SPECIALIZE BEFORE FUSION: MODALITY-SPECIFIC ROUTING FOR MULTIMODAL VIDEO HIGHLIGHT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Highlight detection and video summarization require identifying salient moments from heterogeneous videos where visual and acoustic evidence can vary over time. Existing multimodal methods mainly focus on how modalities are fused, often before adequately modeling their distinct temporal characteristics. We introduce a specialize-before-fusion framework that first learns independently parameterized visual and audio temporal representations and then adaptively combines them using time-varying routing weights, optionally conditioned on a semantic query. Across four benchmarks spanning long-form sports broadcasts and web video, our method achieves state-of-the-art performance, improving over the strongest baseline by up to 6.9 mAP on long-form sports highlight detection while incurring lower inference cost. Ablations consistently favor modality-specific over shared temporal modeling and adaptive over fixed fusion, and representation analyses show visual and audio branches become more distinct specifically near salient moments. These results support specializing heterogeneous modalities before multimodal integration.

1 INTRODUCTION

With recent advances in recording technologies and online media platforms, video content has grown rapidly across a wide range of domains and durations, from short-form videos to YouTube content Kim et al. (2026) and long-form sports broadcasts Lee et al. (2026b). Viewers increasingly favor concise and engaging content, such as sports highlights and video recaps, rather than consuming full-length videos Guan (2024); Violot et al. (2024). This growing demand has made Highlight Detection (HD) and Video Summarization (VS) increasingly important for efficiently navigating large-scale video content. HD and VS aim to identify salient temporal segments that summarize essential content or capture the most engaging moments. Despite substantial progress, learning a generalizable notion of highlight relevance remains challenging across diverse real-world videos Sul et al. (2023). Salient moments can be characterized by very different cues across domains such as sports, movies, and everyday activities, requiring models to reason beyond visually distinctive or temporally sparse events.

Existing approaches have predominantly relied on visual information, while audio is either omitted Lin et al. (2023); Yang et al. (2024); Xu et al. (2024); Zhao et al. (2024); Lei et al. (2021) or incorporated through relatively fixed fusion mechanisms such as concatenation, cross-attention, or pooling Ma et al. (2025); Moon et al. (2023); Gordeev et al. (2026); Sun et al. (2024); Badamdorj et al. (2021); Islam et al. (2026); Ye et al. (2021). However, audio can provide complementary evidence that is difficult to infer from visual content alone Ge et al. (2023); Xu et al. (2020). This is particularly evident in sports, where commentary, crowd reactions, whistles, and impact sounds can strongly indicate salient moments Xu et al. (2026).

More recently, TripleSumm Kim et al. (2026) demonstrated the value of dynamically integrating visual, textual, and audio representations as their relevance varies over time. Beyond determining how much each modality contributes, an important design choice is how its temporal representations are formed before fusion. In sports broadcasts, visual actions, commentary, and crowd reactions provide complementary evidence that can unfold at different temporal scales. Their relative importance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

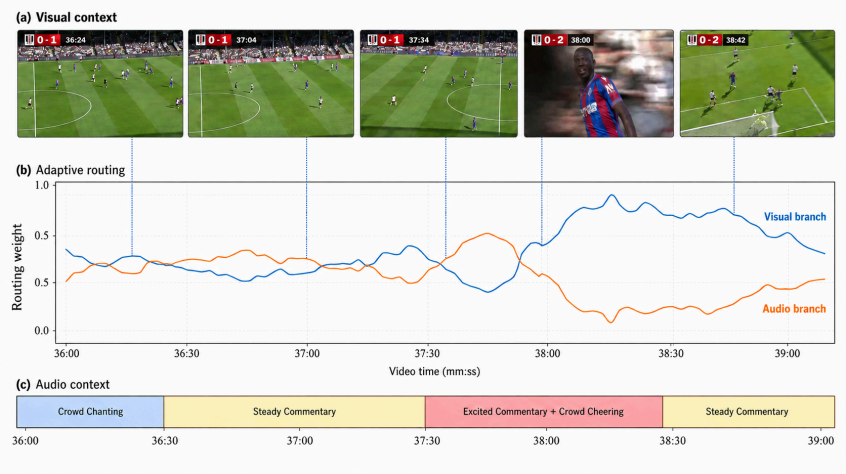


Figure 1: Adaptive multimodal routing over a soccer video. The relative contributions of visual and audio evidence vary over time, motivating modality-specific temporal modeling before adaptive fusion.

also varies throughout a video, as illustrated in Figure 1. This motivates our *specialize-before-fusion* approach: independently model the temporal characteristics of visual and audio streams, then adaptively combine their representations. Unlike the shared temporal processing used in TripleSumm, our modality-specific branches allow each stream to learn temporal transformations tailored to its own signal characteristics.

Motivated by this observation, we introduce a *specialize-before-fusion* formulation for multimodal highlight understanding. Our framework first models visual and audio streams using independently parameterized modality-specific temporal branches, allowing each modality to preserve temporal patterns suited to its own signal characteristics. The architecture further supports multiple branches within each modality to increase specialization capacity when richer representations are available. The resulting representations are conditioned on a semantic query and subsequently integrated through a lightweight adaptive routing mechanism that assigns time-varying contributions to each branch. In this way, our framework explicitly separates *representation specialization* from *multimodal integration*.

We evaluate the proposed framework on four diverse highlight detection and video summarization benchmarks spanning long-form sports broadcasts and shorter web videos. The method achieves strong performance across all four benchmarks while maintaining moderate computational cost. Extensive experiments directly examine the specialize-before-fusion hypothesis through shared versus modality-specific temporal modeling, branch capacity, adaptive versus fixed routing, temporal modeling choices, missing-modality robustness, and representation-level analyses.

Our main contributions are summarized as follows:

- We introduce a **specialize-before-fusion** formulation for multimodal highlight understanding, in which heterogeneous visual and audio streams are first modeled using independently parameterized temporal transformations before multimodal integration.
- We instantiate this principle with modality-specific temporal branches and adaptive routing. The architecture supports multiple independently parameterized branches within each modality to increase specialization capacity, while dynamically determining their contributions at each temporal location.
- We validate the proposed formulation across four highlight detection and video summarization benchmarks. Extensive ablations demonstrate consistent benefits from modality-specific over shared temporal processing and adaptive over fixed fusion, while further analyses examine specialization capacity, modality robustness, and learned representation structure.

2 Related Work

2.1 Highlight Detection

Highlight detection (HD) and moment retrieval (MR) are increasingly studied together in temporal video understanding. MR aims to localize a temporal moment in a video given a natural-language query, whereas HD focuses on estimating the saliency or importance of individual video clips Sun et al. (2024). Their joint formulation was popularized by QVHighlights Lei et al. (2021), which introduced Moment-DETR, a DETR-based framework that retrieves query-relevant moments while simultaneously predicting clip-level highlight scores.

Following this formulation, many recent approaches have adopted DETR-style architectures Carion et al. (2020) for joint MR&HD. QD-DETR Moon et al. (2023) introduces a query-dependent video representation module to make temporal localization more explicitly conditioned on the input query. MH-DETR Xu et al. (2024) incorporates pooling operations into the encoder together with cross-modal interaction modules for joint MR&HD. TR-DETR Sun et al. (2024) further argues that previous methods primarily improve multimodal feature discrimination and interaction while overlooking the reciprocal relationship between MR and HD, and therefore introduces a task-reciprocal Transformer to exploit their mutual benefits. MS-DETR Ma et al. (2025) proposes a motion-semantic disentangled encoder that separately models temporal motion and spatial semantic information before interacting with the text query. More recently, SG-DETR Gordeev et al. (2026) introduces saliency-guided attention, where locally estimated saliency information is used to guide cross-modal interaction.

Despite these advances, most existing MR&HD methods remain predominantly vision-centric, with audio either omitted or incorporated using relatively simple fusion strategies. This motivates a more explicit investigation of how complementary audio and visual evidence should be modeled and integrated for highlight prediction.

2.2 Video Summarization

Video summarization (VS) is closely related to HD, but focuses on selecting a compact set of segments that preserves the most important or representative content of a video. Early deep-learning approaches predominantly relied on visual representations and modeled temporal dependencies using recurrent architectures such as RNNs and LSTMs Zhang et al. (2016); Zhao et al. (2017). Subsequent methods increasingly adopted attention-based temporal modeling, including self-attention and Transformer-style architectures Zhu et al. (2020); Jiang & Mu (2022); Terbouche et al. (2023); Son et al. (2024). Other directions have explored graph-based formulations Zhang et al. (2023); Zhao et al. (2021) and, more recently, diffusion-based models for learning summary representations Yu et al. (2024); Shang et al. (2025); Kim et al. (2025).

As video understanding has become increasingly multimodal, recent work has begun incorporating additional semantic signals beyond visual features. Some methods introduce textual information to improve semantic understanding and summary selection Li et al. (2023); Argaw et al. (2024); Guo et al. (2025), while others additionally exploit audio signals to capture complementary acoustic cues Badamdorj et al. (2021); Liu et al. (2022); Zhao et al. (2022); Xie et al. (2022). These studies demonstrate that non-visual modalities can provide useful information that is not fully captured by appearance and motion alone.

Most recently, TripleSumm Kim et al. (2026) models the time-varying importance of different modalities and dynamically integrates visual, textual, and audio representations at each temporal location. It combines shared-parameter temporal processing with fusion-token cross-attention, enabling content-dependent multimodal aggregation. Our work focuses on the temporal representations underlying this aggregation. We employ independently parameterized visual and audio temporal branches to learn modality-specific temporal patterns, followed by query conditioning and adaptive routing. While TripleSumm reports benefits from parameter sharing within its architecture, we investigate modality-specific temporal processing across sports highlight detection and general video summarization benchmarks.

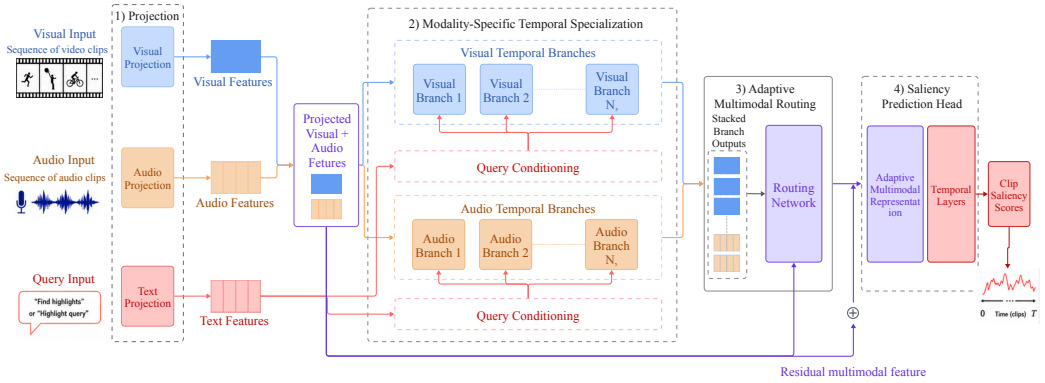


Figure 2: Overview of the proposed framework. Pre-extracted visual, audio, and textual features are first projected into a common representation space. Visual and audio streams are then independently processed by modality-specific temporal modeling and conditioned on the textual query. The resulting query-aware modality representations are dynamically integrated by the adaptive multimodal routing module, which assigns time-varying contributions to each modality. The fused representation is subsequently processed by temporal layers to predict clip-level saliency scores.

