TripleSumm: Adaptive Triple-Modality Fusion for Video Summarization

Sumin Kim* Hyemin Jeong* Mingu Kang* Yejin Kim Yoori Oh† Joonseok Lee†

SEOUL NATIONAL UNIVERSITY

Project Page!

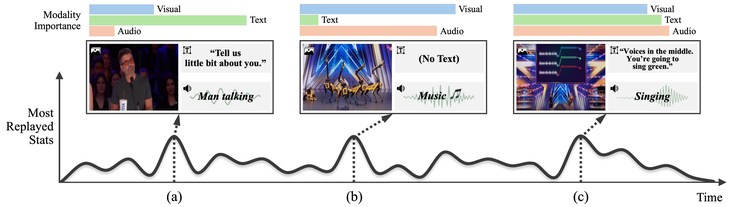
ICLR International Conference On Learning Representations

A. Motivation

"Which modality matters most for a good video summary?" 🤔

Key Challenges

- **Dynamic modality saliency:** The dominant cue shifts over time within the same video.
- **Static fusion limitations:** Existing methods often fuse modalities uniformly or favor visual cues.
- **Temporal + cross-modal reasoning:** Effective summarization requires both long-range temporal modeling and adaptive modality interaction.



Contributions: ***TripleSumm***, ***MoSu***, **and** ***state-of-the-art performance*** **— All Publicly Available!**

B. MoSu Dataset

Existing benchmarks are either too small or lack text/audio modalities.

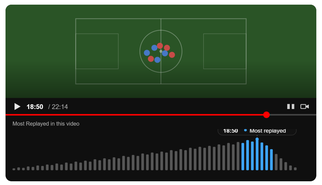
Dataset Construction:

MoSu is the first large-scale dataset to provide **all three modalities** across **diverse categories**.

Frame Importance Annotation:

Ground truth is derived from YouTube **"Most Replayed"** statistics, based on aggregated replay behavior from **50,000+ viewers** per video.

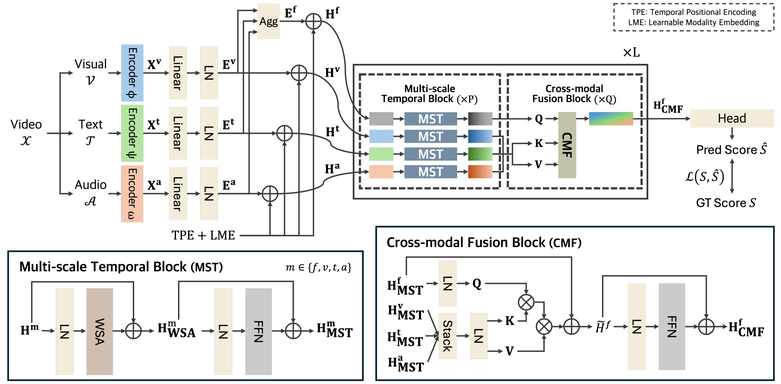
Dataset	Modality			Videos	Duration	Category
	Visual	Text	Audio			
SumMe	✓			25	1.1	–
TVSum	✓			50	3.5	10
Mr.HiSum	✓			31,892	1,788.0	3,509
TL;DW?	✓	✓		12,160	628.5	185
BLiSS	✓	✓		13,303	1,109.0	–
LfVS-P	✓	✓		250K	55,416.6	6,700
MMSum	✓	✓		5,100	1,229.9	170
MoSu	✓	✓	✓	52,678	3,983.7	3,406



YouTube **"Most Replayed"** statistics used for frame-wise importance annotation.

C. TripleSumm Architecture

TripleSumm separates temporal refinement from cross-modal fusion, where a fusion token enables frame-wise adaptive modality selection.



1. **Input:** **Visual**, **text**, and **audio** features are projected into a shared space with a neutral fusion token.
2. **MST:** **Windowed Self-Attn** expands from local to global to capture multi-scale temporal dependencies.
3. **CMF:** The **fusion token** attends to modality-specific tokens at each frame.
4. **Output:** The fused representation is mapped to **frame-level importance scores**.
5. **Training Objective:** The model is trained with an **L2 loss** against ground-truth scores.

D. Quantitative Evaluation

State-of-the-art performance on MoSu with high efficiency!

