative visual guidance using relevant patterns such as shot transition, motion dynamics, visual consistency, face composition, scene change, and semantic change.
Include the relative target, editing action, and reason.

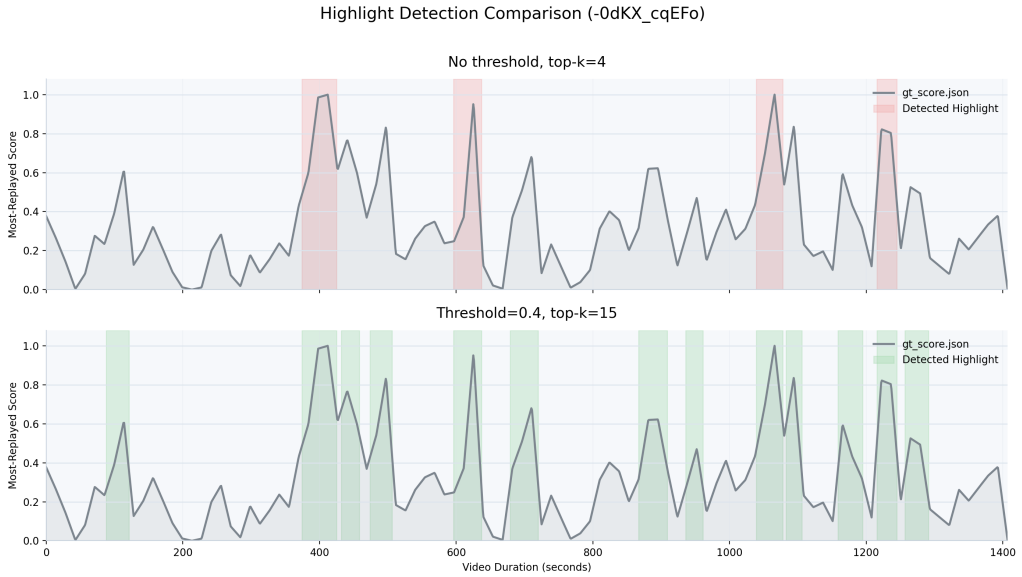
[Text Guide]
Provide relative text guidance using relevant patterns such as subtitle density, text region density, emphasized on-screen text, keyword captions, and subtitle replacement speed.
Include the relative target, editing action, and reason.

[Audio Guide]
Provide relative audio guidance using relevant patterns such as speech dynamics, pitch variation, audio novelty, laughter, applause, background sound control, and reaction bursts.
Include the relative target, editing action, and reason.

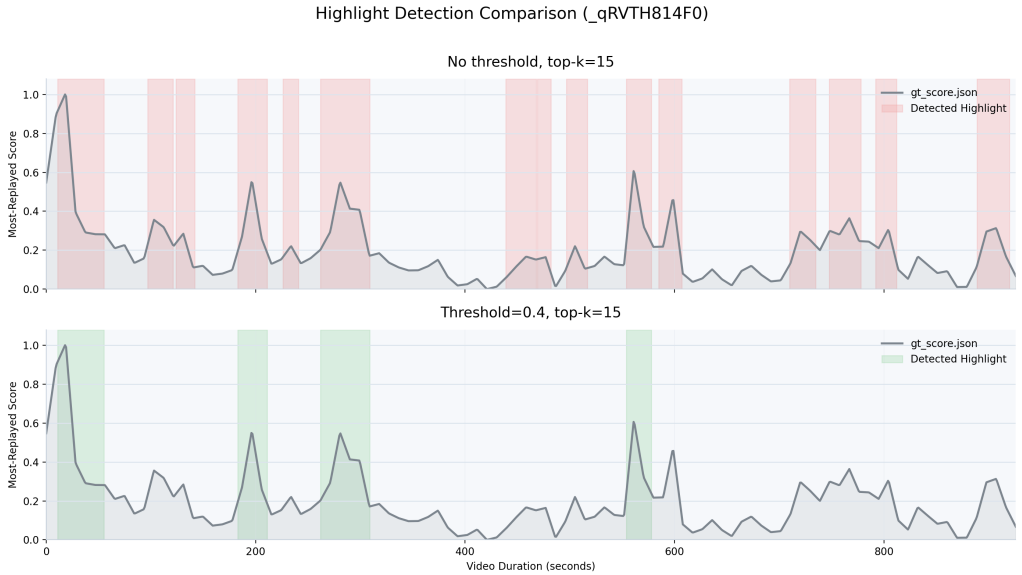
[Priority Guide]
Rank the top three relative editing actions that should be applied first.

[Summary]
Provide a concise summary of the recommended highlight editing style.

[Constraints]
Be concise, structured, and practical.
Focus on relative differences between highlight and non-highlight regions.
Prioritize subcategory analysis and video description.
Avoid exact numerical thresholds and generic recommendations.
Do not simply restate the retrieved analysis; transform it into concrete editing guidance.
Do not output reasoning, analysis steps, or explanations before the final guide.



