

TVHighlights: LLM-Guided Human-Free Collaborative Training for Video Highlight Detection in Movies and TV Dramas

Qi Qiu^{1†} Xuan Wu^{1†} Jiawei Peng^{1,2†} Yuan Miao^{1*} Xu Yang² Yanlong Du¹

¹ HUIJING Digital Media & Entertainment Group

² School of Computer Science & Engineering, Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China

{qimiao.qq, wx193834, pengjiawei.pjw, miaoyuan.my}@alibaba-inc.com

Abstract

Video highlight detection aims to identify the most engaging segments in long-form videos, supporting content editing and recommendation, especially for movies and TV dramas. However, existing methods are ill-suited to cinematic content due to its narrative complexity, while the scarcity of annotated data and the high cost of manual labeling further hinder progress. To bridge this gap, we introduce **TVHighlights**, the first large-scale dataset tailored for video highlight detection in movies and TV dramas, with 1,721 carefully curated videos covering diverse genres. Built on community-driven behaviors, it provides realistic and diverse annotations without human labeling. Based on TVHighlights, we propose **LTV-HD**: a LLM-guided, human-free collaborative training framework for video highlight detection in cinematic content. LTV-HD operates in two stages: (1) weakly supervised pre-training of a lightweight model using video-level labels, followed by (2) iterative refinement through collaboration between large language models (LLMs) and the lightweight model. LLMs generate noisy clip-level pseudo-labels, which the lightweight model learns from under a noise-robust strategy, and its high-confidence predictions are then fed back to guide the LLM in distilling genre-specific highlight patterns through a self-improving loop. Experiments demonstrate that LTV-HD achieves SOTA performance on TVHighlights, validating its effectiveness in real-world scenarios.

1. Introduction

The rise of short-form video platforms has shifted audience preferences toward brief, emotionally charged clips, such as dramatic kisses, fight scenes, or plot twists, over full-length cinematic content. This trend drives growing demand for



Figure 1. Highlighted clips for three TV drama types and their expected detection results. Highlights in movies and TV dramas lack a clear-cut definition, varying across categories.

automatic **video highlight detection**: the task of identifying the most engaging segments in long videos to support content editing, recommendation, and advertising.

However, video highlight detection remains challenging for *movies and TV dramas*. Unlike sports or vlogs, where highlights follow structured patterns (e.g., goals, stunts), cinematic highlights are inherently subjective and diverse (see Figure 1). This lack of a universal definition makes it difficult to design one-size-fits-all models, and even harder to collect consistent human annotations at scale.

Unfortunately, existing datasets and methods fail to address this complexity. Most datasets focus on user-generated or domain-specific videos [41, 43], which differ significantly from narrative-driven cinematic content. More critically, they rely on manual annotation of clip-level highlights, which is not only time-consuming and costly, but also prone to annotator bias due to the subjective nature of the task. As a result, there is a critical gap: *no large-scale, diverse, and reliable dataset exists for video highlight detection in real-world movies and TV dramas*.

To address this, we introduce **TVHighlights**, a cinematic highlight detection dataset derived from user video reuse patterns. By leveraging video fingerprinting, we trace frequently reused clips in short-form content back to original movies and TV dramas, using their popularity as implicit

*Corresponding author.

†These authors contributed equally to this work.

annotations. The resulting dataset contains 1,721 videos across 15 genres, offering a scalable, human-free benchmark for video highlight detection.

Building upon this dataset, we tackle the next challenge: *How to train an effective video highlight detector without manual supervision?* Existing methods are unsuitable for cinematic content, as they either target simple videos [6, 41, 43], rely on fixed queries [21, 27, 34], or require manual labels [11]. While large language models (LLMs) and multimodal LLMs (MLLMs) enable automated labeling, their direct use introduces significant noise due to hallucinations and inconsistent reasoning [17]. Thus, we need a framework that harnesses LLMs reasoning while reducing label noise.

To overcome this, we propose **LTV-HD**: a LLM-guided, human-free collaborative training framework that transforms the noisy outputs of LLMs into reliable supervision through a self-improving learning loop. As shown in Figure 3, our approach proceeds in **two stages**. First, we initialize a lightweight multimodal model via weakly supervised pre-training, using video-level labels within a multiple instance learning framework. This pretrained model then enters the second stage, where it engages in an iterative cycle with the LLM. The LLM generates initial fine-grained pseudo-labels, which the lightweight model learns from under a noise-robust training scheme. High-confidence predictions from the model are then fed back to guide a LLM in inducing genre-specific highlight patterns (e.g., “intense fight scenes” in wuxia dramas). These distilled semantic patterns are used to refine the labeling process, producing cleaner pseudo-labels for subsequent training rounds. By closing this loop, LTV-HD progressively improves quality of annotations and the representational power of the model, effectively converting noisy, unstructured outputs into structured, actionable knowledge.

In summary, our main contributions are as follows:

- We present **TVHighlights**, the first large-scale, community-annotated dataset for cinematic highlight detection, addressing the lack of realistic benchmarks.
- We propose **LTV-HD**, a novel LLM-guided human-free training framework, enabling video highlight detection for movies and TV dramas without manual annotations, which demonstrates new state-of-the-art performance.
- Within this framework, we propose a noise-robust learning strategy and a highlight pattern induction mechanism that turns LLMs into pattern discoverers, enabling iterative label refinement in a feedback loop.

