FSGM-Sum: Factorized Spatio-Temporal Gating Mixer for Efficient Video Summarization

Abstract

Video summarization seeks to produce concise yet informative representations of long videos by selecting segments aligned with human preferences. However, this task remains challenging due to annotation subjectivity and the discrepancy between continuous score regression during training and discrete selection at inference. Existing approaches often rely on dense attention mechanisms for long-range dependency modeling or employ generative frameworks to capture subjectivity, both of which introduce significant computational overhead and limited robustness under scarce supervision. To address this, we propose FSGM-Sum, a lightweight Factorized Spatio-Temporal Gating Mixer for single-pass frame importance prediction. Specifically, our method decomposes importance estimation into complementary temporal and channel saliency signals, where temporal dynamics are modeled via multi-dilation depthwise mixing and channel relevance is captured through global pooling with a compact gating network. These signals are fused through an additive factorization to form a structured gating mechanism, coupled with a temperature-calibrated scoring head for stable optimization and reliable score distributions. This design enables efficient modeling of spatio-temporal dependencies without constructing dense attention maps. We evaluate our approach on standard benchmarks, including SumMe and TVSum under rank-based evaluation protocols. Our method achieves competitive or superior alignment with human annotations while maintaining a significantly simpler and faster inference pipeline. Overall, FSGM-Sum demonstrates that factorized spatio-temporal gating provides an effective and practical alternative to attention-heavy and generative video summarization models. Code is available at: https://github.com/SMRaza1009/FSGM-Sum.git

Keywords

Deep Learning, Factorized Spatio-Temporal Gating Mixer, Lightweight Architecture.

ACM Reference Format:

. 2026. FSGM-Sum: Factorized Spatio-Temporal Gating Mixer for Efficient Video Summarization. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Video summarization aims to condense a long video into a compact representation that preserves the most informative events for efficient browsing, retrieval, and downstream understanding. As large-scale user-generated and web video content continues to grow, supervised video summarization remains an important problem in multimedia analysis, with applications in surveillance, media curation, and recommendation. In standard benchmarks such as SumMe [**?**] and TVSum [21], the task is commonly formulated as learning frame importance scores from human annotations, followed by segment selection under a summary length budget.

Despite substantial progress, supervised video summarization still faces two persistent challenges. First, long-range temporal dependency modeling must be handled under small benchmark sizes, where heavy attention-based models can be computationally expensive and prone to overfitting. Second, the learning and evaluation stages are not perfectly aligned: models are typically optimized with continuous frame-level objectives, whereas final summaries are produced through discrete segmentation and budget-constrained shot selection. This mismatch is especially important when small score perturbations can alter downstream keyshot selection.

Recent methods have explored recurrent models, self-attention, spatiotemporal attention, and generative formulations to improve temporal reasoning and capture annotation ambiguity [**? ? ? ?**]. While effective, these approaches often increase architectural or optimization complexity. Dense attention scales poorly with video length, proposal-based or stochastic designs complicate training and inference, and existing lightweight alternatives do not explicitly separate temporal saliency from feature channel saliency while maintaining a scoring mechanism tailored to downstream summary decoding. As a result, there remains a need for a summarization model that is both computationally economical and structurally better aligned with the way summaries are ultimately generated and evaluated.

To address this need, we propose **FSGM-Sum**, a **F**actorized **S**patio-**T**emporal **G**ating **M**ixer for efficient video summarization. Instead of constructing dense pairwise attention maps, FSGM-Sum models frame importance through two complementary gating signals. A temporal gating branch captures *when* salient events occur using multi-dilation depthwise temporal mixing, while a channel gating branch captures *which feature dimensions* are most informative through global pooling and a lightweight gating network. The two branches are fused by an additive factorized construction, yielding a structured gating field that modulates video representations across time and feature channels with linear complexity in sequence length and feature width. To better support stable downstream segment selection, we further use a calibrated scoring head, together with lightweight training regularization, to produce sharper and more reliable importance estimates.

Extensive experiments on SumMe and TVSum show that FSGM-Sum provides a strong accuracy efficiency tradeoff with a simple single-pass inference pipeline. The results indicate that factorized temporal and feature-channel gating can serve as an effective alternative to heavier attention-based summarization architectures.

In summary, our contributions are as follows:

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, Washington, DC, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

https://doi.org/10.1145/nnnnnnn.nnnnnnn

Unpublished working draft. Not for distribution.

(1) We propose **FSGM-Sum**, a lightweight video summarization architecture that replaces dense self attention with a factorized temporal and feature-channel gating mixer for efficient long range dependency modeling.
(2) We introduce a structured gating formulation that separates temporal saliency and feature-channel saliency, and fuses them through an additive outer-factorized mechanism to modulate frame representations without quadratic attention cost.
(3) We improve learning to evaluation alignment through a calibrated scoring design and lightweight training regularization, leading to more stable importance prediction for downstream budget constrained summary generation.
(4) We demonstrate strong performance on SumMe and TVSum, showing that FSGM-Sum achieves competitive summarization quality with a simpler and more efficient inference pipeline.

2 Related Work

Deterministic supervised summarization. Supervised video summarization has largely been developed under benchmark settings such as SumMe and TVSum, where the goal is to predict frame or segment importance from human annotations under a fixed summary-length budget. Early deep approaches mainly formulated the task as sequence prediction, including recurrent models such as vsLSTM and dppLSTM, attention-based encoder-decoder designs, and fully convolutional sequence labeling frameworks. Later, VASNet showed that self-attention can model long-range dependencies without recurrence, while DSNet reformulated summarization as temporal interest detection with segment-level proposals. These methods established the dominant deterministic predict-and-select pipeline and remain strong supervised baselines. However, they usually learn a single importance trajectory for each video and only indirectly reflect the fact that multiple human summaries may be valid for the same content. [5, 8, 19, 21, 24, 25, 28**?**]