(a) Effect of increasing k. Compared with a small k, the proposed setting captures consecutive highlight regions more reliably, preventing adjacent important modes from being missed.



(b) Effect of minimum peak filtering. Even when k is set to a large value, low score modes are removed if their maximum values do not exceed the minimum peak threshold, leading to more accurate highlight detection.

Figure 10: Examples of the proposed watershed grouping strategy. The method first generates sufficient candidate modes using a large k, and then filters them based on their peak Most-Replayed values. This allows the final number of detected highlight regions to vary across videos while avoiding unreliable low score detections.

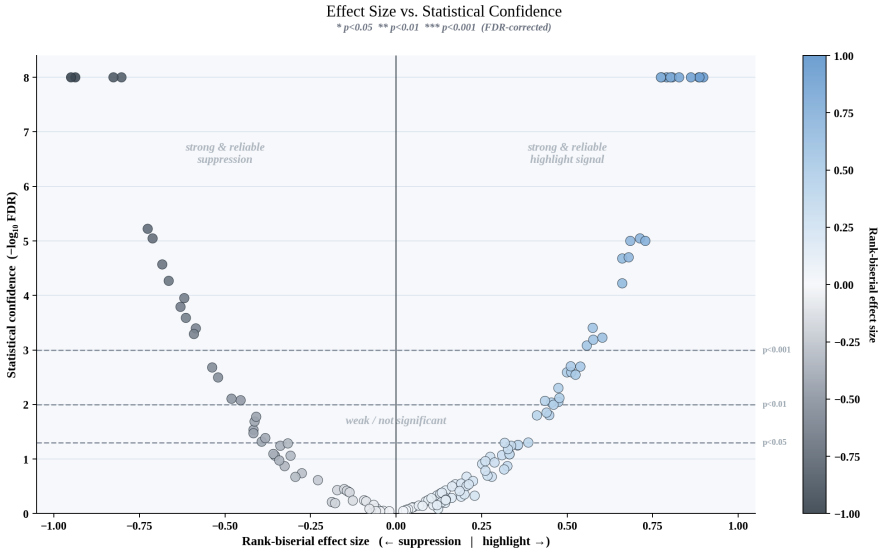
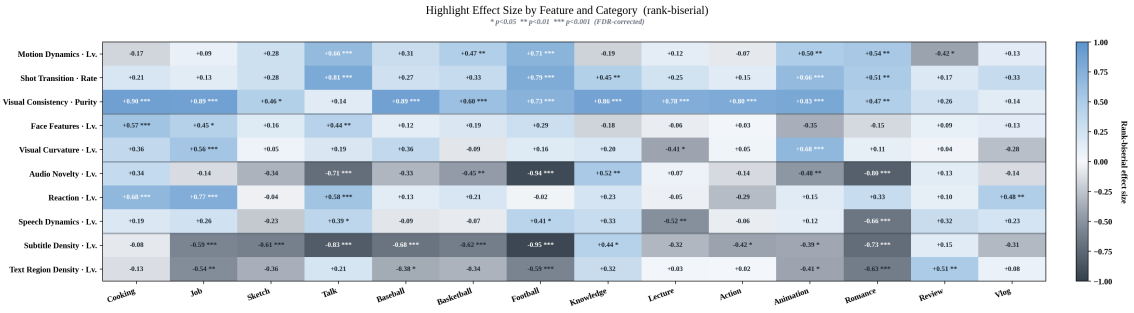
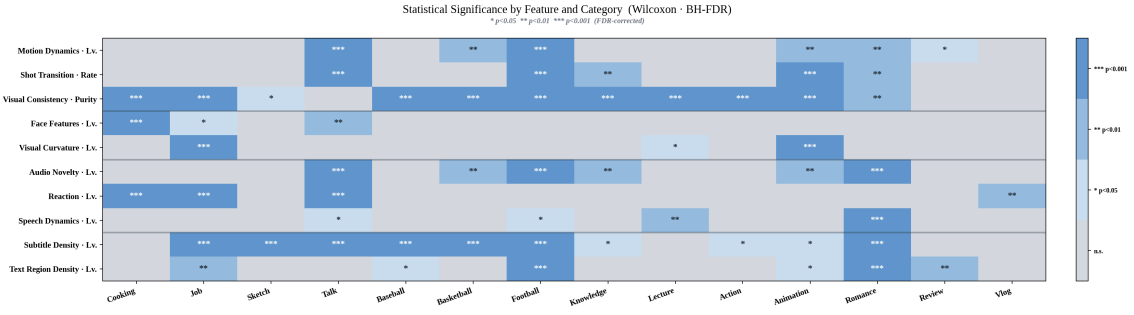


Figure 11: **Effect size versus statistical confidence for every feature–subcategory pair.** The horizontal axis is the rank-biserial effect size of the feature's main metric (right: elevated in highlight regions, left: suppressed), and the vertical axis is the statistical confidence, $-\log_{10}$ of the BH–FDR-corrected Wilcoxon q-value. Marker color repeats the effect size. Strong and reliable cues lie in the upper corners, while the dense cluster near the origin marks pairs with no dependable highlight effect.



