TRINITY: A Multi-Perspective Benchmark for Personal-Style Video Highlight Detection

Qianqian Chen*, Hyun Bin Kim*, Denzel Elden Wijaya, Yang Yi, Bo Liu✉, and Yangkai Ding✉

Huawei Technologies Co., Ltd.

Abstract. Traditional video highlight detection relies on a narrow, event-centric definition of saliency, which often fails to generalize to unconstrained personal videos where highlights are heterogeneous and perspective-dependent. To address this, we introduce TRINITY, a multi-perspective benchmark that decomposes highlight saliency into three complementary dimensions, Event, Emotion, and Nature, within a unified temporal framework. Leveraging this multi-faceted view, we propose a shared-backbone multi-branch architecture designed for parallel multi-perspective prediction via view-specific experts. Comprehensive experiments demonstrate that our method significantly outperforms state-of-the-art baselines, achieving gains of $+7.15/+3.62$ $\text{mAP}_{\rho=15\%/50\%}$ on Mr. HiSum and +10.82 mAP on YouTube Highlights. These results validate that multi-perspective modeling provides a more robust and comprehensive formulation of video saliency, especially for complex real-world scenarios. The benchmark and relevant codes will be released upon acceptance. The benchmark is available at `https://huggingface.co/datasets/vanilladucky/TRINITY`.

Keywords: Highlight Detection · Personal Videos · Multi-Perspective Learning · Benchmark

1 Introduction

Video highlight detection aims to identify moments that strongly shape human attention and memory [13]. Despite rapid progress, existing benchmarks largely adopt a scenario-bound and narrowly defined notion of saliency. Most datasets are constructed around sports competitions, TV programs, or query-conditioned retrieval tasks [19, 34, 36, 38], where highlights are equated with salient semantic events or narrative peaks. While this single-perspective formulation has driven strong performance within controlled settings, it inherently constrains the conceptual scope of what constitutes a highlight. Consequently, models trained under this paradigm often struggle to generalize to unconstrained videos, such as those capturing daily life, where saliency is subtle, heterogeneous, and frequently decoupled from dramatic or high-intensity actions.

* Equal contribution.
✉ Corresponding authors.

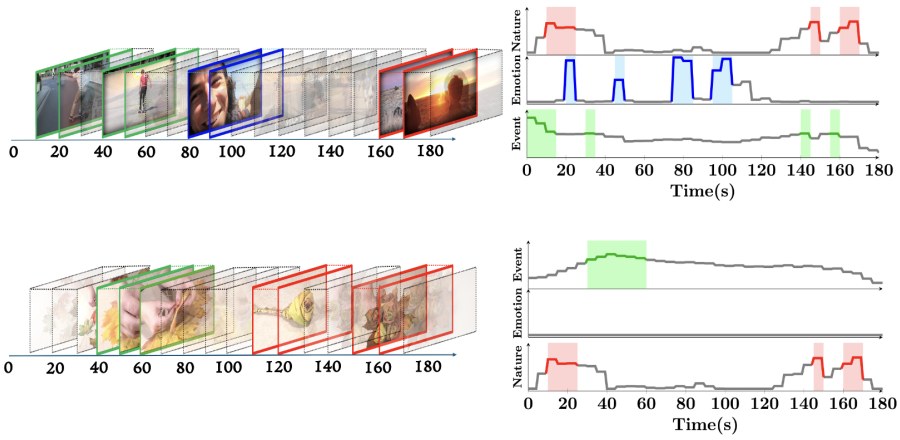


Fig. 1: Visualization of highlight distributions from three perspectives—*Event*, *Emotion*, and *Nature*—for two example videos. The horizontal axis denotes time (in seconds), and the vertical axis represents the highlight score.

In real life, personal-style videos dominate social media. These recordings typically lack explicit story structures or professionally edited climaxes [7, 16, 32]. Engaging moments may instead emerge from affective facial expressions, intimate social interactions, or aesthetically pleasing scenery, where salient signals are often multi-causal and perspective-dependent rather than single one. For instance, in a personal vlog, highlights may stem from both scenic views and the vlogger's joyful expressions. Modeling highlight detection as a single-dimensional prediction task thus reduces a fundamentally multi-faceted phenomenon to a narrow formulation. We argue that advancing highlight detection requires moving beyond scenario-specific supervision toward a paradigm that explicitly models heterogeneous saliency mechanisms within a single video.

To this end, we introduce **TRINITY**, a multi-perspective benchmark that decomposes highlight saliency into three complementary dimensions within a unified temporal framework. Specifically, highlights are annotated along three perspectives: event-driven semantic peaks (denoted as *Event*), affective facial moments (*Emotion*), and aesthetic scenic highlights (*Nature*), which may coexist within a single video as shown in Fig. 1. The Emotion and Nature perspectives are constructed through a scalable automatic annotation pipeline to ensure consistency and reproducibility, while the Event perspective follows established event-centric annotations. Importantly, the three perspectives are intentionally designed to be complementary rather than redundant, giving rise to cross-dimensional orthogonality and natural view sparsity in highlight annotations. By disentangling these heterogeneous saliency signals, TRINITY enables a more structured analysis of highlight patterns beyond the traditional paradigm.

Building on TRINITY, we further propose a novel shared-backbone multi-branch architecture with view-specific experts. The shared backbone models global temporal structure and contextual coherence, while each branch specializes in localizing video highlights under a specific perspective (i.e., Event, Emotion, and Nature). This structural factorization enables efficient parallel multi-perspective prediction from a single input sequence, naturally aligning with the heterogeneous saliency mechanisms in personal videos. Comprehensive experiments demonstrate that our model establishes a strong baseline on TRINITY and achieves state-of-the-art performance on existing event-centric benchmarks and an emotion-centric benchmark. Extensive ablation studies further verify the effectiveness of the proposed components.