2. Related Work

2.1. Video Highlight Detection

Video highlight detection (VHD) aims to score and rank clips within a video, assigning higher scores to high-

lights [2, 3, 13, 37]. While video highlight detection has advanced from sports [33, 35, 42] to social videos [32, 44], the cinematic domain remains understudied due to its complexity and subjectivity. Existing methods rely heavily on trailers as weak labels [11]. Recent progress follows several key trends:

Multimodal Fusion for VHD. Joint-VA [2] pioneered the joint modeling of visual and audio features to enhance semantic understanding. UMT [24] further innovated by incorporating a text query generation mechanism to better facilitate joint tasks. Expanding on this, UniVTG [23] introduced a unified video-language temporal grounding framework, pre-trained using pseudo-labels generated by CLIP.

Weakly-supervised VHD. Some VHD methods use video-level or implicit priors as supervision signals. LIM-s [41] suggests shorter user-generated videos are more likely to contain highlights. MINI-Net [15] approaches video highlight detection as a multi-instance learning (MIL) problem with video-level labels. Qinghao Ye et al. [46] achieved weakly-supervised video highlight detection by leveraging multimodal features.

MLLM-based VHD. Recent work has started to incorporate MLLMs into VHD tasks. For instance, TimeChat [29] fine-tunes an MLLM with LoRA for video highlight detection via a Q&A format. However, the zero-shot performance of these specialized models still lags significantly behind general-purpose, large-scale MLLMs like GPT-4.

2.2. Learning with Noisy Labels

Learning with noisy labels (LNL) is a persistent issue in computer vision, and some VHD methods also frame their task to address noisy labels. The main strategies for LNL include: designing robust loss functions [7, 26, 36] to minimize noise impact; employing sample selection [14, 22, 39] to choose less noisy samples for training; and refurbishing noisy labels [1, 5, 16, 47] to correct errors and reduce their negative effects.

3. TVHighlights Dataset

As shown in Table 1, video highlight detection in movies and TV dramas lacks a suitable benchmark, as existing datasets suffer from critical limitations: domain mismatch (e.g., web videos instead of films) [30, 31], incompatible task designs (e.g., query-guided setups) [11, 21], or restricted accessibility. To bridge this gap, we introduce **TVHighlights**, a large-scale dataset for video highlight detection in movies and TV dramas, covering over 15 genres including wuxia, science fiction and romance.

Dataset Construction. As shown in Figure 2, TVHighlights is built upon a community-driven pipeline that leverages real-world user editing behaviors to identify engaging cinematic content. First, we collect approximately 5,000

Table 1. Existing video highlight detection datasets. H: Highlight detection, S: Video summarization, M: Moment retrieval.

Dataset	Task	Domain	Categories	Query-Guided	Human-Annotated	Publicly-Available
Youtube Highlights[31]	H, S	YouTube	Activity	No	Yes	Yes
TVsum[30]	H, S	YouTube	Vlog / News	No	Yes	Yes
QVhighlights[21]	H, M	YouTube	Vlog / News	Yes	Yes	Yes
Movie Highlight[11]	H	Movies	Open	Yes	Yes	No
TVHighlights(ours)	H	Movies + TV Dramas	Open	No	No	Forthcoming

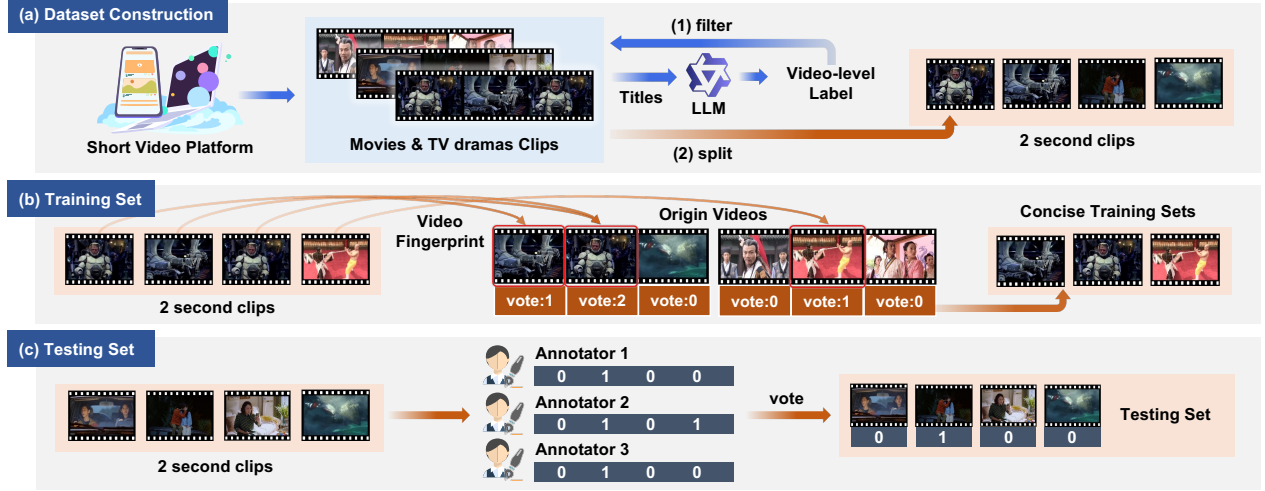


Figure 2. The process of filtering and annotating the TVHighlights dataset. We refine the training set using video fingerprinting technology and collect clip-level highlight labels through multiple human annotations for the test set.

high-engagement videos from short video platforms, all derived from movies and TV dramas. Then we use a LLM to label clips by title and filter out irrelevant ones. Valid clips are split into 2-second non-overlapping clips for fine-grained processing.