Attention-based and structured dependency modeling. To improve contextual reasoning, subsequent work explored richer dependency modeling beyond plain sequential encoders. PGL-SUM combined global and local self-attention with positional encoding to capture both long-range and neighborhood-level relations. Other works investigated more structured interaction mechanisms to encode temporal context without relying on purely dense pairwise attention. In particular, CSTA represented a video as an image-like frame-feature map and used convolutional processing to model inter-frame and intra-feature interactions jointly, while graph-based formulations such as VideoSAGE encoded sparse relations among frames for more structured dependency reasoning. These methods strengthen contextual modeling, but many still rely on comparatively heavy interaction modules or structured relation learning over relatively small supervised benchmarks. [2, 4, 20]

Ambiguity-aware and generative summarization. Because video summarization is inherently subjective, another line of research moves beyond deterministic regression and explicitly models uncertainty, diversity, or one-to-many summary mappings. SUM-GAN and its later variants introduced adversarial and reconstruction-based formulations to learn representative summaries without relying only on direct score regression. More recently, diffusion-based approaches such as SummDiff argued that averaging annotations into a single deterministic target can obscure valid alternative summaries, and instead modeled summarization as a conditional generative process. These methods are appealing because they better reflect annotator ambiguity, but they usually introduce additional architectural, training, or inference complexity compared with single-pass supervised scoring models. [1, 12, 15]

Multimodal and semantic extensions. Recent studies have expanded video summarization beyond purely visual frame scoring by incorporating language supervision, cross-modal alignment, and stronger semantic priors. A2Summ aligned video and text through multimodal attention and contrastive learning, while joint summarization and moment localization frameworks transferred supervision from related temporal understanding tasks to improve summary learning. More recent approaches have explored large-scale pretraining or large language model guidance to inject richer semantic knowledge into importance estimation. These directions suggest that semantic supervision can improve summary quality, but they also move the problem beyond the standard visual-only supervised setting commonly used for direct comparison on SumMe and TVSum. [3, 6, 9, 13]

Lightweight and efficient summarization models. Efficiency remains a central concern because long videos make dense attention increasingly expensive in both memory and computation. Recent efforts therefore seek lighter alternatives that preserve competitive performance while simplifying token mixing and contextual modeling. Efficient DSNet explicitly studies cheaper replacements for self-attention, while methods such as CSTA and VideoSAGE show that convolutional or sparse-graph formulations can reduce the burden of dense pairwise reasoning. Our method is most closely related to this line of work. However, instead of replacing attention with another generic mixer, FSGM-Sum explicitly factorizes importance estimation into temporal saliency and feature-channel saliency, and couples this with calibrated scoring for stable single-pass importance prediction under downstream budget-constrained summary decoding. [4, 18, 20]

3 Method

3.1 Overview

Figure 1 illustrates the overall flow of **FSGM-Sum**. The proposed model replaces backbone-heavy attention generation with a structured gating mechanism that is tailored to supervised video summarization. Instead of learning a dense spatiotemporal importance field through a generic CNN over feature maps, the model decomposes importance into two complementary signals: a *temporal saliency* term that models *when* informative events occur, and a *channel saliency* term that models *which* feature dimensions carry discriminative evidence. These two signals are fused through an outer additive construction to produce a full frame–channel attention logit map, which is then calibrated along both temporal and channel axes before modulating the value stream. The resulting representation is scored by a temperature-calibrated predictor to produce

Unpublished working draft. Not for distribution.