In summary, our main contributions are as follows:

- We introduce TRINITY, a novel multi-perspective benchmark that shifts the highlight detection paradigm from single dimension to heterogeneous saliency modeling. It decomposes highlights into three dimensions, Event, Emotion, and Nature, within a unified temporal framework.
- We propose a shared-backbone multi-branch architecture featuring view-specific experts. This design explicitly factorizes global temporal context from perspective-specific localization, enabling the model to parallelize the detection of diverse saliency signals.
- Comprehensive evaluations demonstrate that our approach not only establishes a robust baseline on TRINITY but also achieves state-of-the-art performance on traditional event-centric and emotion-centric benchmarks, proving its superior generalization across diverse video types.

2 Related Work

2.1 Video Highlight Detection

Video highlight detection aims to automatically localize salient segments in untrimmed videos. Existing methods generally follow two main paradigms in the literature: *query-conditioned* and *text-agnostic* modeling.

Query-conditioned highlight detection: QVHighlights [19] formulates highlight detection as query-conditioned moment retrieval with clip-level saliency prediction. Subsequent works strengthen cross-modal interaction through transformer-based grounding frameworks for query-guided localization [11, 21, 22, 25, 37, 45]. Although such methods reduce semantic ambiguity by anchoring saliency to user intent, they rely on textual queries at inference time and are therefore unsuitable for automatic highlight detection in personal videos.

Text-agnostic highlight detection: Text-agnostic approaches predict saliency directly from visual and audio cues. Early methods rely on edited-video supervision or ranking objectives [38, 42], while more recent works emphasize cross-category generalization and audio–visual modeling [2, 41, 43]. Despite these advances, most

approaches implicitly optimize a single notion of saliency, typically event-driven, and do not explicitly model heterogeneous highlight patterns within a unified framework. In this work, we decompose highlight saliency into complementary visual perspectives under a shared temporal protocol, enabling modeling of multiple heterogeneous patterns without textual conditioning.

2.2 Video Highlight Detection Benchmarks

Highlight detection is inherently subjective, and most benchmarks rely on multiple annotators to stabilize labels. Early datasets such as YouTube Highlights [38] provide segment-level annotations, while video summarization datasets such as TVSum [34] are often repurposed by aggregating shot-level importance scores into highlight labels. Another line of work exploits user-generated artifacts as implicit supervision. Video2GIF [8] treats user-created GIFs as highlight signals, and PHD^2 [24] extends this setting to personalized highlight detection using user histories. More recently, large-scale implicit-feedback datasets such as Mr. HiSum [36] aggregate engagement statistics to derive highlight saliency from massive viewer interactions. Despite differences in supervision sources, most existing benchmarks implicitly assume a single notion of saliency per video. In contrast, our dataset decomposes highlight saliency into three complementary perspectives within a unified temporal framework, enabling structured analysis of heterogeneous highlight patterns across diverse personal videos.

3 TRINITY Dataset

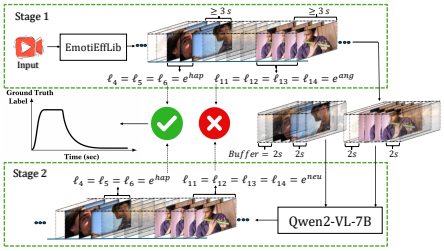
3.1 Dataset Overview

TRINITY is a multi-perspective highlight benchmark built on open-sourced videos, organizing highlights into three perspectives: Event, Emotion, and Nature. Emotion- and nature-driven highlights are newly annotated over the entire video pool using a scalable two-stage pipeline. For the Event dimension, we directly adopt the replay-based annotations from Mr. HiSum [36]. The following subsections describe the construction process for each perspective in detail.

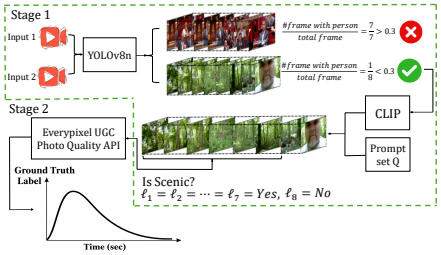
3.2 Event Perspective: Semantic Progression

Highlight Definition. Following the replay-based supervision mechanism introduced in Mr. HiSum, replay frequency derived from aggregated "Most Replayed" statistics is treated as a proxy for event-driven highlights. Segments with consistently high replay intensity are assumed to correspond to temporally localized semantic events that attract collective viewer attention.

Label Construction. Following prior practice in [15], frame-level importance scores are aggregated onto a unified 5-second clip grid via average pooling. To prevent peak dilution, clips containing maximal frame-level values retain the peak score. This preserves the relative ordering of segments.



(a) Two-stage pipeline for emotion-driven highlight annotation.



(b) Two-stage pipeline for nature-driven highlight annotation.

Fig. 2: Overview of the annotation pipelines used to construct the TRINITY dataset. The Emotion pipeline detects temporally stable facial expressions and verifies candidate segments via multimodal reasoning. The Nature pipeline identifies scenic moments through semantic filtering and aesthetic assessment, retaining segments with both landscape relevance and strong visual appeal.

3.3 Emotion Perspective: Affective Facial Moments

Highlight Definition. Emotion-driven highlights are segments where affective signals are conveyed through observable facial expressions and emotional reactions.

Label Construction. Emotion labels are constructed via a two-stage automatic pipeline that combines perceptual-level expression screening with independent semantic-level affect reasoning, ensuring consistent category assignments across models, as shown in Fig. 2 (a).

