SMART: Selective Multimodal Aggregation and Refinement over Time for Video Summarization

Joohyun Chang [1] **Minsung Kim** [1] **Sung-Bin Kim** [2] **Chenshuang Zhang** [1] **Tae-Hyun Oh** [1]

Abstract

The growth of video content has created a strong demand for video summarization that finds key moments from numerous frames. Recent methods mainly use pretrained language models to estimate frame importance based on visual captions, but they often ignore non-visual cues such as speech and audio events. To address this, we propose **SMART**, a selective multimodal aggregation and temporal refinement framework for video summarization. SMART introduces visual-guided modality selection to attend to relevant auxiliary cues, as well as progressive window attention to refine timestep-level representations over a broader temporal context. Experiments show that SMART outperforms 13 state-of-the-art methods on the widely-applied TVSum dataset, demonstrating the effectiveness of selective multimodal integration for video summarization.

1. Introduction

With the rapid advancement of smart devices and online platforms, videos have played a critical role for humans to record and understand the world. Despite the rich information, their long durations bring great challenges to efficiently perceiving the key content from numerous frames. Video summarization (Song et al., 2015; Otani et al., 2019) solves this problem by automatically finding the representative moments from long videos and has achieved great attention. Specifically, previous methods predict an importance score for each frame to indicate its contribution.

Existing methods mainly rely on visual saliency to obtain frame importance scores (Narasimhan et al., 2021; He et al., 2023). As text provides a compact representation of video semantics, recent approaches first convert video content into textual descriptions (e.g., video captions), then use pretrained language models for importance estimation (Lee et al., 2025). However, humans perceive the world through multimodal signals, such as audio, leaving urgent demand for multimodal representations for video summarization.

To address this challenge, we propose **S**elective **M**ultimodal **A**ggregation and **R**efinement over **T**ime (**SMART**). Beyond video captions, SMART introduces complementary multimodal evidence which helps identify important moments, such as speech content, facial emotions, and audio events. To mitigate the dilution of useful information caused by simple cue combination, SMART introduces two components: visual-guided modality selection for selective multimodal aggregation and progressive window attention for temporal refinement. Specifically, **visual-guided modality selection** treats visual captions as the primary semantic reference and uses other signals as auxiliary evidence. Rather than concatenating all cues uniformly, SMART selects auxiliary information conditioned on the central visual context and the local importance query. This design allows the model to emphasize relevant auxiliary cues while reducing the influence of redundant, noisy, or weakly aligned signals. **Progressive window attention** then refines timestep-level representations, gradually expanding the receptive field from local changes to the full video context.

We summarize the key contributions as follows:

- We extend language-based video summarization beyond visual caption only representations by incorporating multimodal textual cues such as audio captions, transcripts, and facial emotions.
- We propose **SMART**, which introduces visual-guided modality selection and progressive window attention to aggregate relevant auxiliary cues and refine temporal representations for video summarization.
- Experimental results show that SMART outperforms visual only and naive multimodal concatenation baselines, and ablations verify the contribution of each component.

2. Method

Given a video $V=\{I_t\}_{t=1}^{T}$ sampled at T timesteps, video summarization aims to predict an importance score $\hat{s}_t \in$

[1]KAIST [2]POSTECH. Correspondence to: Tae-Hyun Oh <tae-hyun.oh@kaist.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