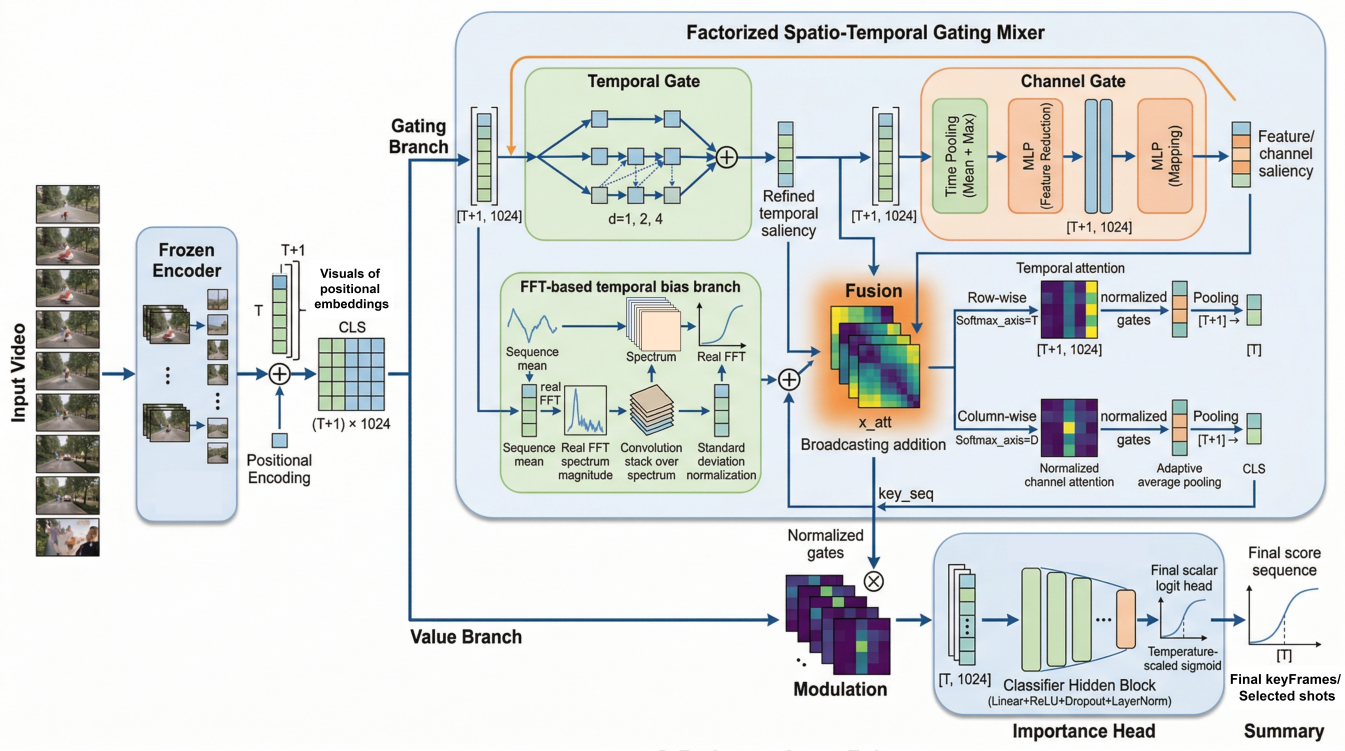


Figure 1: Overview of FSGM-Sum. Frame features from a frozen CNN (GoogleNet) with a CLS token are positionally encoded and projected into value and gating branches. A *Factorized Gating Block* decomposes saliency into temporal (multi-dilation depthwise mixing) and channel (global pooling + MLP) components, which are fused and applied as dual-axis soft gates. A lightweight mixing module and temperature-scaled sigmoid head produce frame-level importance scores, avoiding dense attention via efficient factorized gating.

frame-level importance estimates. This design is lightweight by construction, but more importantly, it is structurally aligned with video summarization: frame selection depends jointly on temporal dynamics and feature-level saliency, and the proposed factorization models these two sources explicitly rather than leaving them implicit inside a generic backbone.

3.2 Problem formulation and notation

Let a video be represented by a sequence of sampled frame descriptors

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^\top \in \mathbb{R}^{T \times D}, \tag{1}$$

where T is the number of sampled timesteps and D is the feature dimension. In our implementation, $D = 1024$, corresponding to pre-extracted frame features. The model predicts a sequence of frame-level importance scores

$$\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_T]^\top \in [0, 1]^T, \tag{2}$$

which are later expanded to the original frame timeline and decoded into a binary summary under a length budget.

For compatibility with the outer model shell, the feature sequence is first replicated across three channels and augmented with a learnable CLS token. Denoting channel replication by $\mathcal{R}(\cdot)$, we form

$$\bar{\mathbf{X}} = \mathcal{R}(\mathbf{X}) \in \mathbb{R}^{1 \times 3 \times T \times D}, \qquad \tilde{\mathbf{X}} \in \mathbb{R}^{1 \times 3 \times (T+1) \times D}, \tag{3}$$

where $\tilde{\mathbf{X}}$ is obtained by prepending the CLS token along the temporal axis. This extra token provides a global context slot during mixing, after which the sequence is pooled back to length T.

3.3 Overall pipeline

The model applies separate key and value embeddings to the CLS-augmented sequence, and uses the key stream to construct a structured spatiotemporal gating field. Specifically, let

$$\mathbf{K} = \phi_K(\tilde{\mathbf{X}}), \qquad \mathbf{V} = \phi_V(\tilde{\mathbf{X}}), \tag{4}$$

where ϕ_K and ϕ_V are linear embeddings applied along the feature dimension. From $\mathbf{K}$, we extract a single sequence

$$\mathbf{H} \in \mathbb{R}^{(T+1) \times D}, \tag{5}$$

which serves as the input to the Factorized Spatiotemporal Gate Mixer (FSGM). The FSGM produces a 2D logit map over timesteps and channels, combines it with the key sequence through a residual key connection, augments it with positional encoding, and normalizes it along both temporal and channel axes. The resulting pair of calibrated attention maps modulates the value stream, yielding mixed frame representations that are mapped to scalar logits and

converted to importance scores through a learnable temperature-calibrated sigmoid. The architectural core is therefore not a generic feature extractor, but a structured gate constructor that explicitly separates temporal and channel saliency before recombining them.

3.4 Factorized Spatiotemporal Gate Mixer

3.4.1 Temporal saliency branch. The temporal branch estimates framewise saliency by explicitly modeling multi-scale temporal variation. Starting from the key sequence $\mathbf{H}$, we first project each token from dimension D to a hidden width H:

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_t + \mathbf{b}_t \in \mathbb{R}^{(T+1)\times H}. \tag{6}$$

This projection reduces the cost of temporal mixing while preserving the information needed for saliency estimation.

We then apply three parallel depthwise-separable temporal convolution blocks with dilation rates $d \in \{1, 2, 4\}$. Let $\mathcal{B}_d(\cdot)$ denote one such block. The aggregated temporal representation is

$$\mathbf{M} = \frac{1}{3}\left(\mathcal{B}_1(\mathbf{Q}) + \mathcal{B}_2(\mathbf{Q}) + \mathcal{B}_4(\mathbf{Q})\right) \in \mathbb{R}^{(T+1)\times H}. \tag{7}$$

The three branches provide receptive fields of different effective scales, allowing the model to respond to abrupt local changes and longer-range temporal transitions within a single lightweight mixer. This is a deliberate departure from implicit temporal modeling in backbone CNNs: summarization requires identifying salient temporal transitions, and dilated temporal convolutions expose this structure directly.

A 1×1 temporal head produces the framewise logits

$$\mathbf{a}^{(t)}_{\text{local}} = \psi_t(\mathbf{M}) \in \mathbb{R}^{T+1}. \tag{8}$$

These logits capture relative temporal importance conditioned on local and mid-range context, but do not yet incorporate a video-level prior on temporal variation.