Training Set Annotation. We leverage the frequent reuse of video clips across user-generated content as an implicit indicator of highlight value, treating repeated adoption as a form of community voting. Using video fingerprinting, we trace such popular clips back to their original source videos and aggregate their occurrence frequencies to compute a voting score for each clip. Clips with higher scores are considered widely recognized as highlights. The final training set consists of 1,368 videos formed from these high-scoring, crowd-validated clips.

Test Set Annotation. To establish a reliable ground truth for evaluation, we construct a held-out subset as the test set. Then, multiple human annotators label highlight moments at 2-second intervals, and the final ground truth is generated by majority voting to ensure reliability. To enable fine-grained analysis, we further classify the test set into five semantic categories, such as action and emotional climaxes. We construct 353 test data, and we will release the dataset and its annotations to support future research. Detailed statistics are provided in the Appendix A.

4. Our Approach

We propose LTV-HD, a LLM-guided human-free collaborative training framework for video highlight detection in movies and TV dramas, as illustrated in Figure 3. LTV-HD trains a lightweight **Local Temporal Relation Multimodal Network (LTRM)** in two stages: (1) weakly-supervised pre-training, and (2) iterative refinement through collaboration between the LTRM and LLM.

Input videos are split into T non-overlapping 2-second clips, denoted as $\{c_t\}_{t=1}^T$. Visual and audio features $\{v_t\}_{t=1}^T$ and $\{a_t\}_{t=1}^T$ are extracted via pre-trained encoders. The LTRM processes these multimodal inputs and outputs clip-level highlight scores $\{s_t\}_{t=1}^T \in [0, 1]$, indicating the likelihood of each clip being a highlight.

In the first stage, we employ **Multiple Instance Learning** to train the LTRM with video-level labels, forming a foundation for refinement. In the second stage, we establish an iterative collaborative training loop: noisy labels are generated via **LLM-guided Label Generation**, and the LTRM learns robustly using a **Noisy Label Cleaning** strategy. Then, high-confidence predictions are used in **Highlight Pattern Induction** to distill genre-specific highlight patterns, which refine subsequent noisy label generation. This closed-loop design enables robust, annotation-free learning with progressively improving label quality.

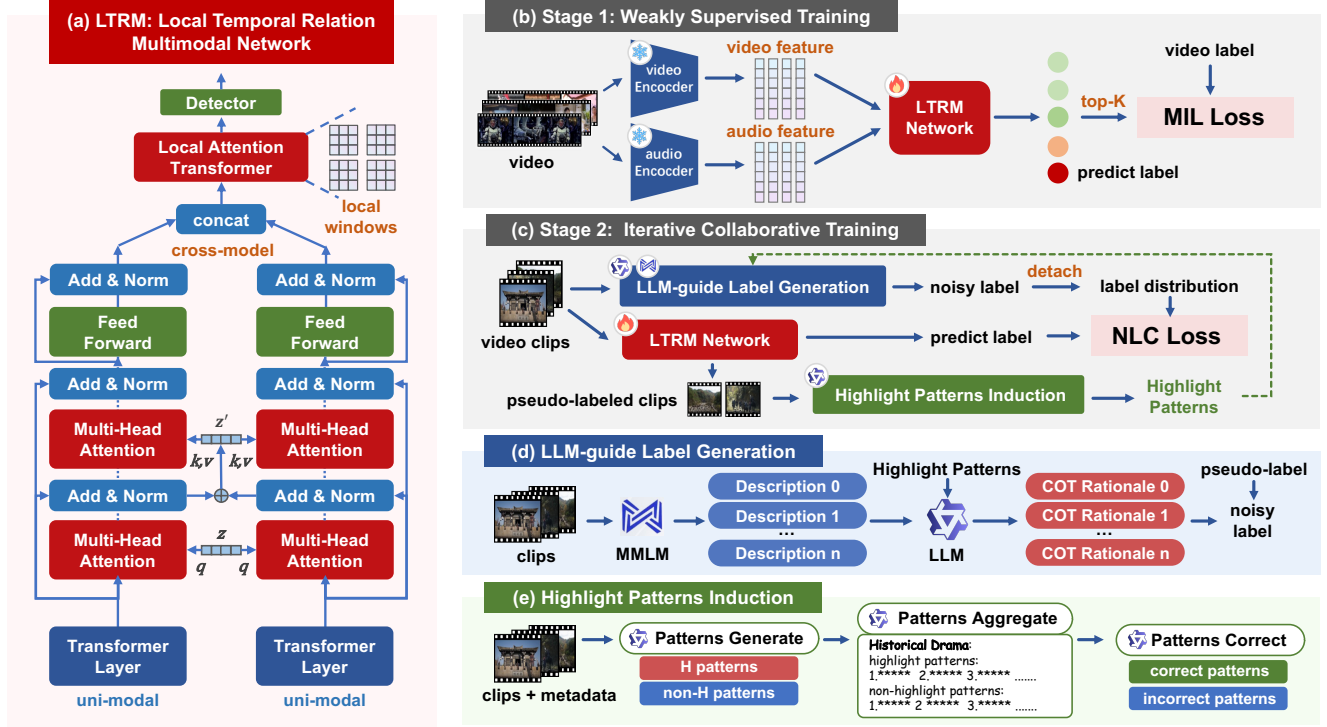


Figure 3. Our LTV-HD framework trains the (a) LTRM network for video highlight detection in two stages: weakly-supervised training and Iterative collaborative training. (b) The first stage uses Multiple Instance Learning (MIL) learning for weakly supervised training. (c) The pretrained model employs Noisy Label Cleaning (NLC) learning with (d) noisy labels generated by LLM. High-confidence data is then selected to derive highlight patterns through (e) Highlight Patterns Induction method, and refined patterns are used for further training.