Stage-1: Temporally consistent expression screening. Videos are uniformly sampled at 1 fps to obtain a sequence of frames $\{x_t\}_{t=1}^{T}$. We apply EmotiEffLib [30,31], a frame-level facial expression recognition model, to assign labels $\ell_t \in \mathcal{E}_1$, where $\mathcal{E}_1 = \{e^{\text{neu}}, e^{\text{hap}}, e^{\text{sad}}, e^{\text{ang}}, e^{\text{fear}}, e^{\text{dis}}, e^{\text{sup}}\}$ denotes the seven basic emotion categories. To suppress noise from transient micro-expressions, the frame x_t is considered affective-valid if it belongs to a stable non-neutral window, that is, $\ell_t = \ell_{t+1} = \ell_{t+2} = c$ with $c \in \mathcal{E}_1 \backslash \{e^{\text{neu}}\}$, thereby requiring emotion labels to remain stable across at least three consecutive frames. Candidate segments $\mathcal{S}$ are then constructed by merging maximal contiguous intervals of affective-valid frames. Each segment $s \in \mathcal{S}$ is defined as $s = \{t \mid t_{\text{start}} \leq t \leq t_{\text{end}}\}$, where all frames share the same sustained non-neutral emotion category $c_s \in \mathcal{E}_1 \setminus \{e^{\text{neu}}\}$.

Stage-2: Independent context-aware multimodal screening. For each candidate temporal segment s, we construct an expanded temporal window $\mathcal{B}(s) = \{t \mid t_{\text{start}} - \delta \leq t \leq t_{\text{end}} + \delta\}$ with $\delta = 2$. The corresponding clip is processed by a large vision-language model (Qwen2-VL-7B [39] we used), denoted as $\mathcal{M}(\cdot)$. It analyzes the visual and contextual cues within the expanded window and predicts an emotion category $\mathcal{M}(\mathcal{B}(s)) \in \mathcal{E}_1$. A segment s is retained as a ground-truth Emotion highlight only when $\mathcal{M}(\mathcal{B}(s)) = c_s$, where c_s denotes the Stage-1 emotion label. For a retained segment s, we assign $y_t = 1$ for all $t \in s$, and $y_t = 0$ otherwise.

Conceptually, Stage-1 detects salient facial dynamics and expression intensity at the perceptual level. Stage-2 complements by performing semantic reasoning over multimodal context, assessing whether the segment conveys a coherent affective meaning. By keeping only segments where both models predict the same category, the pipeline yields more reliable emotion-driven highlight annotations.

Annotation Quality. Annotation consistency is evaluated via cross-annotator agreement among three independent annotators on a randomly sampled set of 50 annotated segments. The inter-annotator agreement reaches 92%, indicating stable labeling consistency across annotators. Under majority voting, 94% of the emotion labels are verified as correct. Detailed agreement analyses are reported in the Appendix.

3.4 Nature Perspective: Scenic Aesthetic Saliency

Highlight Definition. Nature-driven highlights are defined as segments dominated by visually salient scenic compositions with strong aesthetic intensity, independent of explicit human actions.

Label Construction. Nature annotations are constructed via coarse scenic localization followed by frame-level aesthetic scoring within scenic segments, enabling identification of visually appealing landscape moments (Fig. 2(b)).

Stage-1: Coarse scenic localization. Videos are uniformly sampled at 1 fps to obtain frame sequences $\{x_t\}_{t=1}^{T}$. For each frame, YOLOv8n [12] is applied to detect persons, producing a binary indicator h_t. Videos whose person-frame ratio $R_{\text{per}} = \frac{1}{T}\sum_{t=1}^{T} h_t$ satisfies $R_{\text{per}} > \theta_v^{\text{per}}$ with $\theta_v^{\text{per}} = 0.3$ are discarded.

For the remaining videos, CLIP [26] similarities are computed between each frame x_t and a curated prompt set $\mathcal{Q}$ containing landscape and human descriptions. Frame-level probabilities are obtained via softmax over prompt similarities and aggregated into landscape and person scores p_t^{land} and p_t^{per}, respectively. A frame is marked scenic if $l_t = \mathbb{1}\big[p_t^{\text{land}} > 0.4 \wedge p_t^{\text{land}} > p_t^{\text{per}}\big]$, indicating that the frame is more likely to depict landscape content than human-centric scenes. All thresholds are selected based on empirical validation experiments.

Stage-2: Frame-level aesthetic scoring. For each validated scenic segment $s \in \mathcal{S}$, we obtain frame-level aesthetic scores $\{E_t \mid t \in s, E_t \in [0,1]\}$ using the external Everypixel UGC Photo Quality, a pretrained aesthetic assessment service. The model analyzes multiple visual attributes, including sharpness, exposure, and composition, producing a normalized aesthetic score that reflects overall perceptual quality. These scores are directly used as regression targets for the Nature dimension. Specifically, the ground-truth label for a frame t is defined as $y_t = E_t$ if $\exists s \in \mathcal{S}$ such that $t \in s$, and $y_t = 0$ otherwise. This stage is applied only when $\mathcal{S} \neq \emptyset$, ensuring supervision is derived from regions that satisfy both scenic alignment and strong aesthetic quality.

Table 1: Global statistics of the three highlight perspectives in the full TRINITY dataset.

Perspective	#Videos	#Peaks	Avg. Peak Dur. (s)
TRINITY-Nature	15,540	15,540	10.34
TRINITY-Emotion	16,257	32,046	4.87
TRINITY-Event	27,845	55,649	9.90

Annotation Quality. Given the subjective nature of aesthetic judgment, we conduct a human audit with three independent annotators on 50 randomly sampled videos. For scenic validity, annotators label scenic frames for each video, and consensus scenic segments are derived via majority voting over frame-level labels. The average IoU between the consensus segments and the labeled ground-truth segments is 0.89727, while the average pairwise inter-annotator IoU is 0.89744, indicating that the dataset labels align with human annotations at a level nearly identical to the agreement among annotators. For aesthetic quality validation, 94% of the aesthetic highlights are validated by majority voting over the annotators' judgments, with an inter-annotator agreement rate of 90%. To further ensure reproducibility and mitigate proprietary-model bias, we evaluate the cross-scorer agreement of our Everypixel-based[1] Nature labels against alternative open-source aesthetic models (MUSIQ [14] and LAION-Aesthetics[2]) across 5,651 annotated videos. Our labels exhibit strong alignment, achieving an IoU of 0.665 at $\rho = 15\%$ and 0.956 at $\rho = 50\%$ with MUSIQ, and 0.661 at $\rho = 15\%$ and 0.955 at $\rho = 50\%$ with LAION, confirming that the annotations capture general, scorer-agnostic aesthetic signals.