3.4.2 FFT spectral bias. To inject a global prior on temporal dynamics, we compute a scalar spectral bias from the input sequence. We first collapse the channel dimension by averaging over features:

$$s_i = \frac{1}{D}\sum_{j=1}^{D} H_{i,j}, \qquad \mathbf{s} \in \mathbb{R}^{T+1}. \tag{9}$$

We then compute the real Fourier spectrum,

$$\mathbf{r} = \text{rFFT}(\mathbf{s}), \qquad \mathbf{m} = |\mathbf{r}| \in \mathbb{R}^{F}, \tag{10}$$

where $F = \lfloor (T+1)/2 \rfloor + 1$. The spectrum is summarized into a scalar statistic

$$\eta = \frac{1}{F}\sum_{f=1}^{F} m_f, \tag{11}$$

which is passed through a lightweight MLP $g_{\text{fft}}(\cdot)$ to obtain a global bias term

$$b_{\text{fft}} = g_{\text{fft}}(\eta) \in \mathbb{R}. \tag{12}$$

The final temporal logits are

$$\mathbf{a}^{(t)} = \mathbf{a}^{(t)}_{\text{local}} + b_{\text{fft}}\mathbf{1}_{T+1}. \tag{13}$$

This term acts as a frequency-aware temporal variation prior. Importantly, it does not replace temporal modeling; rather, it calibrates the baseline saliency level according to the global spectral energy of the sequence.

3.4.3 Channel saliency branch. Temporal saliency alone is insufficient for summarization because not all feature dimensions are equally informative. The channel branch estimates which descriptor channels consistently or sharply respond to salient content. We compute temporal mean and max pooling over the key sequence:

$$\boldsymbol{\mu} = \frac{1}{T+1}\sum_{i=1}^{T+1} \mathbf{H}_{i,:} \in \mathbb{R}^{D}, \qquad \boldsymbol{\nu} = \max_{1\le i\le T+1} \mathbf{H}_{i,:} \in \mathbb{R}^{D}, \tag{14}$$

and concatenate them to form

$$\mathbf{u} = [\boldsymbol{\mu}; \boldsymbol{\nu}] \in \mathbb{R}^{2D}. \tag{15}$$

A bottleneck MLP then produces channel logits

$$\mathbf{a}^{(c)} = \psi_c(\mathbf{u}) \in \mathbb{R}^{D}. \tag{16}$$

Mean pooling captures persistent evidence across the video, whereas max pooling preserves peak responses that may correspond to rare yet highly informative events. Their combination yields an explicit channel-saliency estimate.

3.4.4 Outer additive fusion. The core modeling decision is to combine temporal and channel saliency through an outer additive construction:

$$\mathbf{A}_{\text{logit}} = \mathbf{a}^{(t)}\mathbf{1}_D^{\top} + \mathbf{1}_{T+1}(\mathbf{a}^{(c)})^{\top} \in \mathbb{R}^{(T+1)\times D}. \tag{17}$$

This fusion imposes an interpretable decomposition in which each entry of the gate logit map is the sum of a timestep-specific term and a channel-specific term. A frame-channel pair attains high logit only when it is supported by both temporal importance and feature saliency. The resulting structure is aligned with the semantics of summarization while avoiding the need to learn a dense attention generator through a heavy CNN.

3.4.5 Key residual, positional encoding, and dual-axis calibration. The fused logit map is combined with the key sequence through a residual key connection:

$$\mathbf{G} = \mathbf{A}_{\text{logit}} + \mathbf{H}. \tag{18}$$

This skip preserves content-dependent structure from the key stream and stabilizes optimization by anchoring the factorized gate to the input representation. A lightweight normalization is then applied, followed by addition of fixed sinusoidal positional encoding $\mathbf{P}$:

$$\mathbf{G} \leftarrow \text{Norm}(\mathbf{G}) + \mathbf{P}. \tag{19}$$

The positional term restores explicit ordering information after factorized gating, which is important because summarization depends on temporal placement as well as saliency magnitude.

Instead of collapsing $\mathbf{G}$ through a single normalization, we calibrate it along both axes:

$$\mathbf{A}^{(t)} = \text{Softmax}(\mathbf{G}; 0), \qquad \mathbf{A}^{(c)} = \text{Softmax}(\mathbf{G}; 1), \tag{20}$$

where $\text{Softmax}(\cdot; 0)$ normalizes over timesteps for each channel, and $\text{Softmax}(\cdot; 1)$ normalizes over channels for each timestep. The temporal normalization answers which timesteps dominate within each feature dimension; the channel normalization answers which features dominate within each timestep.

Unpublished working draft. Not for distribution.

3.4.6 Value modulation and mixed representation. The value stream is modulated by both calibrated attention maps:

$$\tilde{\mathbf{V}} = \mathbf{V} \odot \mathbf{A}^{(t)} + \mathbf{V} \odot \mathbf{A}^{(c)}, \tag{21}$$

where $\odot$ denotes elementwise multiplication. The first term emphasizes temporally dominant evidence, and the second emphasizes channel-dominant evidence; their sum retains both without collapsing them into a single gate.

Because the input was expanded with a CLS token, the mixed sequence has length $T + 1$. We pool it back to the original sampled length,

$$\mathbf{Z} \in \mathbb{R}^{T \times D}, \tag{22}$$

yielding one feature vector per sampled frame.

3.5 Temperature-calibrated importance prediction

Each mixed frame representation $\mathbf{z}_t \in \mathbb{R}^D$ is mapped to a scalar logit

$$\ell_t = f_{\text{cls}}(\mathbf{z}_t), \tag{23}$$

where $f_{\text{cls}}(\cdot)$ denotes the framewise classifier. The final importance score is

$$\hat{y}_t = \sigma\left(\frac{\ell_t}{\tau}\right), \qquad \tau > 0. \tag{24}$$

Here τ is a learnable temperature parameter constrained to a valid positive range. This component is not a cosmetic reparameterization of the sigmoid: summarization is evaluated under budgeted shot selection, so score sharpness directly affects downstream peak selection and segment ranking.

3.6 Training objective

The model is trained on sampled frame-level supervision using a regression objective augmented with auxiliary constraints that target ranking fidelity, shot coherence, and score calibration. Let $\mathbf{y} \in \mathbb{R}^T$ denote the ground-truth target aligned to sampled timesteps. The primary objective is mean squared error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2. \tag{25}$$

This term directly supervises frame-level importance prediction and serves as the anchor objective for training.