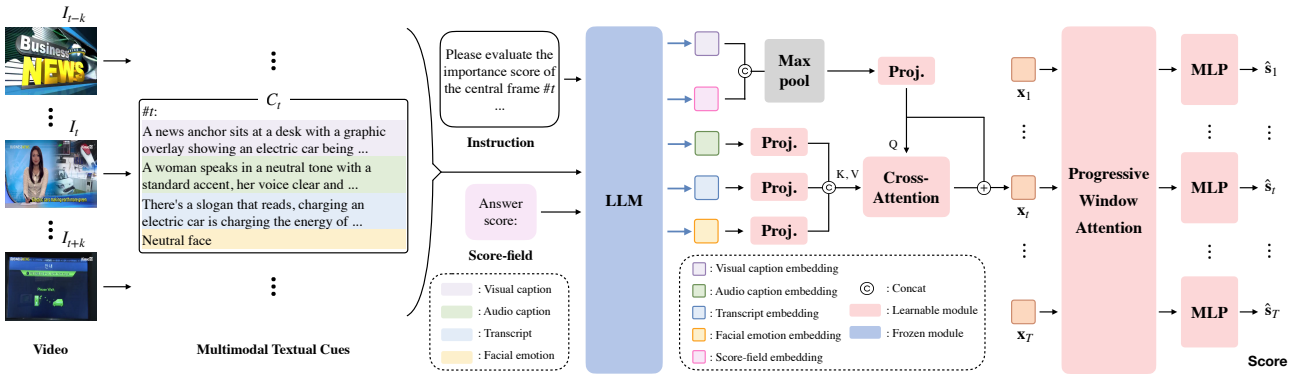

Figure 1. **Overview of SMART.** We first construct aligned multimodal textual cues from visual captions, audio captions, transcripts, and facial emotion semantics. A frozen LLM extracts modality-specific representations from a local temporal window around each timestep t. Visual-guided modality selection uses the visual caption and score-field representations as queries to select relevant auxiliary cues, producing a compact timestep-level representation. Progressive window attention then refines these representations over increasingly broader temporal contexts and predicts timestep-level importance scores.

$[0, 1]$ for each sampled timestep t. The resulting score sequence $\hat{\mathbf{s}} = \{\hat{s}_t\}_{t=1}^{T}$ is used to select representative moments that summarize the video. We propose **SMART**, which selectively aggregates multimodal textual cues and progressively refines temporal representations for video summarization.

Multimodal Textual Cues. For each time step t, we construct a set of textual descriptors $\mathcal{C}_t = \{c_t^v, c_t^a, c_t^r, c_t^f\}$, where c_t^v, c_t^a, c_t^r, and c_t^f denote the visual caption, audio caption, transcript, and facial emotion, respectively. At each time step t, the visual caption describes the visible scene and action, the audio caption describes salient sound events, the transcript captures spoken content, and the facial emotion records the top predicted emotion label when a face is detected. Representing heterogeneous modalities as text enables a unified language model interface while preserving the interpretability of each cue.

Local Contextual Encoding. For each time step t, we form a local context $\mathcal{W}_t = \{\mathcal{C}_{t-k}, \ldots, \mathcal{C}_t, \ldots, \mathcal{C}_{t+k}\}$, where $2k+1$ is the window size. We serialize $\mathcal{W}_t$ into a structured prompt and encode it with a frozen language model. For each modality, we aggregate the hidden states associated with its token span, yielding $\mathbf{v}_t$, $\mathbf{a}_t$, $\mathbf{r}_t$, and $\mathbf{f}_t$ for the visual caption, audio caption, transcript, and facial emotion, respectively. In addition, we obtain a score-field representation $\mathbf{g}_t$ from the hidden states of the answer score field.

Visual-guided Modality Selection. Visual-guided modality selection uses the visual caption and score-field representations as the query for attending to auxiliary modalities. This design lets the model condition auxiliary modality selection on the central visual context and the score field . For each time step t, we obtain a query $\mathbf{q}_t$ by max-pooling the stacked visual caption and score-field representations, followed by a projection. We obtain auxiliary tokens $\mathbf{M}_t$ by projecting and stacking the audio caption, transcript, and facial emotion representations. The fused representation is computed with a residual cross-attention update:

$$\mathbf{x}_t = \mathbf{q}_t + \mathrm{CrossAttn}(\mathbf{q}_t, \mathbf{M}_t, \mathbf{M}_t). \tag{1}$$

During training, we apply modality dropout before forming $\mathbf{M}_t$. Each auxiliary modality is independently replaced with a zero vector with probability p, while the visual caption and score-field representations are always retained. This stochastic masking prevents the model from over-relying on any single auxiliary modality and encourages it to exploit complementary cues across modalities.