Method	Modality			$\tau\uparrow$	$\rho\uparrow$	mAP50 ↑	mAP15 ↑	Params ↓	GFLOPs ↓
	V	T	A						
VASNet (Fajtl et al., 2018)	✓			0.151	0.219	64.49	31.05	8.13M	1.99G
PGL-SUM (Apostolidis et al., 2021b)	✓			0.151	0.218	64.97	30.63	5.31M	1.21G
CSTA (Son et al., 2024)	✓			0.291	0.398	71.77	40.65	10.56M	11.37G
A2Summ (He et al., 2023)	✓	✓		0.181	0.257	66.48	35.70	2.48M	1.35G
SSPVS (Li et al., 2023)	✓	✓		0.190	0.271	66.10	32.65	112.81M	43.64G
Joint-VA (Badamdorj et al., 2021)	✓		✓	0.190	0.272	65.68	32.25	4.21M	1.63G
UMT (Liu et al., 2022)	✓		✓	0.239	0.334	68.83	36.73	4.66M	1.39G
CFSum (Guo et al., 2025)	✓	✓	✓	0.277	0.374	70.97	38.20	19.83M	8.52G
TripleSumm (Ours)	✓	✓	✓	**0.351**	**0.472**	**74.72**	**44.42**	**1.37M**	**0.97G**

(a) Mr. HiSum

Method	TVT			
	τ	ρ	mAP50	mAP15
VASNet	0.069	0.102	58.69	25.28
PGL-SUM	0.097	0.141	61.60	27.45
CSTA	0.128	0.185	63.38	30.42
A2Summ	0.121	0.172	63.20	32.34
SSPVS	0.078	0.113	59.48	26.35
Joint-VA	0.161	0.231	65.88	35.23
UMT	0.178	0.253	66.81	35.65
Ours (Visual)	0.187	0.258	67.16	35.57
Ours (Full)	**0.258**	**0.352**	**70.72**	**40.88**

(b) SumMe

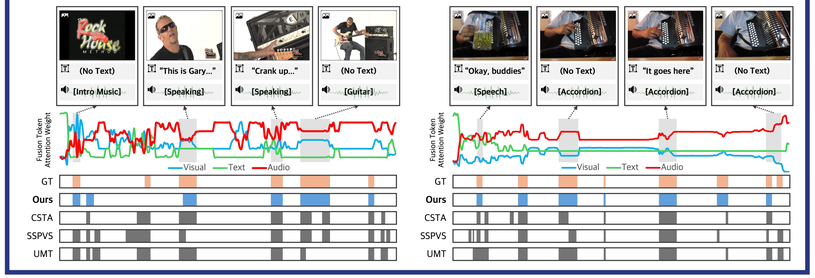
Method	TVT		TV	
	τ	ρ	τ	ρ
VASNet	0.089	0.099	0.160	0.170
PGL-SUM	0.104	0.116	0.192	0.213
CSTA	0.133	0.148	0.246	0.274
A2Summ	0.088	0.096	0.108	0.129
SSPVS	0.142	0.157	0.192	0.257
Joint-VA	0.117	0.129	0.230	0.256
UMT	0.148	0.165	0.241	0.268
Ours (Full)	0.162	0.187	0.265	0.296
Ours (MoSu)	**0.172**	**0.192**	**0.282**	**0.314**

(c) TVSum

Method	TVT		TV	
	τ	ρ	τ	ρ
VASNet	0.153	0.205	0.160	0.170
PGL-SUM	0.141	0.186	0.157	0.206
CSTA	0.168	0.221	0.194	0.255
A2Summ	0.157	0.206	0.137	0.165
SSPVS	0.171	0.226	0.181	0.238
Joint-VA	0.142	0.188	0.166	0.220
UMT	0.144	0.189	0.179	0.235
Ours (Full)	0.198	0.259	0.211	0.275
Ours (MoSu)	**0.200**	**0.262**	**0.217**	**0.282**

E. Qualitative Demonstration

TripleSumm *dynamically shifts attention* across visual, text, and audio cues, producing summaries more aligned with the ground truth.



F. Additional Insights

Zero-shot results on 70.4-min videos suggest long-form video summarization as a promising next direction.

Method	τ	ρ	mAP50	mAP15
Random	0.000	0.000	50.59	16.18
VASNet	0.024	0.036	54.15	18.95
PGL-SUM	0.024	0.035	54.29	17.55
CSTA	0.083	0.123	58.09	22.26
A2Summ	0.042	0.062	54.04	18.66
SSPVS	0.033	0.048	53.60	18.45
Joint-VA	0.052	0.077	53.99	19.90
UMT	0.066	0.097	56.05	23.10
CFSum	0.061	0.089	56.32	20.59
TripleSumm	**0.128**	**0.189**	**59.70**	**23.27**

Scaling training data improves performance, motivating large-scale video summarization research.