To preserve relative ordering between frames, we add a pairwise ranking term

$$\mathcal{L}_{\text{rank}}, \tag{26}$$

which encourages score differences to be consistent with target ordering. To reduce temporal noise within semantically coherent segments, we introduce a shot-level smoothness term

$$\mathcal{L}_{\text{shot-smooth}}, \tag{27}$$

which penalizes excessive variation within the same shot. We further include a shotwise alignment term

$$\mathcal{L}_{\text{shotwise-MSE}}, \tag{28}$$

which matches predicted and target shot-level trends. Finally, to stabilize peak formation under dual-axis attention calibration, we use

$$\mathcal{L}_{\text{dual-softmax}}, \tag{29}$$

which regularizes the calibrated output distribution so that strong selections remain stable under budgeted decoding.

The full objective is

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} &+ \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} \\ &+ \lambda_{\text{smooth}}\mathcal{L}_{\text{shot-smooth}} + \alpha_{\text{align}}\mathcal{L}_{\text{shotwise-MSE}} \\ &+ \lambda_{\text{dual}}\mathcal{L}_{\text{dual-softmax}}.\end{aligned} \tag{30}$$

Together, these terms supervise the model at complementary levels: framewise regression for direct score accuracy, ranking for relative saliency consistency, shot coherence for temporal stability, shotwise alignment for segment-level fidelity, and dual-softmax regularization for calibrated peak selection.

3.7 Inference and summary decoding

At inference time, the model outputs frame-level importance scores on the sampled timeline. These scores are expanded back to the original frame sequence by piecewise-constant assignment between consecutive sampling indices. Let $\{p_t\}_{t=1}^{T}$ be the sampling positions in the original video of length N. We define the frame-expanded score sequence $\mathbf{f} \in \mathbb{R}^N$ as

$$f_n = \hat{y}_t \quad \text{for} \quad n \in [p_t, p_{t+1}), \qquad t = 1, \ldots, T-1, \tag{31}$$

and $f_n = \hat{y}_T$ for $n \in [p_T, N)$. This expansion preserves the sampled prediction while recovering frame-level resolution for segment-based decoding.

Given shot boundaries from temporal segmentation, each shot k is assigned a length w_k and a value v_k:

$$w_k = c_k^{\text{end}} - c_k^{\text{start}} + 1, \qquad v_k = \frac{1}{w_k}\sum_{n=c_k^{\text{start}}}^{c_k^{\text{end}}} f_n. \tag{32}$$

The final summary is obtained by solving a 0/1 knapsack problem under a budget ρN:

$$\max_{\mathbf{u}\in\{0,1\}^M} \sum_{k=1}^{M} u_k v_k \quad \text{s.t.} \quad \sum_{k=1}^{M} u_k w_k \le \rho N. \tag{33}$$

The selected shots are concatenated to form the binary summary.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets. We evaluate on two standard benchmarks for supervised video summarization, TVSum [21] and SumMe [?]. TVSum provides multi-annotator importance annotations, while SumMe provides multiple user summaries per video. We follow the official dataset metadata and evaluation resources (temporal segmentation and change points) required for keyshot-based evaluation on SumMe.

4.1.2 Evaluation Protocol. We follow the standard five-split protocol used in prior work and report the average performance across splits. For each split, the model is trained on the training set and evaluated on the corresponding test set, and the final result is obtained by averaging over all five splits. This setting matches the evaluation regime used by recent rank-correlation based approaches and enables direct comparability [2, 12, 20].

Unpublished working draft. Not for distribution.

Method	SumMe		TVSum	
	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
Human	0.205	0.213	0.177	0.204
A2Summ [7]	0.088	0.096	0.157	0.206
VASNet [5]	0.089	0.099	0.153	0.205
PGL-SUM [2]	0.104	0.116	0.141	0.186
CSTA [20]	0.108	0.120	0.168	0.221
SummDiff [12]	0.133	0.148	0.173	0.226
FSGM-Sum (ours)	**0.220**	**0.246**	**0.229**	**0.298**

Table 1: Comparison of models trained with the TVT (train/val/test) split on SumMe [?] and TVSum [21]. The best results are shown in bold.

4.1.3 Evaluation Metrics. Following recent evaluation practice, we report rank order statistics rather than F1 based summary overlap. Specifically, we measure Kendall's τ [11] and Spearman's ρ [29] between predicted importance scores and human supervision. For TVSum, correlations are computed per video by aggregating across annotators, and then averaged across videos. For SumMe, correlations are computed against the aggregated user summary derived from multiple user annotations. This choice avoids the known sensitivity of F1 to segmentation granularity and summary fragmentation, which has been emphasized in recent literature [20, 22].

4.1.4 Keyshot Summary Construction (SumMe). For SumMe, we convert predicted frame-level scores into a binary keyshot summary under a fixed length budget of 15% of the video duration. We follow the standard segmentation-and-knapsack pipeline: scores are expanded to the original timeline using the sampling indices, shot-level values are computed by averaging within each change-point segment, and a 0/1 knapsack problem is solved to select shots under the budget.

4.1.5 Training and Inference Setup. FSGM-Sum is trained and evaluated using the same input feature representation and split protocol across both datasets. Unless otherwise stated, we use a single configuration for both TVSum and SumMe. The model is optimized with Adam for 100 epochs with learning rate 10^{-3} and weight decay 10^{-7}.

5 Results and Discussion

Tables 1–3 report quantitative comparisons on SumMe and TVSum using Kendall's τ and Spearman's ρ under the standard evaluation protocols. Overall, the results indicate that **FSGM-Sum** consistently improves rank agreement with human supervision while preserving the conventional segmentation-plus-knapsack decoding used at test time as shown in Fig 2. These findings support the central claim that *factorized spatio-temporal gating* can replace dense attention with a substantially lighter token-mixing mechanism, while still producing calibrated importance scores that remain effective after discretization into keyshots.