3 METHODS

3.1 PROBLEM FORMULATION

Given an untrimmed video V consisting of N_v temporal clips, each clip is associated with corresponding visual and audio signals. Following standard practice in highlight detection and video summarization, the visual and audio inputs are first processed using modality-specific pretrained feature extractors. Our formulation supports one or multiple pre-extracted feature representations for each modality, allowing complementary features from different pretrained encoders to be incorporated.

Formally, let $X^v = \{x_i^v\}_{i=1}^{N_v}$ and $X^a = \{x_i^a\}_{i=1}^{N_v}$ denote the resulting visual and audio feature sequences, respectively. We additionally introduce a generic textual query Q, manually specified before training to describe the target notion of highlight relevance. A pretrained text encoder transforms Q into token-level text features $X^q \in \mathbb{R}^{N_q \times d_q}$, where N_q is the number of query tokens and d_q is the text feature dimension.

Given X^v, X^a, and X^q, the goal is to predict clip-level saliency scores $S = \{s_i\}_{i=1}^{N_v}$, where $s_i \in [0, 1]$ denotes the predicted relevance of the i-th clip.

An overview of the proposed framework is shown in Figure 2. The model first projects the pre-extracted modality features into a common representation space, performs modality-specific temporal modeling, conditions the resulting representations on the textual query, and integrates them through adaptive multimodal routing before predicting clip-level saliency scores.

3.2 MODALITY-SPECIFIC TEMPORAL SPECIALIZATION

Our central design principle is to specialize temporal representations within each modality before multimodal fusion. Visual and audio signals provide different forms of evidence for highlight prediction and can exhibit distinct temporal structures. Rather than forcing heterogeneous modalities through a shared temporal transformation, we therefore model them using independently parameterized temporal branches. The framework also supports multiple independently parameterized temporal branches within each modality, providing additional modeling capacity whose benefit we examine empirically through branch-capacity ablations.

Given X^v and X^a, we first project the visual and audio features into a common hidden dimension d using separate learnable projections, $H_0^v = \phi_v(X^v)$ and $H_0^a = \phi_a(X^a)$, where $H_0^v, H_0^a \in \mathbb{R}^{N_v \times d}$.

We denote the representation at temporal location i by h_i^m, such that $H_0^m = \{h_i^m\}_{i=1}^{N_v}$ for $m \in \{v, a\}$.

For each modality m, we employ K_m independently parameterized temporal branches, $\mathcal{T}^m = \{\mathcal{T}_1^m, \ldots, \mathcal{T}_{K_m}^m\}$. Each branch is implemented using a bidirectional GRU over the projected feature sequence, producing $H_k^m = \mathcal{T}_k^m(H_0^m)$ for $k = 1, \ldots, K_m$. Crucially, temporal parameters are not shared between the visual and audio streams. This allows each modality to learn temporal transformations tailored to its own signal characteristics. Multiple branches within the same modality further increase specialization capacity by learning different transformations of the same modality-level input.

We additionally condition each branch representation on the semantic query through multi-head cross-attention. For modality m and branch k, H_k^m serves as the attention query, while the token-level text representation X^q provides the keys and values. The resulting representation is $Z_k^m = \mathcal{C}^m(H_k^m, X^q)$, where $\mathcal{C}^m(\cdot)$ denotes the modality-specific query-conditioning module. Query-conditioning parameters are shared across branches within the same modality, while separate modules are used for visual and audio streams.

The resulting specialized branch sets, $\mathcal{Z}^v = \{Z_1^v, \ldots, Z_{K_v}^v\}$ and $\mathcal{Z}^a = \{Z_1^a, \ldots, Z_{K_a}^a\}$, are subsequently passed to the adaptive routing module. In this way, multimodal integration is performed only after modality-specific temporal specialization.

3.3 Adaptive Multimodal Routing

The relevance of visual and audio evidence can vary across temporal locations, while different branches may capture complementary patterns within each modality. We therefore employ an adaptive routing mechanism that dynamically determines the contribution of each query-conditioned branch.

We combine the visual and audio branch sets into $K = K_v + K_a$ total branches and let $z_{i,k}$ denote the output of the k-th query-conditioned branch at temporal location i. To determine the routing weights, we construct a context representation from three sources: the concatenated projected visual and audio features $u_i = [h_i^v; h_i^a]$, a global textual representation q obtained by masked mean pooling over X^q, and the mean branch representation $\bar{z}_i$ computed across the available query-conditioned branches. The routing input is therefore $r_i = [u_i; q; \bar{z}_i]$.

A lightweight routing network $\mathcal{R}(\cdot)$ produces one logit for each branch as $g_i = \mathcal{R}(r_i)$. After masking unavailable branches, a temperature-scaled softmax produces normalized routing weights $\alpha_{i,k}$. The routed representation is

$$z_i^{\text{route}} = \sum_{k=1}^{K} \alpha_{i,k} z_{i,k}.$$

To preserve the original modality information, we additionally introduce a residual pathway from the projected audio-visual representation. The final multimodal representation is $z_i = \text{LN}(z_i^{\text{route}} + \phi_r(u_i))$, where $\phi_r(\cdot)$ maps the concatenated projected features to the branch output dimension and $\text{LN}(\cdot)$ denotes layer normalization.

Finally, the fused sequence $Z = \{z_i\}_{i=1}^{N_v}$ is processed by a temporal prediction module followed by a saliency head to obtain the final scores $S = \{s_i\}_{i=1}^{N_v}$.

4 Experiments

Datasets We evaluate our method on four benchmarks spanning long-form sports highlight detection and shorter web-video summarization. **SV-Highlights** Lee et al. (2026a) contains 320 long-form videos across eight sports, with an average duration of approximately two hours. We use the provided CLIP Radford et al. (2021) and SlowFast Feichtenhofer et al. (2019) visual features and additionally extract audio features using pretrained AST Gong et al. (2021) and CLAP Elizalde et al. (2023). **SoccerHigh** Díaz-Juan et al. (2025), derived from SoccerNet Giancola et al. (2018), contains 237 paired full-match soccer videos. We extract CLIP and SlowFast visual features and corresponding pretrained audio representations using the same setup. For video summarization, we evaluate

Table 1: Performance comparison on SoccerHigh and SV-Highlights using CLIP+SlowFast visual features and AST audio features. Uni-VTG is evaluated as a visual-textual-only baseline because it does not incorporate an audio modality; all remaining methods use both visual and audio inputs. Efficiency is measured by the number of trainable parameters and GFLOPs. Best results are shown in bold and second-best results are underlined.

	SoccerHigh				SV-Highlights				Efficiency	
Method	mAP ↑	H@1 ↑	H@K ↑	IoU ↑	mAP ↑	H@1 ↑	H@K ↑	IoU ↑	Params. ↓	GFLOPs ↓
Uni-VTG	11.98 ± 1.76	17.41 ± 6.12	13.94 ± 2.53	7.73 ± 1.55	23.06 ± 0.93	38.21 ± 3.46	27.43 ± 1.38	16.32 ± 0.85	51.79	20.35
MS-DETR	34.47 ± 1.03	73.33 ± 1.11	36.42 ± 0.99	22.81 ± 0.73	37.89 ± 1.89	60.25 ± 7.10	40.61 ± 1.22	26.24 ± 0.93	7.78	1.86
JAV	36.91 ± 0.33	72.96 ± 2.32	37.68 ± 0.72	24.25 ± 0.55	32.81 ± 1.59	70.68 ± 2.45	35.57 ± 1.18	22.13 ± 0.87	<u>1.78</u>	**1.18**
QD-DETR	42.93 ± 0.65	81.11 ± 1.11	41.39 ± 0.54	26.75 ± 0.50	37.69 ± 0.77	63.93 ± 7.88	40.47 ± 0.37	26.08 ± 0.25	7.96	4.36
TR-DETR	43.44 ± 0.60	**86.67 ± 2.23**	41.57 ± 0.28	27.08 ± 0.23	39.09 ± 1.32	74.33 ± 8.96	40.33 ± 1.15	25.90 ± 0.90	8.56	3.44
UMT	50.28 ± 0.33	83.70 ± 1.28	45.90 ± 0.40	31.40 ± 0.41	28.61 ± 3.07	62.90 ± 11.62	31.51 ± 2.68	19.10 ± 1.90	11.32	5.45
TripleSumm	51.77 ± 0.50	<u>85.19 ± 1.70</u>	47.87 ± 0.33	33.26 ± 0.36	45.75 ± 0.93	84.28 ± 6.88	45.13 ± 0.50	29.98 ± 0.50	**1.61**	1.93
Ours (1V+1A)	<u>54.31 ± 0.77</u>	<u>85.19 ± 0.65</u>	<u>50.00 ± 0.89</u>	<u>35.35 ± 0.74</u>	<u>52.63 ± 0.83</u>	<u>89.03 ± 4.63</u>	<u>50.31 ± 0.90</u>	**34.68 ± 0.76**	6.49	<u>1.56</u>
Ours (2V+2A)	**54.58 ± 0.44**	<u>85.19 ± 0.65</u>	**50.13 ± 0.47**	**35.49 ± 0.34**	**53.01 ± 1.32**	**90.06 ± 3.92**	**50.33 ± 0.64**	<u>34.59 ± 0.61</u>	8.60	2.51

on **Mr.HiSum** Sul et al. (2023), originally comprising 31,892 YouTube videos, and **MoSu** Kim et al. (2026), containing 52,678 videos across 3,406 categories. For both datasets, we follow the multimodal setup of TripleSumm Kim et al. (2026) and use its released pre-extracted visual, audio, and textual representations. The textual representations are derived from each video's English captions and serve as video-specific semantic queries for conditioning the multimodal representations. This differs from the sports benchmarks, where queries describe general highlight relevance rather than the content of an individual video. Additional dataset statistics, feature extraction settings, and preprocessing details are provided in Appendix A.2.

Evaluation Metrics For SV-Highlights and SoccerHigh, we follow the evaluation protocol of SV-Highlights Lee et al. (2026a) and report mean Average Precision (mAP), Hit@1, Hit@K, and temporal Intersection-over-Union (IoU). mAP measures the overall ranking quality of clip-level saliency predictions, while Hit@1 evaluates whether the highest-scoring predicted clip corresponds to a ground-truth highlight. For Hit@K, K is defined independently for each video as the number of ground-truth highlight clips in that video. Specifically, $\text{Hit@K} = \frac{|\mathcal{P}_K \cap \mathcal{G}|}{|\mathcal{G}|}, \quad K = |\mathcal{G}|$. where $\mathcal{P}_K$ denotes the set of the top-K highest-scoring predicted clips and $\mathcal{G}$ denotes the set of ground-truth highlight clips. Hit@K therefore measures the proportion of ground-truth highlight clips recovered when the number of selected predictions is matched to the number of annotated highlights. IoU measures the temporal overlap between predicted and ground-truth highlight regions. For Mr.HiSum and MoSu, we follow the evaluation protocol used in TripleSumm Kim et al. (2026). We report Kendall's Tau (KTau) and Spearman's rank correlation coefficient (SRho) to measure the agreement between predicted and ground-truth frame-level importance scores. In addition, highlight detection performance is evaluated using the most-replayed statistics as ground-truth saliency signals. Following the same protocol, each video is divided into non-overlapping 5-second segments, which are ranked according to the predicted importance scores. We then report mAP@50 and mAP@15, corresponding to evaluation using the top 50% and top 15% ranked segments, respectively.