4.1. Stage 1: Weakly-Supervised Pre-training

4.1.1. Local Temporal Relation Multi-Modal Network

As shown in the Figure 3 (a), the LTRM network consists of two uni-modal encoders, a cross-modal encoder, and a local temporal relation transformer layer, followed by a detector composed of multiple fully connected layers to output highlight scores.

For the cross-modal encoder, we draw on the implementation from [24], introducing bottleneck tokens $\{z_i\}_{i=1}^{N_z}$ as information bridges across modalities to capture and propagate multimodal features. N_z is much smaller than the number of segments T . Initially, z_i is used as the query in multi-head attention to aggregate modal features:

$$z'_i = z_i + w_z \sum_{t=1}^T \frac{\exp(w_q z_i \times w_k x_t)}{\sum_{n=1}^T \exp(w_q z_i \times w_k x_n)} w_v x_t. \quad (1)$$

Here, z_i and z'_i represent the input and output features of the bottleneck tokens, respectively, and $x \in \{v, a\}$ denotes the visual or audio features. After aggregating multimodal information, we use z'_i as the key and value in multi-head attention to propagate the information into each modality:

$$x'_i = x_i + w_z \sum_{t=1}^T \frac{\exp(w_q x_i \times w_k z_t)}{\sum_{n=1}^T \exp(w_q x_i \times w_k z_n)} w_v z_t. \quad (2)$$

Here, x'_i represents the visual or audio features enhanced by cross-modal interaction.

The highlight degree of a clip can be influenced by its context, and we prefer to focus on adjacent clips to reduce interference from distant, irrelevant ones. To achieve this, we reference the implementation of the Swin Transformer [25], introducing local self-attention windows to restrict the attention computation to a local scope rather than a global one, which enables the modeling of local temporal relations.

4.1.2. Multiple Instance Learning (MIL)

In the absence of accurate clip-level annotations, we train the LTRM using video-level labels $y \in \{0, 1\}$, where $y = 1$ indicates the presence of at least one highlight in the video. This setup follows the Multiple Instance Learning (MIL) paradigm [15], treating the video as a “bag” and its clips as “instances”. A bag with highlight instances is considered positive, while one without them is negative. We select the top K clips with the highest scores from output $\{s_t\}_{t=1}^T$, and the average score of these clips is used for video prediction:

$$\hat{y} = \frac{1}{K} \sum_{i \in \text{top-K}(s)} s_i. \quad (3)$$

We then compute the binary cross-entropy loss between

$\hat{y}$ and the label y to train the model.

4.2. Stage 2: Iterative Collaborative Training

As illustrated in Figure 3 (c), we introduce a collaborative training loop that combines three key components:

- **LLG**: LLM-guided Label Generation produces fine-grained but noisy clip-level labels.
- **NLC**: Noisy Label Cleaning Learning enables robust training from noisy labels.
- **HPI**: Highlight Pattern Induction distills reliable patterns from model predictions to improve label generation.

These components form a self-improving cycle without human annotation: LLG generates labels, NLC enables learning under noise, and HPI extracts knowledge from high-confidence predictions to guide LLG in the next iteration.

4.2.1. LLM-guided Label Generation (LLG)

LLG can automatically generate fine-grained highlight labels, significantly reducing annotation costs. We split the video into T clips and input them into a MLLM to generate descriptions $\{d_t\}_{t=1}^T$. These are then analyzed by a LLM using a chain-of-thought (CoT) [19, 38] approach:

- **video category**: Infer the category from descriptions (e.g., genres like historical, sci-fiction, romance in the TVHighlights dataset).
- **highlight analysis**: Assess highlight degree based on category and provide justification.
- **highlight score**: Output a highlight score s_n with a granularity of 0.1, where values closer to 1 indicate greater highlight degree.

This process yields noisy soft labels $Y_{llm} = \{y_{llm}\}_{t=1}^T$. To optimize, we generate pseudo-labels $Y_p = \{y_p\}_{t=1}^T$ using the model from stage one, and augment Y_{llm} with high-confidence labels from Y_p , resulting in optimized noisy labels $Y_n = \{y_n\}_{t=1}^T$.

4.2.2. Noisy Label Cleaning Learning (NLC)

During the NLC training, we utilize noisy soft labels Y_n . Given the noise in these labels, directly using standard losses (e.g., cross-entropy) could lead to overfitting to noise. Therefore, we employ a label refurbishing method from LNL field (see Section 2.2), which implements a noisy label cleaning approach by adaptively updating the labels to correct the noise.

We first convert the noisy soft labels Y_n into an updatable probability distribution Y_a^p , allowing for gradient computation, where $p = 2$ denotes the number of classes (highlight vs. non-highlight). The model is initialized with weights from the Stage 1 weakly-supervised training to mitigate the impact of noisy data in the early phases of training.

Then, we use the KL divergence to measure the difference between the soft label distribution Y_a^p and the output

of the model, serving as the **classification loss**:

$$\mathcal{L}_{cls} = \frac{1}{n} \sum_{i=1}^n D_{KL}(f(x_i; \theta) \| y_a^p), \quad (4)$$

$$D_{KL}(f(x_i; \theta) \| y_a^p) = f(x_i; \theta) \log \left(\frac{f(x_i; \theta)}{y_a^p} \right), \quad (5)$$

where $x_i = [v_i, a_i]$ represents the input multi-modal features, and f is the model function. During backpropagation, the gradient for Y_a^p is also calculated. We update Y_a^p using a higher learning rate to accelerate denoising while the model learns more reliable representations.