TVT setting. Under the TVT (Table 1) split, FSGM-Sum achieves the strongest performance among all listed methods on both datasets. On **SumMe**, FSGM-Sum reaches $(\tau, \rho) = (0.220, 0.246)$, providing a large margin over deterministic scorers such as CSTA $(0.108, 0.120)$ and also surpassing the diffusion-based SummDiff baseline $(0.133, 0.148)$. On **TVSum**, FSGM-Sum attains $(\tau, \rho) = (0.229, 0.298)$, again improving over CSTA $(0.168, 0.221)$ and SummDiff $(0.173, 0.226)$. The consistent gains across both τ and ρ suggest that the proposed factorized gating produces not only locally correct pairwise ordering (captured by τ) but also stronger global rank consistency (captured by ρ), which is important when scores are later thresholded and pooled by segments.

Method	SumMe		TVSum	
	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
Human	0.205	0.213	0.177	0.204
DSNet-AF [27]	0.037	0.046	0.113	0.138
DSNet-AB [27]	0.051	0.059	0.108	0.129
SUM-GAN [16]	0.049	0.066	0.024	0.031
AC-SUM-GAN [?]	0.102	0.088	0.031	0.041
CLIP-It [17]	-	-	0.108	0.147
iPTNet [10]	0.101	0.119	0.134	0.163
A2Summ [7]	0.108	0.129	0.137	0.165
VASNet [5]	0.160	0.170	0.160	0.170
AAAM [22]	-	-	0.169	0.223
MAAM [22]	-	-	0.179	0.236
VSS-Net [26]	-	-	0.190	0.249
DMASum [23]	0.063	0.089	0.203	0.267
SSPVS [14]	0.192	0.257	0.181	0.238
CSTA [20]	0.246	0.274	0.194	0.255
SummDiff [12]	**0.256**	**0.285**	0.195	0.255
FSGM-Sum (ours)	0.252	0.284	**0.197**	**0.258**

Table 2: Comparison under the standard 5-split protocol on SumMe [?] and TVSum [21]. Kendall's τ and Spearman's ρ are computed using the evaluation pipeline in our codebase. For SumMe, correlations are computed after segmentation-and-knapsack decoding under a 15% budget.

Split	SumMe		TVSum	
	τ	ρ	τ	ρ
Split0	0.220	0.248	0.172	0.226
Split1	0.190	0.210	0.215	0.280
Split2	0.204	0.228	0.204	0.267
Split3	0.299	0.333	0.190	0.251
Split4	0.236	0.263	0.203	0.264
Average	0.252	0.284	**0.197**	**0.258**

Table 3: Split-wise performance of FSGM-Sum on SumMe [?] and TVSum [21] under the standard five-split protocol. The final row reports the average across splits.

Five-split protocol. In Table 2, a similar trend is observed under the standard five-split evaluation. FSGM-Sum achieves $(\tau, \rho) = (0.252, 0.284)$ on **SumMe** and $(0.197, 0.258)$ on **TVSum**. On SumMe, the proposed method improves over strong deterministic baselines such as CSTA $(0.246, 0.274)$ and remains highly competitive with SummDiff $(0.256, 0.285)$, trailing it only marginally. On TVSum, FSGM-Sum provides a measurable gain over both CSTA $(0.194, 0.255)$ and SummDiff $(0.195, 0.255)$, indicating that the benefits of factorized temporal–channel saliency generalize beyond a single split configuration. Notably, the improvements are more pronounced in Spearman's ρ, which emphasizes global ordering; this aligns with the design goal of producing a stable and discriminative score distribution that is less sensitive to downstream selection.

Unpublished working draft. Not for distribution.

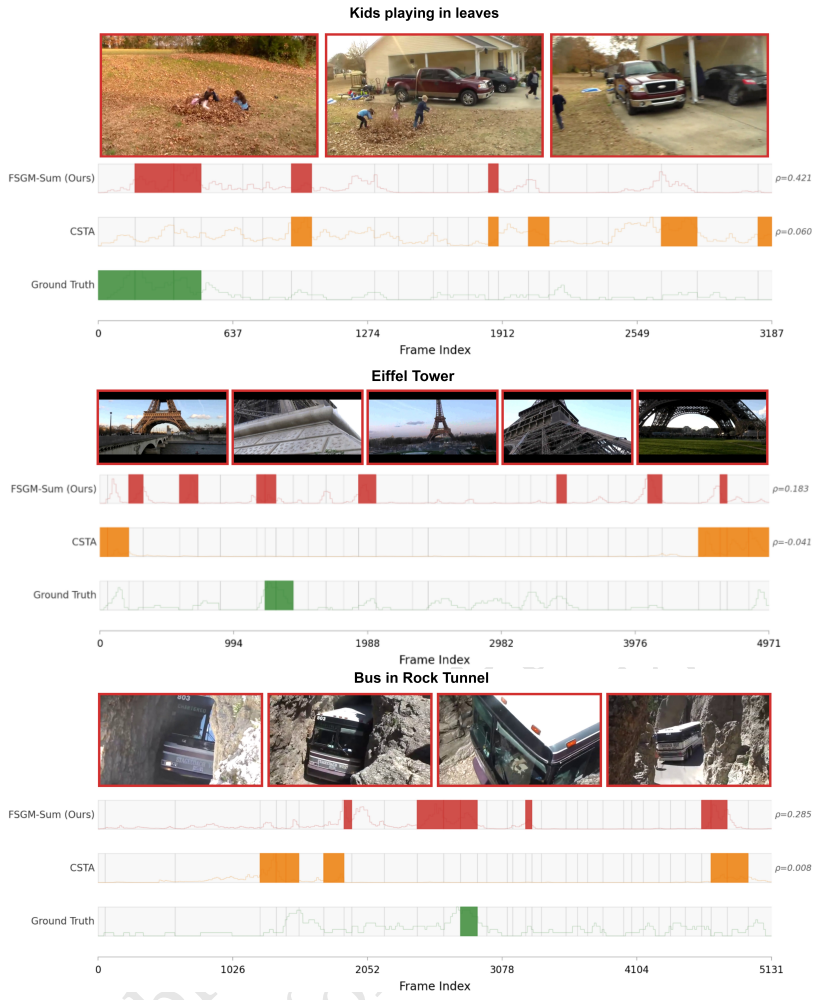


Figure 2: Qualitative comparison of video summaries on SumMe. Each row shows shot-level selections for FSGM-Sum (ours, red), CSTA (orange), and ground truth (green), where filled rectangles are selected shots and the overlaid curve indicates predicted importance scores. Spearman ρ is annotated on the right. FSGM-Sum achieves consistently higher correlation with human annotations across all four videos.