4.1 Implementation Details

All models are implemented in PyTorch using the pre-extracted modality features described above. Following prior work, our default feature configuration uses CLIP and SlowFast for the visual stream and AST for the audio stream. Unless otherwise specified, we use one visual and one audio temporal branch (1V+1A) as the default configuration due to its favorable performance–efficiency trade-off. We additionally report a higher-capacity 2V+2A variant to examine the effect of increased specialization capacity. Alternative feature and branch configurations are analyzed in Appendix A.4.3. We train for 30 epochs using AdamW Loshchilov & Hutter (2017) with an initial learning rate of 1×10^{-4} and a cosine schedule. The model predicts a highlight relevance logit for each temporal clip and is optimized with weighted binary cross-entropy, using weights of 5 for positive clips and 1 for negative clips to address class imbalance. For fair comparison, baselines are retrained using their official implementations whenever available, with matched data splits, feature inputs, and evaluation protocols where supported. All experiments are repeated with three random seeds (42, 2026, and

Table 2: Performance comparison on Mr.HiSum and MoSu. Best results are shown in bold and second-best results are underlined.

Dataset	Method	KTau ↑	SRho ↑	mAP@50 ↑	mAP@15 ↑
Mr.HiSum	MS-DETR	0.2322 ± 0.0005	0.3196 ± 0.0004	69.74 ± 0.19	39.27 ± 0.78
	TR-DETR	0.1712 ± 0.0128	0.2398 ± 0.0179	67.10 ± 0.54	36.34 ± 1.91
	JAV	0.1274 ± 0.0014	0.1846 ± 0.0020	63.96 ± 0.12	32.55 ± 0.25
	UMT	0.2482 ± 0.0013	0.3402 ± 0.0012	70.51 ± 0.22	40.65 ± 0.44
	TripleSumm	0.2473 ± 0.0021	0.3394 ± 0.0025	70.33 ± 0.08	40.42 ± 0.51
	Ours (1V+1A)	$\underline{0.2576 \pm 0.0043}$	$\underline{0.3456 \pm 0.0055}$	$\underline{70.66 \pm 0.20}$	$\underline{41.05 \pm 0.07}$
	Ours (2V+2A)	$\mathbf{0.2596 \pm 0.0052}$	$\mathbf{0.3487 \pm 0.0076}$	$\mathbf{70.71 \pm 0.40}$	$\mathbf{41.30 \pm 0.39}$
MoSu	MS-DETR	0.3167 ± 0.0050	0.4255 ± 0.0051	73.34 ± 0.33	41.52 ± 0.62
	TR-DETR	0.2549 ± 0.0231	0.3496 ± 0.0291	70.16 ± 1.70	35.87 ± 3.78
	QD-DETR	0.2404 ± 0.0207	0.3359 ± 0.0265	69.08 ± 1.27	36.00 ± 1.80
	UMT	0.3194 ± 0.0045	0.4319 ± 0.0065	73.27 ± 0.24	42.14 ± 0.22
	TripleSumm	$\underline{0.3491 \pm 0.0008}$	0.4689 ± 0.0012	74.47 ± 0.11	44.34 ± 0.11
	Ours (1V+1A)	$\mathbf{0.3548 \pm 0.0003}$	$\underline{0.4698 \pm 0.0008}$	$\mathbf{74.82 \pm 0.04}$	$\underline{44.77 \pm 0.08}$
	Ours (2V+2A)	$\mathbf{0.3548 \pm 0.0026}$	$\mathbf{0.4700 \pm 0.0029}$	$\underline{74.78 \pm 0.18}$	$\mathbf{44.86 \pm 0.33}$

3020), and we report mean and standard deviation. Further implementation details are provided in Appendix A.1.

4.2 Results

Long-form Sports Highlight Detection. Table 1 reports the results on SoccerHigh and SV-Highlights. Both configurations of our method outperform the strongest competing approaches on most metrics. On SoccerHigh, the default 1V+1A model achieves 54.31 mAP, 50.00 Hit@K, and 35.35 IoU, outperforming TripleSumm by 2.54, 2.13, and 2.09 points, respectively. Increasing the specialization capacity to 2V+2A further improves performance to 54.58 mAP, 50.13 Hit@K, and 35.49 IoU. Both configurations achieve 85.19 Hit@1, matching TripleSumm while remaining slightly below TR-DETR.

The improvements are more pronounced on SV-Highlights, which contains substantially longer and more diverse sports videos. The default 1V+1A model achieves 52.63 mAP, improving over TripleSumm by 6.88 points, while the higher-capacity 2V+2A variant further increases mAP to 53.01. The 2V+2A model also achieves 90.06 Hit@1 and 50.33 Hit@K, while the default 1V+1A configuration obtains the highest IoU of 34.68. These results demonstrate that the proposed framework provides strong performance even with a single branch per modality, while additional specialization capacity can yield further gains at increased computational cost.

Video Summarization. Table 2 reports the results on Mr.HiSum and MoSu. On Mr.HiSum, the default 1V+1A model achieves a Kendall's Tau of 0.2576, Spearman's correlation of 0.3456, and mAP@50 and mAP@15 scores of 70.66 and 41.05, respectively, outperforming the competing methods across all four metrics. Increasing the branch capacity to 2V+2A further improves Kendall's Tau to 0.2596, Spearman's correlation to 0.3487, and mAP@15 to 41.30. On MoSu, the default model achieves 0.3548 Kendall's Tau, 0.4698 Spearman's correlation, 74.82 mAP@50, and 44.77 mAP@15. The 2V+2A variant provides modest additional gains in Spearman's correlation and mAP@15, reaching 0.4700 and 44.86, respectively, while maintaining comparable performance on the remaining metrics. Overall, the results show that the proposed framework remains effective across both long-form sports highlights and shorter, diverse web-video summarization benchmarks, while increased branch capacity provides additional benefits in selected settings.

Efficiency. As shown in Table 1, the default 1V+1A model contains 6.49M trainable parameters and requires 1.56 GFLOPs, while the higher-capacity 2V+2A variant increases these to 8.60M parameters and 2.51 GFLOPs. Unless otherwise specified, efficiency is measured using an input sequence length of 64 clips and an inference batch size of 8. Although the default model uses more trainable parameters than TripleSumm, its computational cost remains comparable and its GFLOPs are lower (1.56 vs. 1.93), while providing substantially stronger predictive performance. The 2V+2A configuration can provide further accuracy gains at additional computational cost. We therefore adopt 1V+1A as the default configuration, as it provides the most favorable overall trade-off between predictive performance and computational efficiency.

Table 3: Effect of branch capacity under different multimodal feature settings on SV-Highlights. Results are reported as mean ± standard deviation over three runs.

Features	Branches	mAP	Hit@1	Hit@K	IoU	Params. (M) ↓
CLIP + AST	2 Shared Branches	49.15 ± 0.12	86.87 ± 0.86	48.02 ± 0.76	32.78 ± 0.91	6.49
	1V + 1A	51.02 ± 0.45	85.85 ± 4.23	49.07 ± 1.05	33.43 ± 0.82	6.49
	2V + 2A	51.24 ± 0.71	**91.63 ± 1.75**	49.28 ± 0.73	33.58 ± 0.57	8.60
	3V + 3A	51.29 ± 0.28	90.05 ± 0.86	49.49 ± 0.10	33.78 ± 0.06	10.71
	4V + 4A	51.36 ± 0.74	89.55 ± 3.88	49.43 ± 0.89	33.75 ± 0.65	12.81
	1V + 2A	51.47 ± 0.57	87.96 ± 0.80	49.37 ± 0.66	33.71 ± 0.52	7.55
	2V + 1A	**51.61 ± 0.85**	89.02 ± 2.63	**49.66 ± 0.75**	**33.92 ± 0.64**	7.55
CLIP+SlowFast AST+CLAP	2 Shared Branches	52.65 ± 0.23	89.97 ± 0.78	49.12 ± 0.56	34.34 ± 0.78	6.49
	1V + 1A	53.72 ± 0.43	91.11 ± 2.33	51.20 ± 0.70	35.49 ± 0.53	6.49
	2V + 2A	53.85 ± 0.89	92.14 ± 3.13	50.97 ± 0.92	35.25 ± 0.72	8.60
	3V + 3A	**54.16 ± 0.50**	91.63 ± 2.35	**51.55 ± 0.41**	**35.73 ± 0.31**	10.71
	4V + 4A	53.67 ± 0.49	**92.16 ± 3.09**	51.12 ± 0.70	35.39 ± 0.45	12.81
	1V + 2A	53.18 ± 0.53	91.10 ± 3.23	50.90 ± 0.76	35.16 ± 0.48	7.55
	2V + 1A	53.19 ± 0.96	89.02 ± 3.06	50.84 ± 0.91	35.06 ± 0.79	7.55

4.3 Ablation Studies

Shared vs. Modality-Specific Temporal Modeling We examine whether visual and audio streams benefit from dedicated temporal parameters under a matched parameter budget. The shared variant uses two independently parameterized temporal branches, each applied to both modalities, producing four representations for adaptive routing. The 1V+1A variant instead assigns one branch exclusively to each modality and routes the resulting two representations. Both variants therefore contain two sets of temporal parameters and have the same total parameter count of 6.49M at the precision reported in Table 3. Modality-specific allocation improves mAP from 49.15 to 51.02 with CLIP+AST and from 52.65 to 53.72 with the richer feature configuration. This comparison controls parameter count, although the number of routed representations differs.

We further examine temporal parameter sharing within TripleSumm while retaining its fusion mechanism. Independent temporal modules improve mAP by 1.61 points on SV-Highlights and 0.50 points on SoccerHigh. Although this modification increases parameter count, the results complement our parameter-matched comparison by showing that independent temporal processing also benefits another architecture on these sports benchmarks (Appendix A.5).

Effect of Branch Capacity We next vary the number of modality-specific branches to study the effect of specialization capacity. Under the simpler CLIP+AST setting, increasing branch capacity yields only modest gains, with 2V+1A achieving the strongest overall performance at 51.61 mAP. In contrast, richer CLIP+SlowFast and AST+CLAP representations benefit more from increased branch capacity, with 3V+3A reaching 54.16 mAP and achieving the best Hit@K and IoU. Increasing the capacity further to 4V+4A does not improve performance, indicating diminishing returns. Overall, these results suggest that additional branches are most useful when the underlying multimodal representations are sufficiently diverse to support further specialization, while simpler feature combinations can be modeled effectively with a smaller branch bank. Additional comparisons across datasets are provided in Appendix A.5.1.