Since the noisy labels Y_n are largely accurate, we employ a **compatibility loss** to ensure that the updated Y_a^p does not deviate excessively from the original noisy labels Y_n :

$$\mathcal{L}_{cp}(Y_n, Y_a^p) = -\frac{1}{n} \sum_{i=1}^n y_{n,i} \log y_{a,i}^p. \quad (6)$$

Additionally, to prevent premature convergence, where both the output and label distributions stabilize at suboptimal intermediate values, we introduce a **sharpness loss** following [47]. This loss encourages the output to approach extreme values (0 or 1), as formulated below:

$$\mathcal{L}_s(f(x; \theta)) = -\frac{1}{n} \sum_{i=1}^n f(x; \theta) \log f(x; \theta). \quad (7)$$

The total NLC loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{cp} \mathcal{L}_{cp} + \lambda_s \mathcal{L}_s, \quad (8)$$

where λ_{cp} and λ_s are hyper-parameters that control the weights of the losses.

4.2.3. Highlight Patterns Induction (HPI)

To improve the quality of noisy labels generated by LLG, we introduce Highlight Pattern Induction. As shown in Figure 3 (e), HPI takes high-confidence pseudo-labeled clips produced by the model and distills them into structured, category-specific patterns via three steps:

- **patterns generation**: A LLM analyzes the selected clips and their associated metadata, including MLLM-generated descriptions, video categories, and pseudo-labels, to derive highlight and non-highlight patterns.
- **patterns aggregation**: Within each category, semantically similar patterns are merged, and infrequent or redundant ones are removed to form a representative set of patterns.
- **patterns correction**: The aggregated patterns are further filtered for logical coherence and alignment with common sense, pruning potential errors.

The refined patterns generated by HPI provide structured guidance for the LLG, enabling LLM to generate more accurate and consistent highlight labels in subsequent iterations.

Table 2. Comparison of our LTV-HD method with competing baselines on the TVHighlights dataset. We report performance after the first training stage (Stage 1) and for our final model (Stage 2), with results broken down by five video categories. Best results are in **bold** and the second best underlined. All metrics are in percentages (%). MR: direct moment retrieval, CS: per-clip scoring.

Methods	Action		Destruction		CGI		Emotion		Others		Total	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Video Highlight Detection Baselines												
UMT [24]	76.89	45.72	62.33	45.18	56.51	45.03	51.98	31.99	61.12	<u>29.82</u>	60.50	45.13
UniVTG[23]	66.31	34.90	71.90	36.21	56.97	44.65	51.42	32.34	60.80	16.07	55.20	31.71
Multimodal Large Language Models												
TimeChat[29]	50.17	32.93	50.96	36.74	50.82	38.13	50.90	20.91	50.41	12.98	50.77	26.18
Qwen-vl-max(MR)	60.08	41.06	66.38	44.69	66.07	51.73	64.02	31.61	61.82	22.76	64.80	36.62
Qwen-vl-max(CS)	85.03	57.49	80.93	65.39	<u>80.79</u>	62.48	<u>77.75</u>	<u>41.38</u>	79.14	27.74	88.40	47.42
Gemini-2.5-flash-06-17(MR)	54.73	43.39	58.21	45.86	63.24	60.79	58.98	35.27	59.54	29.45	54.87	40.91
Gemini-2.5-flash-06-17(CS)	<u>85.22</u>	58.95	<u>81.53</u>	61.05	79.14	58.18	76.80	38.80	76.27	24.47	87.52	46.10
Our Method												
LTV-HD (Stage 1)	84.25	<u>75.22</u>	80.70	<u>76.36</u>	80.57	<u>73.25</u>	67.72	41.24	63.55	25.96	87.34	<u>62.39</u>
LTV-HD (Stage 2)	90.85	83.97	86.90	84.17	85.63	78.68	85.13	60.00	89.04	53.71	92.74	71.20

5. Experiments

5.1. Experimental Settings

Benchmark. We evaluate on two benchmarks: *TVHighlights*, our proposed dataset for movies and TV dramas, and *YouTube Highlights* for cross-domain generalization and robustness testing. Since TVHighlights lacks clip-level training labels, all models in the experiment are evaluated using their original best-performing checkpoints for fair comparison. For YouTube Highlights, we intentionally avoid query-guided datasets (e.g., QVHighlights) to maintain our focus on query-agnostic detection. Its varied annotation quality makes it ideal for testing the resilience of our NLC module across three settings with increasing label noise: Harvested-highlight (clean), Harvested-Borderline (ambiguous), and MTurk (noisy).

Evaluation Metrics. For YouTube Highlights, we report mean Average Precision (mAP), consistent with prior work [11, 24]. On TVHighlights, we additionally report Area Under the Curve (AUC) due to class imbalance. To reduce bias from variable video lengths, AP and AUC are computed both per category and over the full test set to assess overall performance. We also employ Cohen’s Kappa [10] to measure the agreement between generated labels and human annotations, quantifying the evolution of label quality.