Split-wise robustness and variance. Table 3 highlights both robustness and remaining sensitivity across splits. On **SumMe**, performance ranges from $\tau = 0.190$ (Split1) to $\tau = 0.299$ (Split3), and on **TVSum**, from $\tau = 0.172$ (Split0) to $\tau = 0.215$ (Split1). The spread is expected given the small scale of these benchmarks and the high inter-video variability, where train/test composition can substantially affect generalization. Still, FSGM-Sum remains consistently competitive with the strongest reported baselines in the main comparison, and the strongest split results (e.g., SumMe Split3) suggest that when train/test content distributions align, the structured gating mechanism can yield highly reliable rankings. Importantly, reporting split-wise dispersion provides a clearer robustness signal than the mean alone for subjectively annotated datasets.

5.1 Efficiency Analysis

We also compare efficiency against CSTA [20]. As shown in Table 4, FSGM-Sum is substantially more compact, using only 3.43M parameters (13.09 MB). The FSGM core itself remains lightweight (276.67K parameters), with most capacity allocated to input/output projections. Table 5 shows that FSGM-Sum also has the lowest computational cost across all tested sequence lengths, requiring only 537.61M MACs at $T = 100$ and 2.67G MACs at $T = 500$. The latency results in Table 6 are consistent with this trend: at $T = 100$, FSGM-Sum is the fastest model on both GPU and CPU. Combined with the performance gains in Table 2, these results show that FSGM-Sum offers a better accuracy–efficiency trade-off than existing methods.

5.2 Ablation Study

This ablation isolates the effect of each component in FSGM-Sum by adding modules one at a time (Table 7). Variant (1) disables the factorized gate and collapses the model to a shallow regressor on the pre-extracted descriptors, providing a direct reference point for the benefit of structured token mixing. Introducing the factorized spatio-temporal gate in (2) injects the key inductive bias

Model	Total Params	Trainable Params	Size (MB)
CSTA-GoogleNet	9.78M	9.78M	37.31
CSTA-ResNet	13.01M	13.01M	49.62
CSTA-MobileNet	7.78M	7.78M	29.66
CSTA-EfficientNet	9.56M	9.56M	36.47
FSGM-Sum (ours)	**3.43M**	**3.43M**	**13.09**

Table 4: Parameter comparison with CSTA variants. FSGM-Sum achieves the strongest parameter efficiency, requiring only 3.43M parameters and 13.09 MB, substantially smaller than all CSTA backbones.

T	Model	MACs	FLOPs (∼2×)	vs FSGM
100	CSTA-GoogleNet	3.54G	7.08G	6.6×
	CSTA-ResNet	2.31G	4.61G	4.3×
	CSTA-MobileNet	1.63G	3.26G	3.0×
	CSTA-EfficientNet	1.90G	3.79G	3.5×
	FSGM-Sum (ours)	**537.61M**	**1.08G**	**1.0×**
200	CSTA-GoogleNet	7.13G	14.26G	6.7×
	CSTA-ResNet	4.39G	8.78G	4.1×
	CSTA-MobileNet	3.13G	6.27G	2.9×
	CSTA-EfficientNet	3.62G	7.23G	3.4×
	FSGM-Sum (ours)	**1.07G**	**2.14G**	**1.0×**
500	CSTA-GoogleNet	18.03G	36.06G	6.8×
	CSTA-ResNet	10.63G	21.26G	4.0×
	CSTA-MobileNet	7.68G	15.36G	2.9×
	CSTA-EfficientNet	8.82G	17.63G	3.3×
	FSGM-Sum (ours)	**2.67G**	**5.34G**	**1.0×**

Table 5: Computational complexity at different sequence lengths. FSGM-Sum consistently requires the fewest MACs across all tested sequence lengths, indicating favorable scaling for longer videos.

T	Model	GPU (ms)	CPU (ms)	Relative CPU speed
100	**FSGM-Sum (ours)**	**4.65 ± 1.65**	**51.71 ± 119.47**	**1.0×**
	CSTA-GoogleNet	6.35 ± 1.70	216.10 ± 474.14	4.2× slower
	CSTA-ResNet	5.25 ± 1.58	110.28 ± 276.11	2.1× slower
	CSTA-MobileNet	5.80 ± 1.65	202.50 ± 424.85	3.9× slower
	CSTA-EfficientNet	6.44 ± 1.36	295.57 ± 736.75	5.7× slower
200	**FSGM-Sum (ours)**	**4.64 ± 2.41**	**55.65 ± 136.24**	**1.0×**

Table 6: Inference latency. FSGM-Sum is the fastest model among the compared methods at $T = 100$ on both GPU and CPU. The partial $T = 200$ result available for FSGM-Sum indicates stable runtime as sequence length increases.

Variant	SumMe		TVSum	
	τ	ρ	τ	ρ
(1) Base regressor (no FSGM gate)	0.103	0.114	0.008	0.010
(2) + Factorized spatio-temporal gate (outer-sum logits)	0.190	0.212	0.085	0.112
(3) + Axis-wise norm. (Softmax_axis=TD)	0.208	0.255	0.184	0.241
(4) + Positional encoding	0.219	0.258	0.189	0.248
(5) + CLS token	0.227	0.262	0.192	0.252
(6) + Key/Value embedding	0.243	0.271	0.195	0.255
(7) + Skip connection + normalization	**0.252**	**0.284**	**0.197**	**0.258**

Table 7: Step-by-step ablations for FSGM-Sum (expected values). Each row adds one component. Kendall's τ and Spearman's ρ are reported under the standard 5-split protocol.

of the method: the temporal branch highlights *when* informative events occur, while the channel branch highlights *which* feature dimensions carry salient evidence. Variant (3) then applies axis-wise normalization (TD) to transform gate logits into well-behaved modulation weights along both time and feature axes, which improves rank stability by avoiding gates that are either overly diffuse or excessively concentrated.