3.5 Statistical Overview

To provide a comprehensive view of the proposed benchmark, Table 1 summarizes the global statistics of the three highlight perspectives across the entire TRINITY dataset, including the total number of videos, annotated highlight peaks, and the average duration per peak.

The statistics reveal clear domain and temporal distinctiveness across the three formulated perspectives. The TRINITY-Event perspective comprises the largest video pool and peak count, capturing broad semantic activities. In contrast, the temporal characteristics vary significantly between Emotion and Nature: emotion-driven highlights are typically short and highly concentrated ($\sim$4.87 s), capturing transient facial and behavioral expressions; conversely, nature-driven highlights feature longer sustained durations ($\sim$10.34 s), as aesthetic scenic compositions generally unfold over continuous camera periods.

These macro-level structural variations inherently support our core motivation to model video highlights via a multi-perspective formulation. To further

[1] `https://api.everypixel.com/v1/quality_ugc`
[2] `https://github.com/christophschuhmann/improved-aesthetic-predictor`

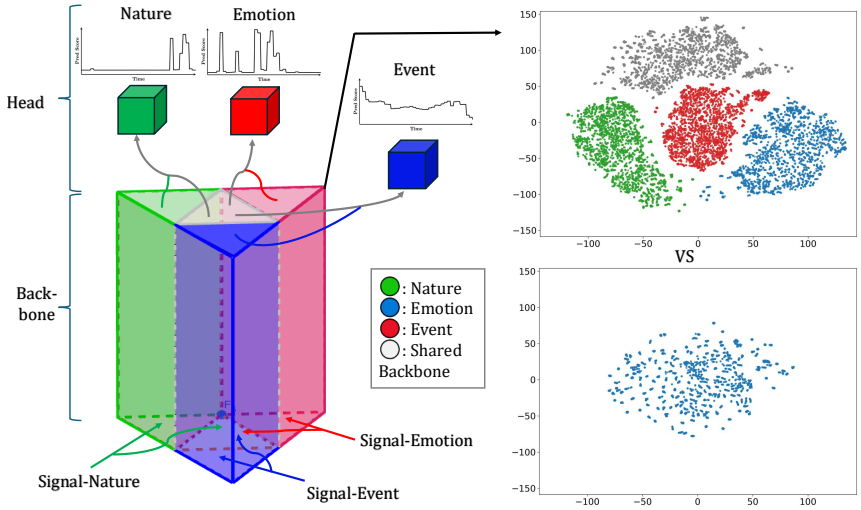


Fig. 3: Left. Overview of the proposed architecture with a shared backbone and perspective-specific experts. Right. Visualization of the learned embedding space, where private experts form distinct clusters while the shared backbone captures global temporal context. Compared with a single-task event-only model (bottom-right), the proposed design yields a structured and separable representation space across perspectives.

investigate how these perspectives interact and validate their statistical orthogonality, we provide a detailed cross-perspective temporal overlap analysis on the co-annotated subset in the Appendix.

4 Methodology

4.1 Overall Architecture

As illustrated in Fig. 3, we formulate video highlight detection as a *multi-task* temporal segment scoring problem with three highlight perspectives: **Event** (t=e), **Emotion** (t=m), and **Nature** (t=n). Given a video v, we split it into N temporal segments $\{s_i\}_{i=1}^{N}$, where each segment represents a non-overlapping 5-second interval. The emotion task ($t = m$) provides a binary highlight label $y_i^t \in \{0, 1\}$, reflecting the discrete presence of facial expressions, while the event ($t = e$) and nature ($t = n$) tasks provide continuous labels $y_i^t \in [0, 1]$ derived from replay statistics and aesthetic scores, respectively. Each segment is represented by a pre-extracted 512-dimensional feature vector $x_i \in \mathbb{R}^{512}$.

Our architecture contains one *shared* temporal backbone T_s and three *private* temporal backbones $\{T_p^t\}_{t\in\{e,m,n\}}$. All backbones are instantiated as Transformer encoders. In contrast to permutation-invariant set encoders, we explicitly model *ordered* temporal segments and inject relative temporal information via 1D RoPE [35] within self-attention layers.

The shared backbone T_s is designed to capture task-agnostic, global temporal context across video timestamps. In parallel, each private backbone T_p^t specializes in task-specific temporal patterns, emphasizing sharp and sparse temporal activations that are characteristic of task t. Let $X = [x_1, \ldots, x_N]$ denote the sequence of segment embeddings. Concretely, for each task t, we parameterize the shared and private streams with distinct query, key, and value projections, i.e.,

$$\begin{aligned} Q_s &= XW_Q^{(s)}, \qquad K_s = XW_K^{(s)}, \qquad V_s = XW_V^{(s)}, \\ Q_p^t &= XW_Q^{(p,t)}, \qquad K_p^t = XW_K^{(p,t)}, \qquad V_p^t = XW_V^{(p,t)}. \end{aligned} \tag{1}$$

These separate projection matrices encourage the shared and task-private representations to occupy different regions of the latent space, enabling complementary modeling of global context and task-specific temporal cues.

The input sequence X is passed through the shared backbone to produce a globally contextualized embedding, while the same sequence is processed by the task-specific private backbone to produce a task-specific embedding. These embeddings are concatenated to capture both global and task-specific semantics. Finally, three independent prediction heads $\{C^t\}$ (one per task) are used, each mapping z_i^t to a highlight probability. Each prediction head consists of multiple linear layers with ReLU activations.

$$\hat{y}_i^t = \sigma\left(C^t[T_s(x_i), T_p^t(x_i)]\right), \quad \text{where } \sigma \text{ is the sigmoid function.} \tag{2}$$

4.2 Task-Conditioned Expert Routing

Unlike soft MoE gating, our routing is *deterministic* and purely task-conditioned. During training, we maintain three different dataloaders (Event/Emotion/Nature). A sample drawn from the dataloader for task t is routed to both the corresponding private backbone T_p^t and the shared backbone T_s.