Implementation Details. Our approach uses distinct feature extraction strategies for each dataset. For TVHighlights, visual features are a concatenation of image features from CLIP (ViT-L-14) [28] and action features from SlowFast (R-50) [9]. Audio features are extracted using the Imagebind audio encoder [12]. For YouTube Highlights, we

follow [24] using I3D [4] for visual features and PANN [20] for audio features. We employ MiniCPM-V 2.6 [45] as the MLLM and DeepSeek-R1-Distill-Qwen [8] as the LLM. In Stage 1, which uses noisy video-level labels, we set top-K = 6 and a learning rate of 0.001, employing early stopping to mitigate noise-induced overfitting. In Stage 2, we set the NLC loss weights to $\lambda_{cp} = 0.1$, $\lambda_s = 0.4$ and execute 2 refinement round. The model is trained for 100 epochs with a learning rate of 0.0001 and a batch size of 32. All models are trained with the Adam optimizer [18]. Notably, we train a single unified video highlight detector for TVHighlights, unlike the per-category setup for YouTube Highlights.

5.2. Results on TVHighlights

We evaluate our LTV-HD framework on the TVHighlights dataset against two groups: established SOTA video highlight detection baselines and advanced multimodal large language models (MLLMs), as presented in Table 2. For MLLMs, we test two distinct prompting strategies: direct moment retrieval (MR) and per-clip scoring (CS), with further details in the Appendix B.

The results demonstrate that **LTV-HD achieves a new SOTA performance across all categories and metrics**. LTV-HD (Stage 2) attains 92.74% AUC and 71.20% AP on the full test set, significantly outperforming the strongest baseline, Qwen-vl-max (CS), which achieves 88.40% AUC and 47.42% AP. Even our initial model (Stage 1) surpasses all baselines in AP (62.39%), demonstrating the strength of the weakly-supervised MIL foundation. The performance gain of +5.40% in AUC and +8.81% in AP from Stage 1 to Stage 2 confirms the efficacy of our two-stage refinement.

Further analysis reveals two key trends. First, **MLLMs struggle with precise temporal localization**, which severely limits their performance under the moment retrieval (MR) paradigm. MR requires the model to not only identify highlight content but also accurately localize its start and end times. In contrast, per-clip scoring (CS) addresses this challenge by reducing the task to clip-level classification. This is evidenced by the consistent superiority of the CS strategy over MR, for example, Qwen-vl-max achieves only 36.62% AP under MR, while improving to 47.42% under CS.

Second, **performance varies significantly across semantic categories**. Most models perform better on visually salient events like Action and Destruction, where appearance and motion cues are strong, than those requiring nuanced understanding such as Emotional Climax. For instance, UMT achieves 45.72% AP on Action but only 31.99% on Emotion. In contrast, LTV-HD (Stage 2) achieves 60.00% AP on Emotion, a +18.62% absolute gain over the best MLLM (Qwen-vl-max, CS: 41.38%), demonstrating that our method, which learns diverse highlight patterns directly from data, is more robust than methods reliant on rigid and predefined semantic categories.

5.3. Results on YouTube Highlights

Our work aims to learn robust video highlight detectors from noisy, large-scale data. We evaluate on YouTube Highlights in two settings: clean dataset (standard benchmark) and three noisy subsets.

As shown in Table 3, we first evaluate our method on six categories of YouTube Highlights. Our method achieves 75.8% average mAP, which is competitive with SOTA methods including UVCOM2 (77.4%) and UniVTG (75.2%), demonstrating strong performance under clean supervision.

More critically, our approach shows superior noise resilience. In Table 4, our approach achieves superior overall performance across three noisy settings. Notably, while UMT drops by 1.54% under MTurk noise, our method degrades only 0.32%, maintaining consistent gains across all categories. This stability validates the effectiveness of our NLC module in distilling reliable signals from noisy label.

5.4. Ablation Study

Training Options and Modules for LTV-HD. The ablation study in Table 5 validates the contribution of each training stage in LTV-HD. Initial weakly supervised training is critical: its inclusion improves AP by 4.24 points, and skipping it results in degraded performance even when later components are enabled (e.g., Row 3 vs. Row 5) Furthermore, refining labels via HPI in Stage 2 is vital: models trained on unrefined pseudo-labels suffer significant performance degradation (Row 4 vs. Row 5), confirming that effective

Table 3. Results on clean YouTube Highlights dataset.

Methods	dog	gym.	park.	ska.	ski.	surf.	Avg.
GIFs [13]	30.8	33.5	54.0	55.4	32.8	54.1	46.4
LSVM [32]	60.0	41.0	61.0	62.0	36.0	61.0	53.6
HighlightMe [3]	63.0	73.0	72.0	64.0	52.0	62.0	64.0
MINI-Net [15]	58.2	61.7	70.2	72.2	58.7	60.1	64.4
SL-Module [43]	70.8	53.2	77.2	72.5	66.1	76.2	69.3
Joint-VA [2]	64.5	71.9	80.8	62.0	73.2	78.3	71.8
PLD [37]	74.9	70.2	77.9	57.5	70.7	79.0	73.0
UMT [24]	65.9	75.2	81.6	71.8	72.3	82.7	74.9
UniVTG [23]	<u>71.8</u>	<u>76.5</u>	73.9	73.3	73.2	82.2	75.2
UVCOM2 [40]	66.5	77.4	82.8	78.7	74.2	84.6	77.4
Ours	66.8	72.6	83.5	<u>74.6</u>	<u>73.3</u>	84.8	<u>75.8</u>

Table 4. Performance comparison on the noisy YouTube Highlights benchmark under three settings with increasing label noise. H-light: Harvested highlight; H-border: Harvested borderline.