Adaptive versus Fixed Fusion Table 5 compares no routing, static routing, and the proposed adaptive routing under both 1V+1A and 2V+2A configurations. Across both datasets, the full adaptive model consistently outperforms the corresponding no-routing and static-routing variants on the majority of metrics. On SV-Highlights, adaptive routing improves mAP from 51.65 to 52.63 in the 1V+1A setting and from 51.28 to 53.01 in the 2V+2A setting, while also improving Hit@K and IoU. A similar trend is observed on Mr.HiSum, where the full 2V+2A model achieves the best performance across all four metrics. These results suggest that allowing branch contributions to vary dynamically over time is more effective than either uniform aggregation or fixed global weighting, particularly when multiple modality-specific branches are available.

Modality Robustness We evaluate the robustness of different methods when only a single modality is available at inference time. Specifically, all models are trained under the standard multimodal

Table 4: Performance comparison on SV-Highlights under visual-only and audio-only settings. Results are reported as mean ± standard deviation over three random seeds. Best results are shown in bold.

	Visual Only				Audio Only			
Method	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
MS-DETR	20.73 ± 1.10	32.51 ± 7.75	23.82 ± 1.26	13.94 ± 0.86	30.06 ± 3.50	52.92 ± 6.46	34.27 ± 2.90	21.08 ± 2.18
QD-DETR	25.88 ± 0.31	41.35 ± 1.45	29.42 ± 0.45	17.70 ± 0.27	31.43 ± 0.87	57.54 ± 8.73	34.58 ± 0.56	21.38 ± 0.40
SG-DETR	24.75 ± 1.62	46.09 ± 3.58	28.43 ± 1.62	17.08 ± 1.06	13.70 ± 6.19	24.56 ± 10.30	14.67 ± 8.42	8.32 ± 4.97
TR-DETR	25.60 ± 0.78	48.23 ± 10.47	28.55 ± 1.03	17.26 ± 0.68	29.16 ± 3.02	68.60 ± 8.18	32.10 ± 2.28	19.51 ± 1.59
TripleSumm	25.75 ± 1.88	42.94 ± 7.92	28.18 ± 1.65	16.98 ± 1.27	35.54 ± 3.49	77.47 ± 3.42	37.46 ± 3.11	23.50 ± 2.36
UMT	17.47 ± 4.33	36.08 ± 6.55	19.77 ± 5.95	11.35 ± 3.68	23.05 ± 9.46	57.11 ± 18.52	25.68 ± 11.15	15.34 ± 7.21
Uni-VTG	25.51 ± 0.96	48.19 ± 4.32	29.55 ± 1.31	17.80 ± 0.80	7.70 ± 0.37	24.10 ± 2.61	10.53 ± 0.37	5.75 ± 0.23
Ours	**42.99 ± 0.77**	**79.61 ± 4.57**	**43.02 ± 1.20**	**28.44 ± 0.71**	**42.27 ± 1.03**	**83.28 ± 9.87**	**42.93 ± 0.88**	**27.91 ± 0.79**

Table 5: Ablation study of routing strategies on SV-Highlights and Mr.HiSum. Results are reported as mean ± standard deviation over three random seeds. Best results within each dataset are shown in bold.

SV-Highlights					
Branches	Variant	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
1V+1A	No Routing	51.05 ± 0.78	87.58 ± 1.50	47.46 ± 0.28	32.77 ± 0.15
1V+1A	Static Routing	51.65 ± 0.54	88.61 ± 2.48	48.98 ± 0.72	33.33 ± 0.50
1V+1A	Full	52.63 ± 0.83	89.03 ± 4.64	50.31 ± 0.92	**34.68 ± 0.76**
2V+2A	No Routing	51.09 ± 1.31	88.02 ± 4.11	48.48 ± 1.21	32.79 ± 0.95
2V+2A	Static Routing	51.28 ± 0.65	87.95 ± 2.41	48.71 ± 0.57	33.03 ± 0.38
2V+2A	Full	**53.01 ± 1.32**	**90.06 ± 3.92**	**50.33 ± 0.64**	34.59 ± 0.61
Mr.HiSum					
Branches	Variant	τ ↑	ρ ↑	mAP@50 ↑	mAP@15 ↑
1V+1A	No Routing	0.2358 ± 0.0061	0.3239 ± 0.0078	68.5843 ± 0.2856	39.4825 ± 0.5156
1V+1A	Static Routing	0.2374 ± 0.0085	0.3257 ± 0.0113	69.6152 ± 0.5029	39.6554 ± 0.6175
1V+1A	Full	0.2576 ± 0.0043	0.3456 ± 0.0055	70.6557 ± 0.1981	41.0469 ± 0.0739
2V+2A	No Routing	0.2418 ± 0.0025	0.3323 ± 0.0041	68.9210 ± 0.0664	39.9771 ± 0.2703
2V+2A	Static Routing	0.2322 ± 0.0024	0.3417 ± 0.0043	69.8849 ± 0.0529	38.4261 ± 0.3360
2V+2A	Full	**0.2596 ± 0.0052**	**0.3487 ± 0.0076**	**70.7120 ± 0.4028**	**41.2980 ± 0.3863**

setting, while during evaluation the unavailable modality is masked by replacing its input features with zeros, resulting in visual-only and audio-only test conditions. As shown in Table 4, our method maintains strong performance under both settings, achieving the best results across all metrics. In the visual-only setting, our model reaches 42.99 mAP, substantially outperforming the strongest baseline. Under the audio-only setting, it achieves 42.27 mAP and similarly remains ahead of competing approaches. These results indicate that the proposed modality-specific temporal modeling learns useful representations within each modality and that the model remains effective even when one modality is unavailable at inference time.

5 Conclusion

We presented a multimodal framework built around a *specialize-before-fusion* principle for highlight detection and video summarization. Rather than directly combining generic modality representations, the proposed approach first models visual and audio streams using independently parameterized temporal branches, allowing each modality to preserve its own temporal characteristics before multimodal integration. The resulting representations are then adaptively combined according to the current video context. Across four diverse benchmarks, the proposed approach achieves strong performance while maintaining moderate computational cost. Ablation studies show that modality-specific temporal modeling consistently outperforms shared temporal processing, while adaptive routing further improves over fixed or uniform fusion. Additional analyses indicate that visual and audio representations become more distinct around salient moments and that their relative contributions vary across temporal locations. Overall, these results support the central hypothesis that heterogeneous modalities should be temporally specialized before fusion, providing a simple and effective design principle for multimodal video highlight understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with developing and refining theoretical models and conceptual frameworks, proposing and refining research hypotheses, suggesting experimental parameters, providing feedback on research methodology and experimental design, assisting with method implementation, interpreting experimental results, identifying potentially relevant literature, sourcing and searching for background information, drafting portions of the manuscript, and editing the manuscript to improve clarity and readability.

All AI-assisted outputs were critically reviewed by the authors. Suggested hypotheses, methodological choices, experimental parameters, and interpretations were evaluated through our own reasoning, literature review, and empirical experiments. AI-assisted code and method implementations were inspected, tested, and validated by the authors before use. Literature suggestions and sourced information were independently checked against the original publications or other primary sources before being incorporated into the manuscript. Drafted or edited text was reviewed and revised by the authors to ensure technical accuracy, appropriate attribution, and consistency with the actual methods and experimental results.

All reported results were produced from our own implementations and experiments, and all scientific claims and interpretations were independently verified by the authors. We take full responsibility for the final content of this work, including all text, claims, methods, experimental results, references, and artifacts produced with the aid of generative AI.

REFERENCES

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

Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Dernoncourt, and Joon Son Chung. Scaling up video summarization pretraining with large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8332–8341. IEEE, 2024.

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8107–8117. IEEE, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Artur Díaz-Juan, Coloma Ballester, and Gloria Haro. Soccerhigh: a benchmark dataset for automatic soccer video summarization. In *Proceedings of the 8th International ACM Workshop on Multimedia Content Analysis in Sports*, pp. 121–130, 2025.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 6201–6210. IEEE, 2019.

Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. Learning event-specific localization preferences for audio-visual event localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3446–3454, 2023.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1792–179210. IEEE, 2018.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 907–916. IEEE, 2026.

Qihao Guan. The impact of short videos on long video engagement: A comparative analysis of promotional and non-promotional content on youtube. *Available at SSRN 4979201*, 2024.

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

Zahidul Islam, Sujoy Paul, and Mrigank Rochan. Test-time adaptation for video highlight detection using meta-auxiliary learning and cross-modality hallucinations. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6893–6902. IEEE, 2026.

Hao Jiang and Yadong Mu. Joint video summarization and moment localization by cross-task sample transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16367–16377. IEEE, 2022.

Kwanseok Kim, Jaehoon Hahm, Sumin Kim, Jinhwan Sul, Byunghak Kim, and Joonseok Lee. Summdiff: generative modeling of video summarization with diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15096–15106. IEEE, 2025.

Sumin Kim, Hyemin Jeong, Mingu Kang, Yejin Kim, Yoori Oh, and Joonseok Lee. Triplesumm: Adaptive triple-modality fusion for video summarization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Donggyu Lee, Youngbin Ki, Jeonghun Kang, and Taehwan Kim. Svhighlights: Towards extremely long sport video highlight detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9254–9265, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817564. URL https://doi.org/10.1145/3770855.3817564.

Donggyu Lee, Youngbin Ki, Jeonghun Kang, and Taehwan Kim. Svhighlights: Towards extremely long sport video highlight detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9254–9265, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817564. URL https://doi.org/10.1145/3770855.3817564.

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

Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. Progressive video summarization via multimodal self-supervised learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5573–5582. IEEE, 2023.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2782–2792. IEEE, 2023.

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

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

Hongxu Ma, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4514–4523, 2025.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23023–23033. IEEE, 2023.

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

Zirui Shang, Yubo Zhu, Hongxi Li, Shuo Yang, and Xinxiao Wu. Video summarization using denoising diffusion probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6776–6784, 2025.

Jaewon Son, Jaehun Park, and Kwangsu Kim. Csta: Cnn-based spatiotemporal attention for video summarization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18847–18856. IEEE, 2024.

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

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4998–5007, 2024.

Hacene Terbouche, Maryan Morel, Mariano Rodriguez, and Alice Othmani. Multi-annotation attention model for video summarization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3143–3152. IEEE, 2023.

Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM web science conference*, pp. 213–223, 2024.

Jiehang Xie, Xuanbai Chen, Shao-Ping Lu, and Yulu Yang. A knowledge augmented and multimodal-based framework for video summarization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 740–749, 2022.

Hao Xu, Meenakshi Sarkar, Vishnu Raj, and David Gunawan. Audio for sports highlight detection: A comparative empirical study. In *9th Int. Workshop on Multimedia Content Analysis in Sports (MMSports' 26)@ ACM Multimedia 2026*, 2026.

Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 3893–3901, 2020.

Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Youyao Jia, and Sidan Du. Mh-detr: Video moment and highlight detection with cross-modal transformer. In *2024 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2024.

Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18308–18318. IEEE, 2024.

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)*, pp. 7930–7939. IEEE, 2021.

Qinghao Yu, Hui Yu, Ying Sun, Derui Ding, and Muwei Jian. Unsupervised video summarization based on the diffusion model of feature fusion. *IEEE Transactions on Computational Social Systems*, 11(5):6010–6021, 2024.

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

Yunzuo Zhang, Yameng Liu, Weili Kang, and Ran Tao. Vss-net: Visual semantic self-mining network for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2775–2788, 2023.

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

Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2793–2801, 2021.

Bin Zhao, Maoguo Gong, and Xuelong Li. Hierarchical multimodal transformer to summarize videos. *Neurocomputing*, 468:360–369, 2022.

Henghao Zhao, Kevin Qinghong Lin, Rui Yan, and Zechao Li. Diffusionvmr: Diffusion model for joint video moment retrieval and highlight detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14522–14535, 2024.

Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020.

A Appendix

A.1 Implementation Details

All models are implemented in PyTorch. We use the pre-extracted visual, audio, and textual features described in Sec. 4. When multiple feature extractors are available for the same modality, their representations are first combined at the feature level and then projected into a shared hidden dimension of $d = 256$. The projected visual and audio sequences are temporally aligned before being passed to the modality-specific temporal branches.

For each modality, we employ one visual branch and one audio branch in the default configuration. Each branch is implemented using a single-layer bidirectional GRU with hidden dimension 256 and dropout 0.1. The temporal branches are independently parameterized, while the query-conditioning module is shared across branches within the same modality. Query conditioning is implemented using multi-head cross-attention with 4 attention heads and a single attention layer.

The adaptive router takes as input the concatenated projected visual and audio features, the masked mean-pooled textual query representation, and the mean representation of the available query-conditioned branch outputs. The routing network consists of LayerNorm followed by a two-layer MLP with GELU activation and dropout, using a hidden dimension of 128. Routing weights are obtained using a temperature-scaled softmax with temperature $\tau = 1$. The weighted branch representation is combined with a projected residual connection from the concatenated visual and audio features, followed by LayerNorm.

The fused representation is subsequently processed by a temporal classifier consisting of a single GRU layer followed by a saliency prediction head, which produces a highlight relevance logit for each temporal clip. For training, we use temporal sequences of 256 clips with a stride of 32 clips. Models are trained for 30 epochs using AdamW with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and a cosine learning-rate schedule. We use a batch size of 256. To account for the imbalance between highlight and non-highlight clips, we optimize a weighted binary cross-entropy loss in which positive clips are assigned a weight of 5 and negative clips a weight of 1. Training is performed on an NVIDIA A100 40GB GPU.

During inference, we use a shorter temporal window of 64 clips with the same stride of 32 clips and an inference batch size of 8. Predictions from overlapping temporal windows are combined to recover clip-level saliency scores over the full video.

For fair comparison, we retrain baseline methods using their official implementations whenever available. We keep the same dataset splits, input features, training budget, and evaluation protocol across methods whenever supported by their architectures. For methods that do not support all modalities, we retain their original modality configuration rather than modifying the architecture. All experiments are repeated using three random seeds (42, 2026, and 3020), and we report the mean and standard deviation across runs.

A.2 Dataset Details

SV-Highlights. SV-Highlights Lee et al. (2026a) contains 320 long-form sports videos covering eight sports: American football, baseball, basketball, ice hockey, racing, rugby, soccer, and volleyball, with 40 videos per sport. Due to YouTube availability restrictions, three videos could not be downloaded, and the original videos could not be redistributed by the dataset authors. We therefore conduct our experiments on the remaining 317 videos. Following the dataset protocol, each video is divided into non-overlapping 2-second clips. Each clip is assigned a binary label, where 1 indicates that the clip belongs to a highlight segment and 0 otherwise. The ground-truth annotations are derived from the corresponding official highlight videos. We use the provided CLIP Radford et al. (2021) and SlowFast Feichtenhofer et al. (2019) visual features. Since audio features are not included in the original release, we additionally extract audio representations using pretrained AST Gong et al. (2021) and CLAP Elizalde et al. (2023). All modalities are temporally aligned at the clip level before being provided to the model.

SoccerHigh. SoccerHigh Díaz-Juan et al. (2025) is derived from SoccerNet Giancola et al. (2018) and contains 237 paired full-match soccer videos from three major European leagues. Each match

Table 6: Per-sport performance on SV-Highlights. Results are reported as mean ± standard deviation over three random seeds.

Sport	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
American Football	68.75 ± 0.21	87.50 ± 0.00	63.29 ± 0.66	46.62 ± 0.84
Baseball	36.87 ± 1.82	87.50 ± 12.50	37.80 ± 1.36	23.41 ± 1.04
Basketball	58.19 ± 1.11	91.67 ± 14.43	56.23 ± 1.78	39.80 ± 1.25
Ice Hockey	65.63 ± 4.31	100.00 ± 0.00	59.66 ± 4.28	42.92 ± 4.36
Racing	44.02 ± 6.14	91.67 ± 14.43	44.22 ± 4.96	28.96 ± 4.22
Rugby	59.83 ± 4.64	95.83 ± 7.22	55.61 ± 1.25	38.80 ± 1.34
Soccer	56.63 ± 7.08	100.00 ± 0.00	53.59 ± 6.44	37.60 ± 5.29
Volleyball	31.48 ± 1.03	58.33 ± 7.22	32.44 ± 1.72	19.60 ± 1.11
Overall	**52.63 ± 0.83**	**89.03 ± 4.63**	**50.31 ± 0.90**	**34.68 ± 0.76**

consists of two halves and is processed according to the temporal annotations provided by the benchmark. To maintain a consistent feature configuration with SV-Highlights, we extract CLIP and SlowFast visual representations and corresponding audio features using pretrained audio encoders. Visual and audio features are aligned to the same temporal clip boundaries before training and evaluation. We use the official dataset split provided by SoccerHigh without modifying the video-level partitioning.

Mr.HiSum. Mr.HiSum Sul et al. (2023) is a large-scale video summarization benchmark derived from YouTube-8M Abu-El-Haija et al. (2016). It contains 31,892 videos spanning 3,509 semantic entities, with video durations below 300 seconds. The benchmark provides frame-level importance annotations derived from YouTube's most-replayed statistics. For our experiments, we follow the multimodal setup of TripleSumm Kim et al. (2026) and use its released pre-extracted visual, textual, and audio representations. This additionally provides audio and textual modalities that are not included in the original Mr.HiSum release. We follow the official train, validation, and test partitions used in TripleSumm to ensure direct comparability.

MoSu. MoSu Kim et al. (2026) contains 52,678 YouTube videos spanning 3,406 semantic categories. The videos have an average duration of approximately 272.3 seconds, with durations ranging from 120 to 501 seconds. Similar to Mr.HiSum, the dataset uses most-replayed statistics as a saliency signal for video summarization. We directly use the pre-extracted visual, textual, and audio representations released with TripleSumm and follow its official data splits and preprocessing protocol.

Temporal Alignment and Preprocessing. For all datasets, visual, audio, and textual representations are aligned to the temporal units used for saliency prediction. When multiple pretrained encoders are used for the same modality, their features are temporally matched before being provided to the corresponding modality stream. Videos are padded within each mini-batch when necessary, and validity masks are used to ensure that padded temporal locations do not contribute to temporal modeling, routing, or evaluation.

A.3 Detailed Results

Per-sport Analysis. Table 6 reports performance across the eight sports in SV-Highlights. The model performs particularly strongly on American football and ice hockey, achieving mAP scores of 68.75 and 65.63, respectively. Strong performance is also observed for rugby, basketball, and soccer. In contrast, volleyball and baseball remain more challenging, with substantially lower mAP and temporal overlap scores. The variation across sports suggests that highlight characteristics differ considerably between sporting domains, particularly in the relative usefulness and temporal structure of visual and acoustic cues.

Table 7: Effect of different prompt formulations on SoccerHigh and SV-Highlights. Results are reported as mean ± standard deviation over three runs.

Dataset	Prompt Setting	mAP	Hit@1	Hit@K	IoU
SoccerHigh	No Prompt	53.04 ± 0.58	84.93 ± 1.70	48.72 ± 0.32	34.15 ± 0.29
	Single Prompt	54.19 ± 0.43	**86.30 ± 0.64**	49.78 ± 0.41	35.18 ± 0.35
	Soccer-Specific Prompt	**54.31 ± 0.68**	84.81 ± 0.65	**49.93 ± 0.55**	**35.31 ± 0.44**
	Visual-Audio Pool (30)	53.71 ± 0.15	85.93 ± 1.29	49.45 ± 0.14	34.88 ± 0.20
	General Highlight Pool (30)	54.27 ± 0.48	85.19 ± 1.70	49.85 ± 0.41	35.22 ± 0.28
	Descriptive Pool (30)	54.10 ± 1.16	85.93 ± 0.64	49.76 ± 0.96	35.16 ± 0.87
SV-Highlights	No Prompt	52.49 ± 0.83	88.02 ± 3.06	49.13 ± 0.69	33.46 ± 0.47
	Single Prompt	52.84 ± 0.90	90.06 ± 1.73	50.20 ± 1.13	34.55 ± 0.83
	Visual-Audio Pool (30)	52.89 ± 0.67	**91.63 ± 1.75**	50.21 ± 0.69	34.53 ± 0.48
	General Highlight Pool (30)	52.93 ± 0.93	88.50 ± 3.87	**50.31 ± 0.87**	**34.64 ± 0.64**
	Descriptive Pool (30)	**52.94 ± 0.91**	90.06 ± 1.73	50.14 ± 0.94	34.47 ± 0.71
	Sports Descriptive Pool (32)	52.63 ± 1.10	85.87 ± 1.46	50.06 ± 1.27	34.32 ± 1.07

A.4 More Ablation Studies

A.4.1 Effect of Semantic Query Conditioning

We investigate whether the semantic formulation of the textual query affects highlight prediction by comparing models trained without a prompt, with a single generic prompt, and with several domain-specific or diversified prompt sets. In the single-prompt setting, the same textual query Q is fixed for all videos within a dataset throughout training. For the multi-prompt settings, a prompt is randomly sampled from the corresponding prompt pool on a per-batch basis during training, introducing variation in the semantic conditioning while keeping the underlying highlight objective unchanged. During validation and inference, prompt selection is deterministic: a single fixed prompt is used for all videos to ensure consistent evaluation.

As shown in Table 7, prompt conditioning provides modest but generally positive gains over the no-prompt baseline, although the differences between prompt formulations remain small. On SoccerHigh, the soccer-specific prompt achieves the best mAP, Hit@K, and IoU, suggesting that domain-aligned semantic guidance can be beneficial. On SV-Highlights, however, the general and descriptive prompt pools perform similarly, with no single formulation consistently dominating across all metrics. Overall, these results indicate that textual conditioning is useful, while performance is relatively robust to the exact prompt formulation. Representative prompts include *"Find the highlight moments in this sports video"* for the generic setting, *"Find important soccer moments such as goals, saves, dangerous scoring opportunities, or major referee decisions"* for the domain-specific setting, and *"Find highlights based on what can be seen and heard in the video"* for the audio-visual setting.