In each iteration, we process three mini-batches at a time (one per dataloader) in a round robin fashion, yielding three different task losses and their gradients. The losses are backpropagated only after the three mini-batches are passed through. This design ensures that: (i) each private backbone only specializes in its own supervision, (ii) the shared backbone is trained in all tasks, and (iii) any possible catastrophic forgetting [23] between datasets is avoided.

4.3 Total Loss

We adopt a binary cross-entropy (BCE) objective for all tasks. For a mini-batch of size B and sequence length N, the task loss is

$$\mathcal{L}_t = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} \mathrm{BCE}\big(y_{b,i}^t, \hat{y}_{b,i}^t\big), \qquad t \in \{e, m, n\}. \tag{3}$$

The model is then trained using the sum of these three individual task losses with equal weights.

$$\mathcal{L}_{total} = \sum_{t \in \{e,m,n\}} \lambda_t \mathcal{L}_t, \quad \text{where } \lambda_e = \lambda_m = \lambda_n = 1. \tag{4}$$

4.4 Inference Strategy

During inference, the segment embeddings of a video are fed to the model and processed by the shared backbone as well as each of the three private backbones. Subsequently, for each task, the shared backbone output and the corresponding private backbone output are fused and passed through the corresponding prediction head. As a result, multi-perspective highlight scores $\hat{y}_i^e, \hat{y}_i^m, \hat{y}_i^n$ are produced for each temporal segment.

5 Experiments

5.1 Datasets and Metrics

Pretraining. We pretrain on TRINITY, a unified multi-perspective highlight dataset consisting of three perspective groups: Event, Emotion, and Nature. These groups are derived by grouping videos according to highlight perspectives. TRINITY-Event contains **27,845** videos, TRINITY-Emotion contains **16,257** videos, and TRINITY-Nature contains **15,540** videos, corresponding to event-, emotion-, and nature-centric annotations, respectively. We adopt a joint multi-dataset training strategy across the three perspectives. Each video is segmented into fixed-length 5-second temporal segments following prior practice in [15]. Each segment is assigned a normalized highlight score within the range $[0, 1]$, reflecting the degree of saliency under the corresponding perspective.

Benchmarks and Evaluation Metrics. We assess the event-centric capability of our model by fine-tuning only the Event branch while freezing all other components. Performance is evaluated on YouTube Highlights [38] and Mr. HiSum [36]. Crucially, we adopt the official train/validation/test splits of Mr. HiSum to ensure fair comparability with prior work. Since part of our pretraining data is sourced from Mr. HiSum, all test videos are removed from the pretraining pool to prevent data leakage. Mr. HiSum is released under the CC BY 4.0 license, allowing modifications; to adhere to this, we will distribute only high-level features and our new annotations. For Mr. HiSum and TRINITY, we report $\text{mAP}_{\rho=50\%}$ and $\text{mAP}_{\rho=15\%}$, and for YouTube Highlights, we report the standard mAP metric. For emotion modeling, we fine-tune only the Emotion branch and freeze the remaining modules. We evaluate on VEATIC [27], a context-aware video emotion dataset with continuous valence and arousal annotations. Its character-centric and temporal affective supervision aligns well with our Emotion perspective. Crucially, the Emotion branch focuses on emotion-driven highlight localization under skewed web distributions rather than balanced recognition, using VEATIC

Table 2: Fine-tuned highlight detection results on Mr. HiSum, YouTube Highlights, and VEATIC. **Bold** indicates the best result and underlined values indicate the second best.

Mr. HiSum			**YouTube Highlights**							**VEATIC**		
Method	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$	Method	Dog	Gym	Park.	Skate.	Skii.	Surf.	Avg.	Method	SAGR↑	RMSE↓
SAVS-Net [1]	-	57.12	Temp.-Cue [43]	55.38	62.66	70.88	69.06	60.05	59.76	62.97	RB-SFP [28]	-	0.2423
SL-Module [41]	24.95	58.63	Mr. HiSum [36]	50.80	67.20	78.30	42.00	61.80	71.60	62.00	VEATIC [27]	0.7637	0.2410
PGL-SUM [17]	27.45	61.60	RASL [20]	64.20	75.80	70.90	45.60	66.50	66.70	65.10	VDFS [29]	0.804	0.182
HierSum [4]	32.60	63.80	C2 [46]	-	-	-	-	-	-	67.20	-	-	-
Aha [6]	32.66	64.19	AV-Recur. [10]	-	-	-	-	-	-	68.30	–	–	–
SummDiff [15]	33.83	65.44	SL-Module [41]	70.80	53.20	77.20	72.50	66.10	76.20	69.30	–	–	–
–	–	–	PLD-VHD [40]	**74.90**	70.20	77.90	57.50	70.70	79.00	73.00	–	–	–
[HTML]F2F2F2 **Ours**	**40.98**	**69.06**		74.87	**91.89**	**87.70**	**83.95**	**77.20**	**87.33**	**83.82**		**0.8185**	**0.1749**

as auxiliary evidence for temporally grounded affective cues. While common affective patterns are well-supported, addressing rare emotions and direct external validation remains future work. We report SAGR and RMSE following the official protocol.

5.2 Implementation Details

Our model is implemented in PyTorch. Segment-level features are extracted using a frozen CLIP ViT-B/32 encoder. Temporal backbones are Transformer encoder layers with hidden size 512, 8 attention heads, and depth 5, using 1D RoPE for temporal position encoding. For each perspective, the forward pass consists of 4 shared attention heads and 4 task-specific private heads, yielding 8 heads per branch. During TRINITY pretraining, one shared backbone and three perspective-specific private backbones (Event, Emotion, Nature) are jointly optimized with three task heads. Optimization uses AdamW, with early stopping if validation performance does not improve for more than three consecutive epochs across all three tasks. The learning rate, weight decay, batch size, and dropout ratio are set to $5e^{-5}$, 0.01, 16, and 0.2, respectively.