(a)



(b)

Figure 12: **Highlight effect of each feature across the subcategories.** (a) Effect size: each cell is the rank-biserial effect size of a feature–subcategory pair (blue: elevated in highlight regions, gray: suppressed), annotated with its value and BH–FDR significance ($^{*}q<0.05$, $^{**}q<0.01$, $^{***}q<0.001$). (b) Significance: the same grid colored only by the BH–FDR significance level (n.s., $*$, $**$, $***$). Across subcategories, visual consistency is the most consistently positive cue and subtitle density is reliably negative, whereas most other features reach significance only in specific subcategories, confirming that highlight cues are largely category-dependent rather than universal.

Input:

[Category]: Lifestyle
[Sub-category]: vlog
[Video Description]: This is a daily vlog video. There are two people in the video: one main character who is filming the vlog and one friend who joins in the middle. At the beginning, the main character sits alone in a cafe, drinks a beverage, and briefly talks about the day's schedule and their mood. Then, the scene moves to the street, showing surrounding scenery, signs, and walking shots. After meeting the friend, the video includes scenes of them talking and laughing together. The topic is not a special event, but an ordinary day and the small emotional changes within it. The main actions include moving around, ordering, drinking a beverage, talking, and looking around. The mood is comfortable and bright, combining quiet moments alone with slightly more lively moments shared with the friend. Scenes with different characteristics, such as the cafe scene, street walking scene, and conversation with the friend, are naturally connected.
[Editing Direction]: I want to edit this into a natural and bright highlight. I want the scene transitions to flow smoothly, and I want to add short and light on-screen text to create a stylish rhythm.

EXAONE-Deep:

[Category Information]
Main Category: Lifestyle
Sub Category: vlog

[Video Context]
This is a daily vlog video featuring the main character and a friend. The video begins with the main character drinking a beverage at a cafe while briefly talking about the day's schedule and emotions, then moves to the street to show scenery, signs, and activity shots, and later includes moments of conversation and laughter with the friend. The theme centers on an ordinary day and subtle emotional changes, with everyday actions such as moving around, ordering, drinking, talking, and looking around. The mood is comfortable and bright, and quiet solo moments naturally connect with more active moments shared with the friend. Different scenes, such as the cafe scene, street walking scene, and conversation with the friend, are smoothly connected.

[Dominant Modality]
Visual + Audio. Visual scene transitions and changes in the characters' emotional states form the main flow of the highlight, while in audio, laughter and conversation should be naturally integrated into the scene transitions. In particular, the natural conversation and laughter that occur during the scenes with the friend are highly important.

[Visual Guide]
*Condition: Prioritize compositions where the main character's solo scene before the friend joins and the conversation scene with the friend are smoothly connected within the same sequence.
* Editing Action: To create a natural transition between the end of the cafe scene and the beginning of the friend's appearance, start with a wider shot that includes both the main character and the friend, then gradually move into the conversation scene using cuts where the distance between the characters becomes closer. In the street scenes, when capturing signs and people moving in the surroundings, track the motion within a stable frame to catch moments of change. For the laughter moment with the friend, capture the action clearly and connect it to a following shot that continues to show the emotional shift.
* Reason: The analysis indicates that visual transition alignment is the strongest signal, and changes in character composition during the conversation with the friend serve as a key element of semantic change.

[Text Guide]
* Condition: On-screen text should be used mainly when showing order numbers or location information, and the content should remain concise.
* Editing Action: In the cafe scene, briefly apply a text box when the order number appears on screen. In the street walking scene, add text only when sign or location information is necessary, and keep the text small. The text should not dominate the scene, but should only help guide the temporal flow of the video.
* Reason: The text-based analysis shows that text box replacement is a frequently used signal, especially in scenes related to information transitions.

[Audio Guide]
* Condition: Prioritize sections where natural laughter and conversation occur between the main character and the friend.
* Editing Action: In the conversation scene, preserve the audio when laughter occurs, and avoid suppressing the surrounding sound too much so the moment feels natural. In scenes without dialogue, minimize background sound to emphasize the quiet cafe atmosphere or the main character's solitary mood. Especially in the first meeting scene with the friend, keep the silence before and after the conversation begins to enhance the natural feeling.
* Reason: The reaction cue analysis indicates that laughter is a strong signal, and audio features play an important role in scene transitions.

[Priority Guide]
1. First, select scenes that smoothly connect the main character's quiet solo moment in the cafe with the first conversation scene with the friend.
2. In the street walking scene, track surrounding signs and people's movements in a wide and natural way so the camera transitions feel smooth.
3. Capture the laughter moment during the conversation with the friend and connect it to a following shot that continues the same emotional flow.

[Summary]
The vlog highlight should focus on naturally connecting the main character's quiet solo moments with more active interactions with the friend. By balancing visual transitions and audio reactions, the edit can maintain a comfortable and bright mood while vividly conveying everyday actions and emotional changes. In particular, the natural laughter and conversation in the scenes with the friend should be treated as key highlight moments.

Figure 13: Example output of the LLM generated editing guideline.