Variants (4) and (5) add lightweight context cues. Positional encoding in (4) helps separate frames with similar appearance but different temporal roles, which is particularly useful in long or repetitive sequences. Adding a CLS token in (5) provides a compact global signal that regularizes the scoring head and typically improves generalization when supervision is limited. Variant (6) further introduces key/value embeddings to decouple the features used to *produce* the gate from those being *modulated*, allowing more flexible scaling while retaining linear-time complexity. Finally, (7) adds a skip connection and normalization to stabilize optimization, anchoring the gating operation to the input stream and reducing sensitivity to small score perturbations an issue that matters for SumMe where segmentation-and-knapsack decoding can magnify minor ranking changes.

6 Conclusion

This paper presented **FSGM-Sum**, a lightweight supervised video summarization model designed for efficient and stable frame-importance prediction under the standard segmentation-plus-knapsack evaluation pipeline. In contrast to attention-heavy temporal models, FSGM-Sum replaces dense pairwise interactions with a *Factorized Spatio-Temporal Gating Mixer* that decomposes importance estimation into complementary *temporal* and *channel* saliency branches. Extensive experiments on **SumMe** and **TVSum** using rank-based metrics validate the effectiveness of the proposed design. Under the standard five-split protocol, FSGM-Sum achieves strong Kendall and Spearman correlations and consistently improves over prior deterministic baselines, while remaining competitive with recent diffusion-based formulations. These results indicate that structured factorized gating can serve as a practical alternative to dense attention for spatiotemporal importance modeling, offering a favorable trade-off between accuracy and efficiency for long-video summarization. Future work can extend this direction in three ways: (i) incorporating explicit modeling of annotator disagreement through distributional or uncertainty-aware scoring while retaining single-pass inference, (ii) exploring tighter decoder-aligned objectives that directly optimize score stability under segmentation and knapsack selection, and (iii) adapting the factorized gating design to multimodal settings or higher-resolution clip representations to further improve semantic relevance without sacrificing efficiency.

References

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2020. Unsupervised Video Summarization via Attention-Driven Adversarial Learning. In *MultiMedia Modeling*. Springer, 492–504. doi:10.1007/978-3-030-37731-1_40

[2] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2021. Combining Global and Local Attention with Positional Encoding for Video Summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*. 226–234. doi:10.1109/ISM52913.2021.00045

[3] Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Dernoncourt, and Joon Son Chung.

Unpublished working draft. Not for distribution.

2024. Scaling Up Video Summarization Pretraining with Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8332–8341.

[4] Jose M. Rojas Chaves and Subarna Tripathi. 2024. VideoSAGE: Video Summarization with Graph Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2527–2534.

[5] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2019. Summarizing Videos with Attention. In *Computer Vision – ACCV 2018 Workshops*. Springer, 39–54. doi:10.1007/978-3-030-21074-8_4

[6] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and Attend: Multimodal Summarization With Dual Contrastive Losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14867–14878.

[7] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14867–14878.

[8] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2020. Video Summarization With Attention-Based Encoder-Decoder Networks. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 6 (2020), 1709–1717. doi:10.1109/TCSVT.2019.2904996

[9] Hao Jiang and Yadong Mu. 2022. Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16388–16398.

[10] Hao Jiang and Yadong Mu. 2022. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16388–16398.

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

[12] Kwanseok Kim, Jaehoon Hahm, Sumin Kim, Jinhwan Sul, Byunghak Kim, and Joonseok Lee. 2025. SummDiff: Generative Modeling of Video Summarization with Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15096–15106.

[13] Min Jung Lee, Dayoung Gong, and Minsu Cho. 2025. Video Summarization with Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18981–18991.

[14] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. 2023. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5584–5593.

[15] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised Video Summarization With Adversarial LSTM Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 202–211.

[16] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 202–211.

[17] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems* 34 (2021), 13988–14000.

[18] Ashish Prasad, Pranav Jeevan, and Amit Sethi. 2024. EDSNet: Efficient-DSNet for Video Summarization. *arXiv preprint arXiv:2409.14724* (2024).

[19] Mrigank Rochan, Linwei Ye, and Yang Wang. 2018. Video Summarization Using Fully Convolutional Sequence Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 347–363.

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

[21] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing Web Videos Using Titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5179–5187. doi:10.1109/CVPR.2015.7299154

[22] Hacene Terbouche, Maryan Morel, Mariano Rodriguez, and Alice Othmani. 2023. Multi-Annotation Attention Model for Video Summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3142–3151.

[23] Junyan Wang, Yang Bai, Yang Long, Bingzhang Hu, Zhenhua Chai, Yu Guan, and Xiaolin Wei. 2020. Query twice: Dual mixture attention meta learning for video summarization. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4023–4031.

[24] Ke Zhang, Wei-Lin Chao, Fei Sha, and Kristen Grauman. 2016. Video Summarization with Long Short-term Memory. In *Computer Vision – ECCV 2016*. Springer, 766–782. doi:10.1007/978-3-319-46478-7_47

[25] Ke Zhang, Wei-Lin Chao, Fei Sha, and Kristen Grauman. 2016. Video Summarization with Long Short-term Memory. In *Computer Vision – ECCV 2016*. Springer, 766–782. doi:10.1007/978-3-319-46478-7_47 Includes the dppLSTM variant.

[26] Yunzuo Zhang, Yameng Liu, Weili Kang, and Ran Tao. 2023. VSS-Net: Visual Semantic Self-mining Network for Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

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

[28] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. 2021. DSNet: A Flexible Detect-to-Summarize Network for Video Summarization. *IEEE Transactions on Image Processing* 30 (2021), 948–962. doi:10.1109/TIP.2020.3039886

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

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Unpublished working draft. Not for distribution.