5.3 Quantitative Results

Performance on Public Benchmarks. We compare our method with recent state-of-the-art approaches on Mr.HiSum, YouTube Highlights, and VEATIC. Results are summarized in Table 2. On Mr.HiSum, our method achieves 40.98 $mAP_{\rho=15\%}$ and 69.06 $mAP_{\rho=50\%}$, outperforming the recent SummDiff [15] by $+$**7.15** and $+$**3.62** points, respectively. The gains are consistent under both strict and relaxed evaluation ratios. On YouTube Highlights, under the text-agnostic setting, we achieve the best average performance and leading results on most categories, reaching an average score of 83.82, exceeding the strongest prior baseline PLD-VHD [40] by $+$**10.82** points. Large improvements are observed in dynamic categories such as Gym, Skating, and Surfing, indicating strong cross-domain generalization. On VEATIC, our model obtains 0.8185 SAGR and 0.1749 RMSE, surpassing previous methods on both correlation and regression precision. This demonstrates that our framework generalizes beyond ranking-based metrics.

Table 3: State-of-the-art comparison under the three TRINITY highlight perspectives (Nature, Emotion, and Event), reflecting the multi-perspective formulation of TRINITY. **Bold** indicates the best result.

Method	TRINITY-Nature		TRINITY-Emotion		TRINITY-Event	
	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$
SL-Module [41]	38.68	62.23	50.55	67.57	31.46	62.71
PGL-SUM [17]	46.40	61.84	62.47	**73.81**	33.61	61.84
CSTA [9]	36.47	39.67	64.96	67.29	35.97	40.77
SumDiff [15]	35.68	59.63	37.17	61.38	26.21	57.45
Aha† [6]	38.69	40.02	21.02	28.29	32.66	64.19
Qwen3-VL-8B† [3]	24.89	56.11	22.99	53.57	17.92	52.46
Qwen3-VL-235B† [3]	27.91	58.29	26.42	56.90	18.10	52.82
[HTML]F2F2F2 **Ours**	**58.74**	**68.05**	**67.42**	71.32	**40.98**	**69.06**

† Zero-shot due to fine-tuning cost.

Across three heterogeneous benchmarks with distinct evaluation protocols, our method consistently establishes new state-of-the-art results, validating the effectiveness of our shared–private highlight modeling design.

Evaluation on TRINITY. We evaluate several representative highlight detection methods [9, 17, 41] under the three highlight perspectives defined in TRINITY (Nature, Emotion, and Event), as shown in Table 3. Our framework consistently achieves the best overall performance across the three perspectives under stricter evaluation ratios. Under the Nature perspective, our method attains 58.74 $mAP_{\rho=15\%}$ and 68.05 $mAP_{\rho=50\%}$, substantially outperforming prior methods and demonstrating a stronger capability in capturing scenery-driven saliency. Under the Emotion perspective, we obtain the best $mAP_{\rho=15\%}$ score of 67.42 while achieving competitive performance at $mAP_{\rho=50\%}$ (71.32). Under the Event perspective, our model also achieves the best performance under both metrics, reaching 40.98 $mAP_{\rho=15\%}$ and 69.06 $mAP_{\rho=50\%}$.

Notably, although all methods are trained on TRINITY, baseline architectures originally designed for single-definition saliency modeling exhibit limitations when handling heterogeneous highlight signals, particularly for the Nature and Emotion perspectives. In contrast, our shared–private architecture separates perspective-specific representations while maintaining a unified temporal backbone, enabling robust highlight localization across diverse saliency definitions.

5.4 Ablation Study

Architecture ablation. Table 4 analyzes the effect of architectural decomposition. Moving from single-task training (S1) to a shared-backbone multi-task model (S2) improves the average performance by +3.50 $mAP_{\rho=15\%}$, indicating that cross-task supervision benefits shared temporal representations. Using fully independent experts (S3) improves performance under $mAP_{\rho=50\%}$ compared with shared-backbone MTL (S2), indicating improved ranking when broader temporal coverage is evaluated, as the relaxed threshold ($\rho = 50\%$) rewards highlights

Table 4: Architecture ablation on TRINITY. We analyze the impact of shared backbones, private experts, and multi-task learning. Results are reported using $\text{mAP}_{\rho=15\%}$ and $\text{mAP}_{\rho=50\%}$ (denoted as 15% and 50% in the header).

Setting	Arch. Config.					Event		Nature		Emotion		Avg.	
	#B.B.	Sha.	Pri.	#Head	MTL	15%	50%	15%	50%	15%	50%	15%	50%
S1: Single-Task Baseline	1	✓	×	1	×	32.85	63.06	43.16	65.57	61.05	75.56	45.69	68.06
S2: Shared-Backbone MTL	1	✓	×	3	✓	37.97	66.86	44.41	64.35	65.20	75.71	49.19	68.97
S3: Independent Experts	3	×	×	1	×	37.54	67.51	44.78	65.26	63.27	76.34	48.53	69.70
S4: Shared–Private Experts	4	✓	✓	3	✓	40.98	69.06	58.74	68.05	67.42	71.32	55.71	69.48

Table 5: Dataset contribution ablation under a fixed architecture (S4: Shared–Private Experts). Models are trained with progressively richer supervision.