A.4.2 Effect of Temporal Context Length

Table 8 studies the effect of temporal clip length on SoccerHigh and SV-Highlights. The optimal context length differs across the two datasets. On SoccerHigh, performance generally improves as the clip length increases from 32 to 256, with a length of 256 achieving the best mAP, Hit@K, and IoU. In contrast, SV-Highlights performs best with a shorter clip length of 32, while longer contexts provide no consistent improvement. This suggests that the amount of useful temporal context is dataset-dependent: SoccerHigh benefits from broader temporal modeling, whereas the more diverse SV-Highlights videos can be effectively modeled using shorter local contexts. Overall, performance remains relatively stable across a broad range of clip lengths, indicating that the proposed framework is not highly sensitive to this hyperparameter.

A.4.3 Feature Extractors

Table 9 examines the effect of different visual and audio feature combinations. The results show that feature complementarity is dataset-dependent. On SoccerHigh, CLIP combined with AST+CLAP achieves the strongest overall performance, reaching 55.74 mAP, 50.68 Hit@K, and 35.89 IoU, while adding SlowFast does not provide further gains. In contrast, SV-Highlights benefits more clearly from richer multimodal representations, where combining CLIP+SlowFast with AST+CLAP achieves the best mAP, Hit@K, and IoU.

Table 8: Effect of clip length on SoccerHigh and SV-Highlights. Results are reported as mean ± standard deviation over three random seeds. Best results for each dataset are shown in bold.

Dataset	Clip Length	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
SoccerHigh	32	52.39 ± 0.93	84.07 ± 1.29	48.38 ± 0.51	33.83 ± 0.42
	64	53.97 ± 0.30	**86.67 ± 1.11**	49.46 ± 0.62	34.94 ± 0.53
	128	53.98 ± 0.47	85.18 ± 1.29	49.83 ± 0.34	35.26 ± 0.39
	256	**54.31 ± 0.77**	85.19 ± 0.65	**50.00 ± 0.89**	**35.35 ± 0.74**
	512	53.86 ± 0.29	85.19 ± 1.70	49.62 ± 0.56	35.09 ± 0.54
SV-Highlights	32	**53.48 ± 0.14**	**92.15 ± 0.08**	**50.52 ± 0.39**	**34.83 ± 0.28**
	64	52.70 ± 0.57	91.10 ± 0.83	50.26 ± 0.83	34.57 ± 0.57
	128	52.83 ± 0.68	89.02 ± 3.06	50.27 ± 0.48	34.57 ± 0.38
	256	52.63 ± 0.83	89.03 ± 4.64	50.31 ± 0.92	34.68 ± 0.76
	512	52.65 ± 0.91	90.07 ± 3.20	49.99 ± 0.72	34.34 ± 0.54

Table 9: Effect of different visual and audio feature extractor combinations on SoccerHigh and SV-Highlights. Results are reported as mean ± standard deviation over three random seeds. Best results for each dataset are shown in bold.

Dataset	Visual Features	Audio Features	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
SoccerHigh	CLIP	AST	54.15 ± 0.73	86.30 ± 1.70	49.16 ± 1.28	34.58 ± 1.15
	CLIP + SlowFast	AST	54.31 ± 0.77	85.19 ± 0.65	50.00 ± 0.89	35.35 ± 0.74
	CLIP	CLAP	55.47 ± 0.42	**88.52 ± 0.64**	50.27 ± 0.35	35.41 ± 0.30
	CLIP + SlowFast	CLAP	53.77 ± 0.70	86.30 ± 2.31	49.13 ± 0.78	34.33 ± 0.65
	CLIP	AST + CLAP	**55.74 ± 0.90**	87.78 ± 1.11	**50.68 ± 0.75**	**35.89 ± 0.66**
	CLIP + SlowFast	AST + CLAP	54.66 ± 0.37	87.41 ± 1.70	50.19 ± 0.28	35.56 ± 0.32
SV-Highlights	CLIP	AST	51.02 ± 0.45	85.85 ± 4.23	49.07 ± 1.05	33.43 ± 0.82
	CLIP + SlowFast	AST	52.63 ± 0.83	89.03 ± 4.64	50.31 ± 0.92	34.68 ± 0.76
	CLIP	CLAP	52.84 ± 0.44	87.45 ± 3.05	50.47 ± 0.33	34.75 ± 0.26
	CLIP + SlowFast	CLAP	52.85 ± 1.18	**91.60 ± 3.72**	50.24 ± 1.12	34.55 ± 0.91
	CLIP	AST + CLAP	52.16 ± 0.93	89.04 ± 5.59	49.94 ± 0.73	34.26 ± 0.61
	CLIP + SlowFast	AST + CLAP	**53.72 ± 0.43**	91.11 ± 2.33	**51.20 ± 0.70**	**35.49 ± 0.53**

These results suggest that using additional pretrained representations is not uniformly beneficial; rather, their effectiveness depends on whether they provide complementary information for the target dataset. In particular, the gains on SV-Highlights indicate that its more diverse sports content can benefit from combining appearance, motion, and multiple acoustic representations, whereas the more homogeneous SoccerHigh setting can be effectively modeled with a simpler visual representation.

A.4.4 Zero-Shot Performance

We further evaluate cross-dataset generalization by training each model on one sports highlight benchmark and directly testing it on the other without target-domain fine-tuning. This evaluation is feasible because both SV-Highlights and SoccerHigh use compatible clip-level highlight annotations, allowing the same clip-level prediction formulation to be applied across datasets. To avoid introducing dataset-specific semantic guidance during transfer, the textual query is fixed to the same generic prompt used for validation and inference in both transfer directions.

As shown in Table 10, our method achieves the strongest performance in both directions. When trained on SV-Highlights and evaluated on SoccerHigh, it reaches 44.86 mAP, outperforming TripleSumm by 1.90 points. The reverse setting is more challenging, yet our method still achieves 31.98 mAP compared with 28.35 for TripleSumm, while also obtaining the best Hit@1, Hit@K, and IoU. These results suggest that the learned multimodal representations transfer reasonably well across sports domains and are less dependent on dataset-specific highlight patterns.

A.4.5 Effect of Temporal Modeling

Tables 11 and 12 compare different temporal modeling choices, including MLP, GRU, LSTM, and self-attention. Recurrent models provide the strongest overall performance across the four benchmarks. On SoccerHigh, GRU achieves the best mAP, Hit@K, and IoU, while on SV-Highlights GRU and LSTM perform similarly, with LSTM obtaining slightly higher mAP and Hit@1 and GRU

Table 10: Cross-dataset generalization between SV-Highlights and SoccerHigh. Models are trained on one dataset and directly evaluated on the other without target-domain fine-tuning.

Train → Test	Method	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
SV-Highlights → SoccerHigh	MS-DETR	34.19	71.11	34.62	22.00
	QD-DETR	39.38	75.56	39.43	25.68
	TR-DETR	39.61	78.89	38.60	25.22
	TripleSumm	42.96	78.89	40.27	26.63
	UMT	40.20	75.56	38.14	25.02
	Ours	**44.86**	**80.00**	**43.27**	**29.30**
SoccerHigh → SV-Highlights	MS-DETR	22.76	48.44	24.73	14.89
	QD-DETR	22.68	45.31	24.20	14.43
	TR-DETR	19.50	43.75	20.46	11.98
	TripleSumm	28.35	68.75	29.27	18.03
	UMT	25.06	50.00	27.85	16.66
	Ours	**31.98**	**76.56**	**32.40**	**20.25**

Table 11: Effect of different temporal modeling layers across four benchmarks. Results are reported as mean ± standard deviation over three random seeds. Best results for each dataset are shown in bold.

SoccerHigh				
Temporal Layer	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
MLP	40.53 ± 0.15	80.74 ± 1.28	40.35 ± 0.20	26.33 ± 0.16
GRU	**54.31 ± 0.77**	85.19 ± 0.65	**50.00 ± 0.89**	**35.35 ± 0.74**
LSTM	53.83 ± 0.53	83.70 ± 1.70	49.50 ± 0.46	34.95 ± 0.39
Attention	49.00 ± 0.07	**86.30 ± 0.64**	45.60 ± 0.11	30.97 ± 0.12
SV-Highlights				
Temporal Layer	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
MLP	41.01 ± 1.04	82.72 ± 0.16	41.65 ± 1.00	26.88 ± 0.76
GRU	52.63 ± 0.83	89.03 ± 4.64	**50.31 ± 0.92**	**34.68 ± 0.76**
LSTM	**52.85 ± 0.46**	**89.51 ± 3.38**	49.89 ± 0.34	34.24 ± 0.28
Attention	42.85 ± 2.63	85.86 ± 4.69	42.92 ± 1.96	28.05 ± 1.64
Mr.HiSum				
Temporal Layer	τ ↑	ρ ↑	mAP@50 ↑	mAP@15 ↑
MLP	0.2112 ± 0.0013	0.2964 ± 0.0021	68.3630 ± 0.2260	37.6948 ± 0.8584
GRU	**0.2576 ± 0.0043**	**0.3456 ± 0.0055**	70.6557 ± 0.1981	**41.0469 ± 0.0739**
LSTM	0.2349 ± 0.0028	0.3217 ± 0.0032	69.1807 ± 0.1885	38.9640 ± 0.3675
Attention	0.2477 ± 0.0023	0.3367 ± 0.0051	**70.7425 ± 0.0196**	40.3521 ± 0.3488
MoSu				
Temporal Layer	τ ↑	ρ ↑	mAP@50 ↑	mAP@15 ↑
MLP	0.3042 ± 0.0005	0.4164 ± 0.0014	72.5344 ± 0.0605	40.6957 ± 0.0662
GRU	0.3548 ± 0.0003	0.4698 ± 0.0008	74.8219 ± 0.0432	44.7737 ± 0.0782
LSTM	**0.3549 ± 0.0015**	**0.4713 ± 0.0009**	**74.8272 ± 0.1382**	**45.0283 ± 0.1002**
Attention	0.3302 ± 0.0089	0.4417 ± 0.0115	73.7701 ± 0.3513	43.3358 ± 0.4128

achieving better Hit@K and IoU. GRU also performs best on most Mr.HiSum metrics, whereas LSTM provides a small advantage on MoSu.

In contrast, the MLP variant, which processes each temporal location independently, shows a substantial performance drop across all datasets despite being considerably more efficient. This highlights the importance of explicitly modeling temporal dependencies for highlight prediction. Self-attention offers higher throughput than recurrent models but generally underperforms GRU and LSTM in predictive accuracy. Overall, we adopt GRU as the default temporal layer because it provides the most consistent performance across datasets while maintaining moderate computational complexity.

A.5 Shared versus Modality-Specific Temporal Modeling in TripleSumm

TripleSumm Kim et al. (2026) reports that sharing temporal parameters across modalities improves performance in its video summarization setting. We examine whether this finding extends to sports

Table 12: Efficiency comparison of different temporal modeling layers.