Label	Method	dog	gym	park	ska	ski	surf	Avg.
H-light	UMT	65.90	75.20	81.60	71.80	72.30	82.70	74.90
	Ours	66.80	72.60	83.50	74.60	73.30	84.80	75.80
H-border	UMT	65.93	74.31	81.58	71.84	70.24	82.46	74.39
	Ours	66.57	72.97	81.71	72.26	71.59	83.47	74.76
MTurk	UMT	63.78	76.16	75.02	73.62	69.99	81.59	73.36
	Ours	66.75	72.94	83.76	75.63	70.56	83.24	75.48

Table 5. Ablation study on the training stages and modules of LTV-HD. We assess the impact of utilizing different training rounds and key components. s: training stage, r: refinement round.

Stage 1	Stage 2			Performance	
Weak Sup.	NLC(r1)	HPI(r2)	NLC(r2)	AUC	AP
✗	✗	✗	✗	89.12	63.45
✓	✗	✗	✗	90.20	67.69
✗	✓	✗	✗	90.36	66.72
✗	✓	✓	✓	91.85	70.01
✓	✓	✓	✓	92.74	71.20

highlight patterns are essential for fine-tuning. Finally, incorporating NLC learning provides a consistent gain, surpassing the standard cross-entropy baseline. This demonstrates that our noise-aware learning framework effectively enhances label reliability and model generalization.

Losses in NLC Learning. Table 6 illustrates the performance of LTV-HD under different loss combinations and label update strategies in NLC learning. Notably, performance drops significantly when labels are not updated, likely due to overfitting to noise during training. Excluding the sharpness loss also degrades results, possibly because the training converges too early under label update influence. Finally, the compatibility loss provides an additional performance boost.

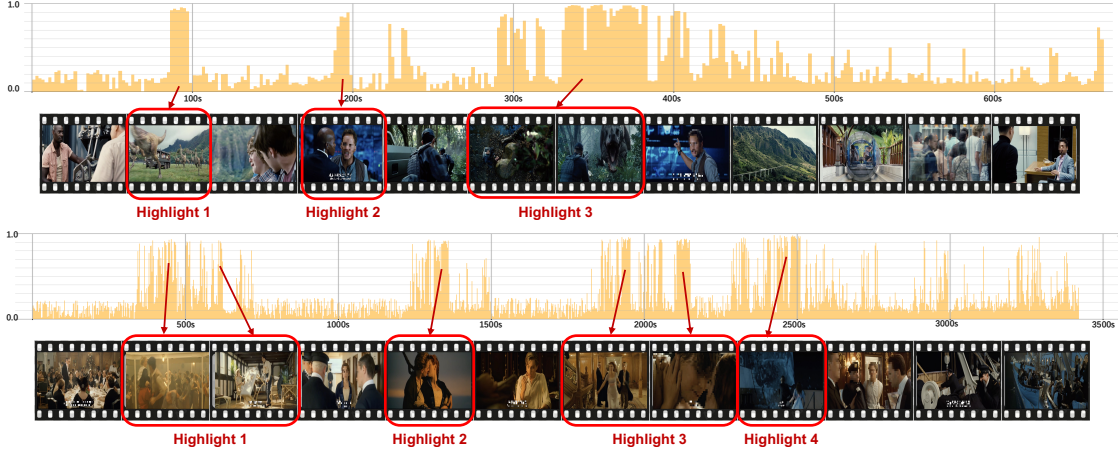


Figure 4. Results of our method on Jurassic World and Titanic. For Jurassic World at the top, highlights include: (1) running dinosaurs, (2) a argument, and (3) indominus rex attacking soldiers. For Titanic at the bottom, highlights include: (1) a party dance and objects being knocked over, (2) a romantic kiss, (3) a chase and intimate moments, and (4) the iceberg collision.

Table 6. Ablation studies on NLC Learning.

Label Update	Loss _{cls}	Loss _{cp}	Loss _s	AUC	AP
✗	✓	✗	✗	91.01	66.05
✓	✓	✗	✗	90.76	70.21
✓	✓	✓	✗	90.60	70.85
✓	✓	✗	✓	93.31	69.03
✓	✓	✓	✓	92.74	71.20

5.5. Label Quality and Model Evolution

To quantitatively validate Our LTV-HD framework, we evaluate the alignment between predicted labels and human annotations across stages, as shown in Table 7. The initial weak labels show moderate agreement with human annotations (video AUC: 86.81, Kappa: 62.47), but lack clip-level supervision. Our Stage 1 model leverages these to achieve strong clip-level performance (AP: 62.39), already surpassing general-purpose MLLMs like Qwen and Gemini, proving the advantage of a specialized architecture under weak supervision. In Stage 2, iterative refinement drives both higher metrics and better label alignment with humans: from s1 to s2-r2, clip AP increases by +8.81, while Kappa rises from 52.41 to 60.54, which provides direct evidence of effective denoising. Further improvements from r1 to r2 confirm that multi-round refinement yields consistent benefits. These results validate our core strategy: start with a robust, task-specific baseline, then progressively refine both labels and model predictions.

5.6. Visualization

Figure 4 visualizes the video highlight detection results of our lightweight LTV-HD model. Each frame sequence is paired with the corresponding highlight score curve, where detected highlights are marked by red bounding boxes. The

Table 7. Quantitative analysis of label quality and model performance at different stages. We compare initial weak labels, model predictions (LTV-HD, MLLMs), and human annotations.