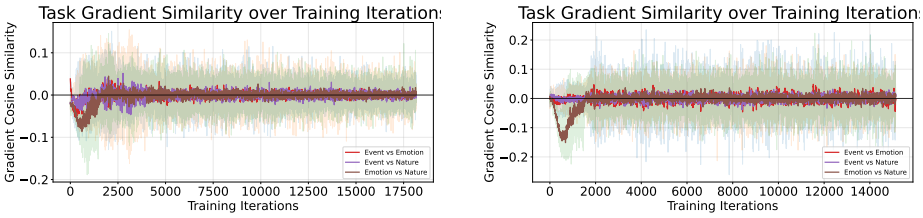
Setting	Event		Nature		Emotion		Avg.	
	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$	$mAP_{\rho=15\%}$	$mAP_{\rho=50\%}$
D1: Event	40.25	69.45	31.25	65.58	40.83	65.76	37.44	66.93
D2: Event + Nature	39.98 ↓0.27	68.70 ↓0.75	57.89 ↑26.64	67.71 ↑2.13	45.98	61.06	48.25	64.32
D3: Full (Event + Nature + Emotion)	40.98 ↑1.00	69.06 ↑0.36	58.74 ↑0.85	68.05 ↑0.34	67.42 ↑21.44	71.32 ↑10.26	55.71	69.48

that capture a larger portion of the ground-truth segments. However, the absence of shared representations limits overall cross-perspective generalization. Our shared–private expert design (S4) achieves the best overall performance (55.71 / 69.48). Relative to the independent-expert setting (S3), S4 improves Event from 37.54 to 40.98, Nature from 44.78 to 58.74, and Emotion from 63.27 to 67.42 under $\text{mAP}_{\rho=15\%}$, demonstrating the benefit of combining shared representations with perspective-specific experts. Consistent with this observation, as illustrated in Fig. 3, the t-SNE visualization shows clearer perspective-wise clustering of temporal segment features under S4, indicating more structured temporal representations across perspectives.

Dataset contribution ablation. Table 5 examines the impact of progressively incorporating multi-perspective supervision under a fixed architecture (S4). Training with Event-only data (D1) yields moderate performance across all three dimensions. Adding Nature supervision (D2) substantially improves Nature performance from 31.25 to 57.89 under $\text{mAP}_{\rho=15\%}$, while Event performance remains largely stable, indicating limited negative interference across different perspectives. Finally, incorporating all three perspectives (D3) provides consistent improvements across dimensions, with the largest gain on Emotion The full multi-perspective training achieves the highest overall average performance (55.71 / 69.48), validating the benefit of jointly modeling heterogeneous highlight cues.

5.5 Gradient Conflict Observation

A notable problem in multi-task learning is gradient conflict between the losses of different tasks. Conflicting gradients were shown to have negative transfer effects [18] and there have been multiple attempts to mitigate this effect [5,33,44]. Since



(a) With Aligned-MTL **(b)** Without Aligned-MTL

Fig. 4: Comparison of gradient cosine similarity during training with and without Aligned-MTL.

our problem is also formulated as a multi-task learning problem, we performed gradient conflict analysis to ensure there were no adversarial effects. In order to measure the gradient conflict, we employed cosine similarity of gradients [44] between our three different task losses.

The gradient cosine similarity of our model can be seen in Fig. 4b where the $\cos(\theta_{i,j}) < 0$ for the early parts of training, after which it stabilizes to near 0, indicating limited gradient interference between tasks. Even after employing Aligned-MTL [33], as shown in Fig. 4a, only smaller fluctuations in similarity values are observed at the beginning, while the final stabilization near zero remains unchanged, suggesting that our model suffers negligible gradient conflicts.

5.6 Attention Visualization

We demonstrate our model's capacity to capture multiple semantic cues from a single video by visualizing its temporal attention mechanism. Unlike spatial attention in standard Vision Transformers, our model processes temporal frames as tokens. Consequently, its attention mechanism reveals the specific temporal segments prioritized for different tasks.

As Figure 5 shows, clear disentanglement between the three heads is observed, suggesting the model learns distinct, task-specific representations from the shared video backbone. We validate this semantic grounding by examining frames where attention values peaked. The model effectively localizes facial expressions for emotion, physical transitions for events, and landscape vistas for nature. Notably, the low cross-correlation between these head-specific distributions indicates that the model appears to partition the latent space to minimize task interference. Furthermore, the mechanism acts as a temporal filter, assigning low attention weights to uninformative segments, such as motion blur or camera resets, to focus purely on causally relevant evidence.

6 Conclusion

We revisit video highlight detection under realistic personal-style settings and identify the limitations of event-centric formulations. To address this gap, we introduce TRINITY, a multi-perspective benchmark that decomposes highlight

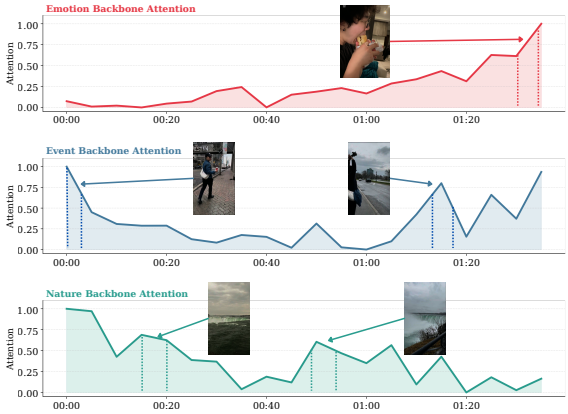


Fig. 5: Attention visualization of a personal vlog video. The visualization shows temporal localization across different semantic perspectives. The Emotion backbone highlights human-centric contentment, the Event backbone captures conversational interaction segments, and the Nature backbone focuses on environmental landmarks (Niagara Falls), illustrating the specialization of perspective-specific backbones.

saliency into event, emotion, and nature dimensions. We further propose a shared-backbone multi-expert architecture to model perspective-specific saliency. Extensive experiments show strong performance on TRINITY and consistent gains on existing benchmarks, demonstrating improved generalization across heterogeneous highlight definitions. We hope this work encourages more realistic and multi-dimensional formulations of highlight detection.