Temporal Layer	Params. ↓	GFLOPs ↓	Latency ↓	Throughput ↑	Memory ↓
MLP	**3.43M**	**1.41**	**0.55 ms**	**1829.58**	**138.44 MB**
GRU	6.49M	1.56	2.38 ms	419.81	217.83 MB
LSTM	7.41M	1.56	2.31 ms	433.74	238.48 MB
Attention	8.03M	2.96	0.76 ms	1308.91	208.53 MB

Table 13: Effect of temporal parameter sharing in TripleSumm on sports highlight detection. Results are reported as mean ± standard deviation over three random seeds. Bold indicates the higher mean within each dataset.

Dataset	Parameters	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
SV-Highlights	Shared	45.75 ± 0.93	84.28 ± 6.88	45.13 ± 0.50	29.98 ± 0.50
	Independent	$\mathbf{47.36 \pm 0.76}$	$\mathbf{87.12 \pm 3.29}$	$\mathbf{46.82 \pm 0.68}$	$\mathbf{30.23 \pm 0.65}$
SoccerHigh	Shared	51.77 ± 0.50	85.19 ± 1.70	$\mathbf{47.87 \pm 0.33}$	33.26 ± 0.36
	Independent	$\mathbf{52.27 \pm 0.23}$	$\mathbf{87.29 \pm 1.43}$	47.62 ± 0.25	$\mathbf{34.32 \pm 0.12}$

highlight detection by comparing its shared-parameter configuration with a variant that assigns an independently parameterized temporal module to each modality. We retain the remaining architecture, including the fusion mechanism, and use the same input features, dataset splits, training protocol, and evaluation procedure for both variants. Results are reported as mean and standard deviation over three random seeds.

As shown in Table 13, independent temporal parameters improve TripleSumm's mAP by 1.61 points on SV-Highlights and 0.50 points on SoccerHigh. All four metrics improve on average on SV-Highlights. On SoccerHigh, mAP, Hit@1, and IoU improve, while Hit@K decreases slightly from 47.87 to 47.62. Thus, the advantage of parameter sharing reported in TripleSumm's original setting does not uniformly extend to these sports benchmarks.

These findings complement our within-framework ablations by showing that independent temporal processing can also benefit a different architecture. Since removing parameter sharing increases the number of trainable parameters, this comparison does not isolate specialization from additional capacity; our parameter-matched shared-branch comparison provides complementary evidence. Furthermore, independent TripleSumm remains below our default 1V+1A model by 5.27 mAP on SV-Highlights and 2.04 mAP on SoccerHigh, indicating that removing parameter sharing alone does not account for the full performance advantage of our framework.

A.5.1 More Shared vs. Modality-Specific Temporal Modeling

We further compare shared and modality-specific temporal modeling across all four benchmarks using the same CLIP+SlowFast visual features and AST audio features on four benchmarks as shown in Table 14 and Table 15. In the shared setting, the visual and audio streams reuse the same temporal branch parameters, whereas the modality-specific variants assign independent temporal branches to each modality. Across all datasets, the modality-specific formulation consistently outperforms the shared baseline. On SoccerHigh, mAP improves from 53.73 with two shared branches to 54.31 with 1V+1A, while on SV-Highlights the gain is larger, from 50.61 to 52.63. Similar trends are observed for video summarization: on Mr.HiSum, Kendall's Tau increases from 0.2391 to 0.2576, and on MoSu from 0.3476 to 0.3548. Increasing the number of modality-specific branches can provide further gains, although the optimal branch allocation varies across datasets and the improvement is not monotonic with branch count. These results further support our central hypothesis that visual and audio streams benefit from independent temporal transformations before multimodal fusion, while additional branch capacity provides a flexible mechanism for increasing specialization when beneficial.

Table 14: Effect of modality-specific branch capacity on SoccerHigh and SV-Highlights using CLIP+SlowFast visual features and AST audio features. Results are reported as mean ± standard deviation over three random seeds. Best results for each dataset are shown in bold.

Dataset	Branches	mAP ↑	Hit@1 ↑	Hit@K ↑	IoU ↑
SoccerHigh	2 Shared Branches	53.73 ± 0.20	83.70 ± 0.64	48.24 ± 0.26	33.57 ± 0.24
	1V+1A	54.31 ± 0.77	85.19 ± 0.65	50.00 ± 0.89	35.35 ± 0.74
	2V+2A	54.58 ± 0.44	85.19 ± 0.65	50.13 ± 0.47	35.49 ± 0.34
	3V+3A	**54.65 ± 0.44**	**87.04 ± 0.64**	49.74 ± 0.81	35.10 ± 0.73
	4V+4A	54.35 ± 0.60	86.67 ± 1.11	**50.18 ± 0.78**	**35.56 ± 0.71**
	2V+1A	53.66 ± 0.52	84.82 ± 1.29	49.22 ± 0.43	34.73 ± 0.47
	1V+2A	54.02 ± 0.59	84.44 ± 0.00	49.62 ± 0.32	35.02 ± 0.42
SV-Highlights	2 Shared Branches	50.61 ± 1.03	88.09 ± 3.66	49.94 ± 0.67	34.30 ± 0.45
	1V+1A	52.63 ± 0.83	89.03 ± 4.64	50.31 ± 0.92	**34.68 ± 0.76**
	2V+2A	**53.01 ± 1.32**	90.06 ± 3.92	**50.33 ± 0.64**	34.59 ± 0.61
	3V+3A	52.57 ± 1.33	87.97 ± 2.31	50.00 ± 1.57	34.31 ± 1.29
	4V+4A	52.47 ± 0.76	88.99 ± 1.66	49.92 ± 1.17	34.25 ± 0.84
	2V+1A	52.80 ± 0.52	**91.10 ± 0.83**	50.10 ± 0.71	34.43 ± 0.45
	1V+2A	52.65 ± 0.84	86.38 ± 0.97	50.19 ± 0.91	34.57 ± 0.70

Table 15: Effect of modality-specific branch capacity on Mr.HiSum and MoSu using released features. Results are reported as mean ± standard deviation over three random seeds. Best results for each dataset are shown in bold.

Dataset	Branches	KTau ↑	SRho ↑	mAP@50 ↑	mAP@15 ↑
Mr.HiSum	2 Shared Branches	0.2391 ± 0.0044	0.3378 ± 0.0053	70.10 ± 0.17	40.17 ± 0.35
	1V+1A	0.2576 ± 0.0043	0.3456 ± 0.0055	70.66 ± 0.20	41.05 ± 0.07
	2V+2A	0.2596 ± 0.0052	0.3487 ± 0.0076	70.71 ± 0.40	**41.30 ± 0.39**
	3V+3A	0.2575 ± 0.0031	0.3467 ± 0.0049	70.64 ± 0.27	41.22 ± 0.29
	4V+4A	**0.2613 ± 0.0033**	0.3500 ± 0.0055	70.86 ± 0.13	41.28 ± 0.33
	2V+1A	0.2590 ± 0.0059	0.3474 ± 0.0085	70.72 ± 0.42	41.04 ± 0.69
	1V+2A	0.2610 ± 0.0074	**0.3503 ± 0.0096**	**70.84 ± 0.37**	41.21 ± 0.47
MoSu	2 Shared Branches	0.3476 ± 0.0017	0.4541 ± 0.0018	74.21 ± 0.12	43.18 ± 0.25
	1V+1A	0.3548 ± 0.0003	0.4698 ± 0.0008	74.82 ± 0.04	44.77 ± 0.08
	2V+2A	0.3548 ± 0.0026	0.4700 ± 0.0029	74.78 ± 0.18	44.86 ± 0.33
	3V+3A	0.3612 ± 0.0055	0.4780 ± 0.0062	75.14 ± 0.34	45.48 ± 0.47
	4V+4A	0.3617 ± 0.0026	0.4783 ± 0.0043	75.19 ± 0.23	45.40 ± 0.08
	2V+1A	0.3618 ± 0.0028	0.4783 ± 0.0031	75.16 ± 0.15	**45.53 ± 0.16**
	1V+2A	**0.3620 ± 0.0013**	**0.4791 ± 0.0023**	**75.21 ± 0.07**	45.33 ± 0.13

A.5.2 Branch Representation Diversity

To examine what independently parameterized temporal branches learn, we analyze their outputs using both UMAP and Centered Kernel Alignment (CKA). For each annotated highlight segment, the video is divided into 2-second clips, and we extract the output representation of every visual and audio branch at each temporal location. All comparisons therefore use temporally aligned representations corresponding to the same video clips.

We first use UMAP to qualitatively visualize the learned feature spaces. For each selected highlight, representations from all branches are jointly projected into two dimensions. As shown in Figure 3, branches form clearly separated clusters across representative baseball and soccer examples, and this separation persists as branch capacity increases from 2V+2A to 4V+4A.

We further quantify this behavior using linear CKA, which measures representational similarity between pairs of branch outputs while being invariant to isotropic scaling and orthogonal transformations. For each branch configuration, CKA is computed between temporally aligned branch representations aggregated across highlight clips, and separately over three temporal windows relative to the annotated highlight: *Local* (clips immediately surrounding the highlight boundary), *Highlight*

(clips within the annotated highlight span), and *Context* (surrounding non-highlight clips). Figure 4 reports the resulting pairwise CKA matrices for configurations ranging from 1V+1A to 4V+4A.

Two consistent patterns emerge across all branch configurations. First, branches within the same modality exhibit high representational similarity (CKA of 0.91-0.99), indicating that visual branches learn correlated transformations of the visual stream, and likewise for audio branches, rather than diverging arbitrarily. Second, and more notably, similarity *between* modalities is substantially lower and varies systematically with temporal context: cross-modal CKA is lowest within Local and Highlight windows (0.67-0.78) and increases within Context windows (0.79-0.86) across every branch configuration we tested. This indicates that visual and audio branches produce more distinct representations specifically near salient moments, while converging toward more similar representations during non-highlight content. This pattern is consistent with the time-varying routing behavior illustrated in Figure 1, where the relative contribution of visual and audio evidence shifts around highlight boundaries, and offers a representational counterpart to that routing behavior: branches do not merely receive different routing weights, but encode more modality-distinctive information exactly where that distinction is most useful for the task.

Together, the UMAP and CKA analyses provide complementary evidence: UMAP reveals the geometric separation of branch representations within individual highlight segments, while CKA quantifies both within- and cross-modality similarity over a larger set of temporally aligned samples and across varying branch capacities. We emphasize that these analyses characterize representational structure rather than assigning a specific semantic function to each individual branch, and that the high within-modality similarity suggests some redundancy among branches of the same modality, consistent with the diminishing returns observed when increasing branch capacity beyond 3V+3A in Table 3.