Modeling	Video-level			Clip-level		
	AUC	AP	Kappa	AUC	AP	Kappa
Week label (s1)	86.81	81.71	62.47	-	-	-
Qwen-vl-max	78.35	78.10	47.63	88.40	47.42	49.15
Gemini	76.28	86.10	65.13	87.52	46.10	52.97
LTV-HD (s1)	83.05	82.45	52.91	87.34	62.39	52.41
LTV-HD (s2-r1)	86.68	89.11	56.28	92.54	69.78	55.72
LTV-HD (s2-r2)	90.62	89.39	65.68	92.74	71.20	60.54

examples demonstrate that our model accurately localizes meaningful highlight moments even under complex visual conditions (e.g., fast motion, low lighting), confirming the robustness and generalization capability of LTV-HD. More illustrative cases are provided in Appendix C.

6. Conclusion

In this paper, we introduce a new TVhighlights dataset for video highlight detection in movies and TV dramas, and we propose a LTV-HD framework, which includes weakly supervised and collaborative training stages. We design a LLM-guided label generation process to produce fine-grained annotations. In collaborative training, we develop a Noisy Label Cleaning Learning method to robustly train with noisy labels, and introduce a Highlight Patterns Induction method to derive highlight-related patterns for optimized labels. Experiments on TVhighlights and YouTube Highlights datasets confirm the effectiveness of our framework. Future work will focus on expanding the applicability of LTV-HD for broader video highlight detection.

References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 312–321. PMLR, 2019. 2
- [2] Taivanbat Badamdorj, Mrigank Roachan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8107–8117. IEEE, 2021. 2, 7
- [3] Uttaran Bhattacharya, Gang Wu, Stefano Petrangeli, Viswanathan Swaminathan, and Dinesh Manocha. Highlightme: Detecting highlights from human-centric videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8137–8147. IEEE, 2021. 2, 7
- [4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017. 6
- [5] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11442–11450. AAAI Press, 2021. 2
- [6] Runnan Chen, Penghao Zhou, Wenzhe Wang, Nenglun Chen, Pai Peng, Xing Sun, and Wenping Wang. Pr-net: Preference reasoning for personalized video highlight detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7960–7969. IEEE, 2021. 2
- [7] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2
- [8] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 6
- [9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6
- [10] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971. 6
- [11] Bei Gan, Xiujun Shu, Ruizhi Qiao, Haoqian Wu, Keyu Chen, Hanjun Li, and Bo Ren. Collaborative noisy label cleaner: Learning scene-aware trailers for multi-modal highlight detection in movies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24070–24079. IEEE, 2023. 2
- [12] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 6
- [13] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1001–1009. IEEE Computer Society, 2016. 2, 7
- [14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546, 2018. 2
- [15] Fa-Ting Hong, Xuanteng Huang, Weihong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, pages 345–360. Springer, 2020. 2, 4, 7
- [16] Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2
- [17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 2
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [19] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 5
- [20] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28: 2880–2894, 2020. 6
- [21] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, pages 11846–11858. Curran Associates, Inc., 2021. 2, 3
- [22] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. DISC: learning from noisy labels via dynamic instance-specific selection and correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24070–24079. IEEE, 2023. 2

- [23] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univt: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 2, 6, 7
- [24] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3032–3041. IEEE, 2022. 2, 4, 6, 7
- [25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 4
- [26] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 6543–6553. PMLR, 2020. 2
- [27] WonJun Moon, Sangeek Hyun, Sanguk Park, Dongchan Park, and Jae-Pil Heo. Query - dependent video representation for moment retrieval and highlight detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23023–23033. IEEE, 2023. 2
- [28] 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. 6
- [29] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14313–14323. IEEE, 2024. 2, 6
- [30] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5179–5187. IEEE Computer Society, 2015. 2, 3
- [31] Min Sun, Ali Farhadi, and Steven M. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 787–802. Springer, 2014. 2, 3
- [32] Min Sun, Ali Farhadi, and Steven M. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 787–802. Springer, 2014. 2, 7
- [33] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. Detecting highlights in sports videos: Cricket as a test case. In *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, ICME 2011, 11-15 July, 2011, Barcelona, Catalonia, Spain*, pages 1–6. IEEE Computer Society, 2011. 2
- [34] Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Watch video, catch keyword: Context-aware keyword attention for moment retrieval and highlight detection. *CoRR*, abs/2501.02504, 2025. 2
- [35] Jinjun Wang, Changsheng Xu, Chng Eng Siong, and Qi Tian. Sports highlight detection from keyword sequences using HMM. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan*, pages 599–602. IEEE Computer Society, 2004. 2
- [36] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 322–330. IEEE, 2019. 2
- [37] Fanyue Wei, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Learning pixel-level distinctions for video highlight detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3063–3072. IEEE, 2022. 2, 7
- [38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5
- [39] Xiaobo Xia, Bo Han, Yibing Zhan, Jun Yu, Mingming Gong, Chen Gong, and Tongliang Liu. Combating noisy labels with sample selection by mining high-discrepancy examples. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1833–1843. IEEE, 2023. 2
- [40] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18709–18719, 2024. 7
- [41] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1258–1267. Computer Vision Foundation / IEEE, 2019. 1, 2
- [42] Changsheng Xu, Jinjun Wang, Hanqing Lu, and Yifan Zhang. A novel framework for semantic annotation and personalized retrieval of sports video. *IEEE Trans. Multim.*, 10 (3):421–436, 2008. 2
- [43] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *2021 IEEE/CVF*

- International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7950–7959. IEEE, 2021. [1](#), [2](#), [7](#)
- [44] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 982–990. IEEE Computer Society, 2016. [2](#)
 - [45] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. [6](#)
 - [46] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7930–7939. IEEE, 2021. [2](#)
 - [47] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7017–7025. Computer Vision Foundation / IEEE, 2019. [2](#), [5](#)