References

1. Alharbi, F., Habib, S., Albattah, W., Jan, Z., Alanazi, M.D., Islam, M.: Effective video summarization using channel attention-assisted encoder–decoder framework. Symmetry **16**(6), 680 (2024)
2. Badamdorj, T., Rochan, M., Wang, Y., Cheng, L.: Joint visual and audio learning for video highlight detection. In: ICCV. pp. 8127–8137 (2021)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL technical report (2025)
4. Beedu, A., Essa, I.: HierSum: A global and local attention mechanism for video summarization. arXiv preprint arXiv:2504.18689 (2025)
5. Cha, C., You, T., Yeh, R.A., Vitelli, S., Wang, S.H.: Conflict-averse gradient descent for multi-task learning. In: Advances in Neural Information Processing Systems. vol. 34, pp. 18803–18816 (2021)

6. Chang, A., De Melo, C., Lukin, S.M.: Aha! - predicting what matters next: Online highlight detection without looking ahead. In: NeurIPS (2025)
7. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4D: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022)
8. Gygli, M., Song, Y., Cao, L.: Video2GIF: Automatic generation of animated GIFs from video. In: CVPR. pp. 1001–1009 (2016)
9. Hong, S.E., Choi, G.M., Kim, J.H.: Csta: Cnn-based spatiotemporal attention for video summarization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18859–18868 (2024)
10. Islam, Z., Paul, S., Rochan, M.: Unsupervised video highlight detection by learning from audio and visual recurrence. In: Proc. WACV. pp. 8702–8711 (2025)
11. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: ICCV. pp. 13846–13856 (2023)
12. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023)
13. Kahneman, D., Fredrickson, B.L., Schreiber, C.A., Redelmeier, D.A.: When more pain is preferred to less: Adding a better end. Psychological Science **4**(6), 401–405 (1993)
14. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: ICCV (2021)
15. Kim, K., Hahm, J., Kim, S., Sul, J., Kim, B., Lee, J.: SummDiff: Generative modeling of video summarization with diffusion. In: ICCV. pp. 15096–15106 (2025)
16. Kirk, D., Sellen, A., Harper, R., Wood, K.: Understanding videowork. In: Proc. CHI. pp. 61–70 (2007)
17. Koutras, P., Maragos, P.: Combining global and local attention with positional encoding for video summarization. In: 2021 IEEE International Symposium on Multimedia (ISM). pp. 41–48 (2021)
18. Lee, H.B., Yang, E., Hwang, S.J.: Deep asymmetric multi-task feature learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2956–2964. PMLR (2018)
19. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. NeurIPS **34**, 11846–11858 (2021)
20. Li, T., Sun, Z., Xiao, X.: Unsupervised modality-transferable video highlight detection with representation activation sequence learning. IEEE TIP **33**, 1911–1922 (2024)
21. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: UniVTG: Towards unified video-language temporal grounding. In: ICCV. pp. 2794–2804 (2023)
22. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: CVPR. pp. 3042–3051 (2022)
23. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of Learning and Motivation, vol. 24, pp. 109–165. Academic Press (1989)
24. Garcia del Molino, A., Gygli, M.: PHD-GIFs: Personalized highlight detection for automatic GIF creation. In: ACM MM. pp. 600–608 (2018)
25. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: CVPR. pp. 23023–23033 (2023)

26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
27. Ren, Z., Ortega, J., Wang, Y., Chen, Z., Guo, Y., Yu, S.X., Whitney, D.: VEATIC: Video-based emotion and affect tracking in context dataset. In: Proc. WACV. pp. 4467–4477 (2024)
28. Ren, Z., Wang, Y., Ke, T.W., Guo, Y., Yu, S.X., Whitney, D.: Region-based emotion recognition via superpixel feature pooling. In: CVPRW (2024)
29. Saigo, S., Hayami, T., Watanabe, H.: Enhancing continuous emotion recognition via visually diverse frame selection. In: Proc. IEEE Global Conference on Consumer Electronics (GCCE). pp. 1275–1278 (2025)
30. Savchenko, A.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: ICML. pp. 30119–30129 (2023)
31. Savchenko, A.V., Savchenko, L.V., Makarov, I.: Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Trans. Affect. Comput. (2022)
32. Sellen, A.J., Fogg, A., Aitken, M., Hodges, S., Rother, C., Wood, K.: Do life-logging technologies support memory for the past? An experimental study using SenseCam. In: Proc. CHI. pp. 81–90 (2007)
33. Senushkin, D., Belikov, N., Galimaldinov, A., Konushin, A.: Independent component alignment for multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20083–20093 (2023)
34. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: TvSum: Summarizing web videos using titles. In: CVPR. pp. 5179–5187 (2015)
35. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864 (2021)
36. Sul, J., Han, J., Lee, J.: Mr. HiSum: A large-scale dataset for video highlight detection and summarization. In: NeurIPS (2023)
37. Sun, H., Zhou, M., Chen, W., Xie, W.: TR-DETR: Task-reciprocal transformer for joint moment retrieval and highlight detection. In: AAAI. vol. 38, pp. 4998–5007 (2024)
38. Sun, M., Farhadi, A., Seitz, S.: Ranking domain-specific highlights by analyzing edited videos. In: ECCV. pp. 787–802 (2014)
39. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
40. Wei, F., Wang, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Learning pixel-level distinctions for video highlight detection. In: CVPR. pp. 3073–3082 (2022)
41. Xu, M., Wang, H., Ni, B., Zhu, R., Sun, Z., Wang, C.: Cross-category video highlight detection via set-based learning. In: ICCV. pp. 7970–7979 (2021)
42. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: CVPR. pp. 982–990 (2016)
43. Ye, Q., Shen, X., Gao, Y., Wang, Z., Bi, Q., Li, P., Yang, G.: Temporal cue guided video highlight detection with low-rank audio-visual fusion. In: ICCV. pp. 7950–7959 (2021)
44. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 5824–5836 (2020)

45. Zeng, X., Yang, Y., Zeng, P., Yin, W., Liu, B., Wu, X., Wang, Y.: Promptemo: Learning emotion with bilateral textual prompts in multi-domain open-set scenarios. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12322–12330 (2026)
46. Zhao, C., Zhao, Z., Zhao, X.: Weakly-supervised video highlight detection by characteristic and commonality modeling. In: ICASSP. pp. 1–5 (2025)