A.5.3 Qualitative Generalization to Unseen Sports

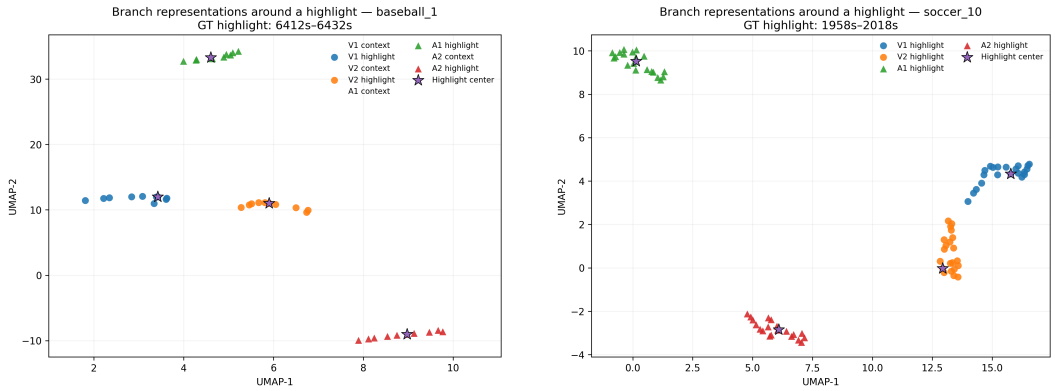
To further examine out-of-domain behavior, we apply the trained model to videos from sports categories that are never observed during training, including e-sports and badminton. These videos are not used for training or model selection and do not contain ground-truth highlight annotations; we therefore use them only for qualitative analysis. The same generic textual query used during validation and inference is applied without any domain-specific adaptation.

Figure 5 presents representative examples together with the predicted saliency and routing weights. In the League of Legends example, the model assigns high saliency to an intense team-fight sequence while placing greater weight on the audio branch. This behavior is consistent with the strong commentary response accompanying the event, where excited speech provides a salient cue despite the substantial visual-domain shift from conventional sports footage. A similar pattern is observed in the badminton example around the championship-winning moment, where heightened crowd reaction and excited commentary coincide with increased reliance on the audio branch and a high predicted highlight score.

These examples suggest that the learned routing mechanism can respond to informative modality cues even in previously unseen sports domains. In particular, the router does not rely exclusively on visual appearance learned from the training benchmarks, but can increase its reliance on audio when acoustic reactions provide stronger evidence of highlight relevance. Since these videos lack ground-truth annotations, we treat these observations as qualitative evidence rather than quantitative measures of generalization.

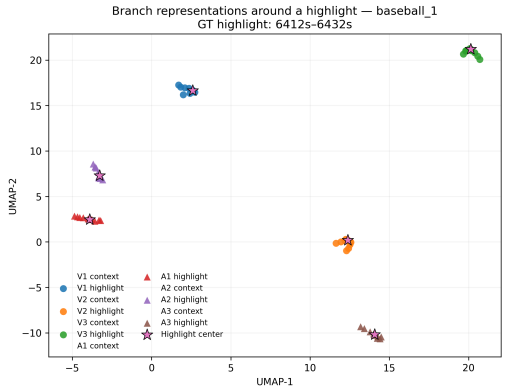
A.5.4 Modality-Masking Analysis

To further examine whether the adaptive router responds to changes in modality availability, we evaluate the model under single-modality masking. As shown in Fig. 6, when the audio modality is masked, the routing weights shift strongly toward the visual expert. Conversely, when the visual modality is masked, substantially greater weight is assigned to the audio expert. This behaviour indicates that the router does not rely on a fixed expert allocation, but instead adjusts the contribution of different experts according to the information available at inference time. These results provide additional evidence that the proposed routing mechanism can adaptively reallocate expert importance when one modality is unavailable.

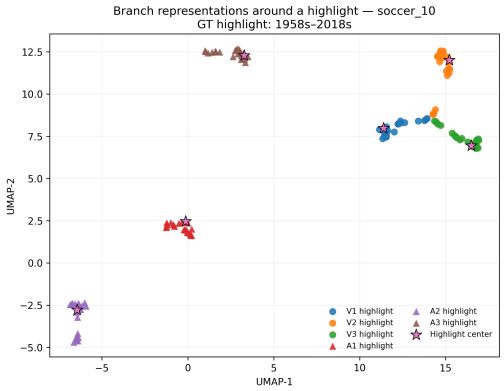


(a) Baseball, 2V+2A

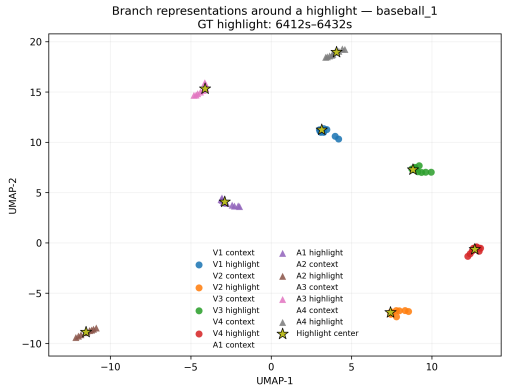
(b) Soccer, 2V+2A



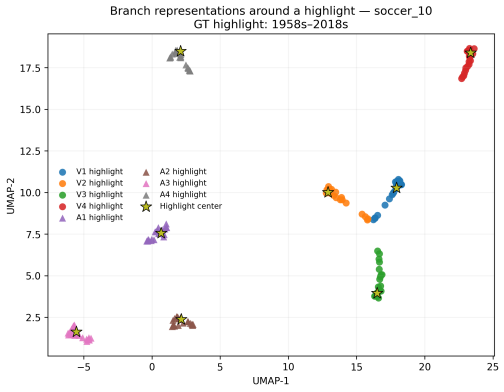
(c) Baseball, 3V+3A



(d) Soccer, 3V+3A

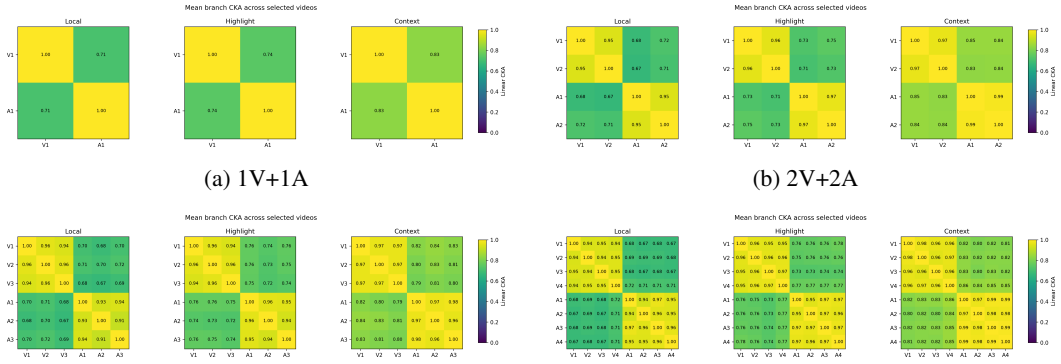


(e) Baseball, 4V+4A



(f) Soccer, 4V+4A

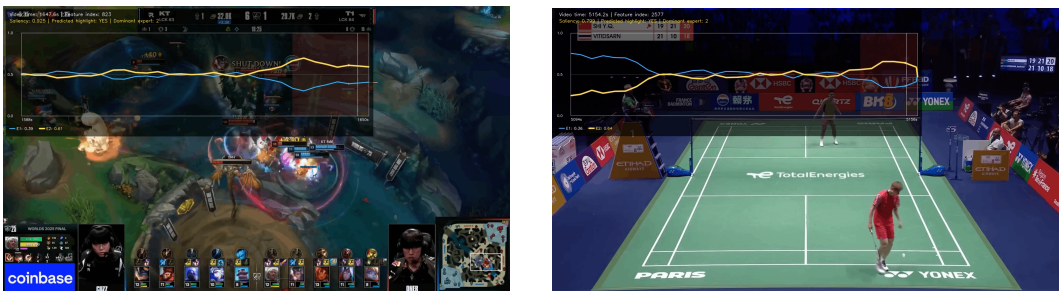
Figure 3: UMAP visualization of temporal branch representations within representative ground-truth highlight segments from baseball and soccer. Each point denotes the representation of a 2-second clip produced by one visual (V) or audio (A) branch, with all branch outputs for a given example jointly projected into the same UMAP space. Across 2V+2A, 3V+3A, and 4V+4A configurations, the branch representations occupy distinguishable regions of the projected space, providing qualitative evidence that independently parameterized branches do not produce identical representations. Stars indicate the representation corresponding to the temporal center of each highlight.



(a) 1V+1A (b) 2V+2A

(c) 3V+3A (d) 4V+4A

Figure 4: Pairwise linear CKA similarity between temporal branch representations under different branch configurations. Higher values indicate greater representational similarity between two branches, while lower values indicate more distinct learned representations.



(a) League of Legends (b) Badminton

Figure 5: Qualitative generalization to previously unseen sports domains. The plots overlay the predicted highlight saliency and temporal routing weights on representative video frames. In the League of Legends example, increased audio-branch weight coincides with an intense team fight accompanied by excited commentary. In the badminton example, the model similarly assigns greater weight to the audio branch around the championship-winning moment, where strong crowd and commentary reactions provide salient acoustic evidence. Both videos are unseen during training and are evaluated without domain-specific adaptation.

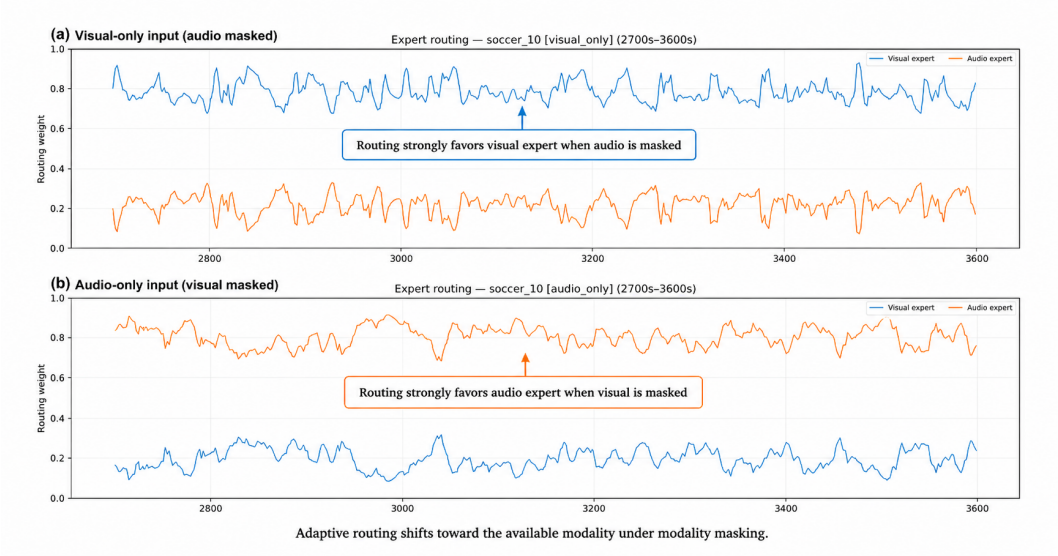


Figure 6: Modality-masking analysis. When audio is masked, the routing weights shift strongly toward the visual expert, whereas masking the visual modality causes the routing weights to shift toward the audio expert. This demonstrates that the adaptive router can reallocate expert contributions according to modality availability.