Progressive Window Attention. The fused representations form a video-level sequence $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$. We refine this sequence with progressive window attention and obtain $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_T]$. The attention window expands with depth, allowing early layers to capture short-range temporal changes and deeper layers to aggregate broader context up to the full sequence. Finally, an MLP prediction head maps each $\tilde{\mathbf{x}}_t$ to the importance score $\hat{s}_t$.

Training Objective.. Let $\mathbf{s} = \{s_t\}_{t=1}^{T}$ denote the ground-truth importance score sequence aligned with the sampled timesteps. During training, the cue extraction models and the language model are kept frozen. We optimize only the learnable projection modules, visual-guided modality selection, progressive window attention, and the MLP prediction head. We use mean squared error between the predicted and ground-truth timestep-level importance scores as the training loss. Detailed configurations, temporal alignment, and prompt templates are provided in Appendix A.

3. Experiments

In this section, we validate the effectiveness of SMART on the standard video summarization benchmark following

Table 1. Comparison with prior methods on TVSum, where τ and ρ denote Kendall's tau and Spearman's rho, respectively. V, T, A, and E denote visual, text, audio and facial emotion cues.

Method	Modality	TVSum	
		τ	ρ
Human (Otani et al., 2019)	-	0.177	0.204
visual-only methods			
VASNet (Fajtl et al., 2019)	V	0.160	0.170
DSNet-AB (Zhu et al., 2020)	V	0.108	0.129
DSNet-AF (Zhu et al., 2020)	V	0.113	0.138
DMASum (Wang et al., 2020)	V	0.203	0.267
PGL-Sum (Apostolidis et al., 2021)	V	0.206	0.157
MSVA (Ghauri et al., 2021)	V	0.190	0.210
iPTNet (Jiang & Mu, 2022)	V	0.134	0.163
CSTA (Son et al., 2024)	V	0.194	0.255
visual-text methods			
CLIP-It (Narasimhan et al., 2021)	V + T	0.108	0.147
A2Summ (He et al., 2023)	V + T	0.137	0.165
SSPVS (Li et al., 2023)	V + T	0.181	0.238
Argaw et al. (2024)	V + T	0.155	0.186
language-based methods			
LLMVS (Lee et al., 2025)	V + T	0.211	0.275
SMART	V + T + A + E	**0.225**	**0.292**

existing studies. Our experiments address two questions: (i) Can SMART achieve superior results over state-of-the-art methods? (Sec. 3.2) (ii) How does each module in SMART contribute to the final performance? (Sec. 3.3) We first describe the experimental setup, followed by comparison results and ablation studies.

3.1. Experiment Setup

Dataset. We evaluate SMART on TVSum (Song et al., 2015), a standard video summarization benchmark containing 50 videos from diverse web domains. The videos range from 2 to 10 minutes, with an average length of approximately 4 minutes, and each video is annotated by 20 users with temporal importance scores.

Evaluation. We follow the standard 5-fold cross-validation protocol widely adopted in prior work (Lee et al., 2025; Apostolidis et al., 2021; Jiang & Mu, 2022). For each test video, SMART predicts a sequence of importance scores over sampled timesteps. TVSum annotations are originally collected over temporal segments, and the standard preprocessed benchmark represents them as score sequences aligned with sampled frame indices. We align our one-second multimodal cue representations to this preprocessed timestep sequence before training and evaluation.

For each test video, we compare the predicted score sequence with the annotation from each user and average the results across annotators. Following (Otani et al., 2019), we report Kendall's τ and Spearman's ρ, which measure rank-order correlation between predicted and human importance scores. Implementation details are provided in Appendix A.

Table 2. Ablation study on TVSum. Visual captions are used in all settings, and τ and ρ denote Kendall's tau and Spearman's rho, respectively. VMS and PWA denote visual-guided modality selection and progressive window attention, respectively.

Auxiliary Modalities			Components		TVSum	
Audio Cap.	Transcript	Facial Emo.	VMS	PWA	τ	ρ
Baseline						
✗	✗	✗	✗	✗	0.208	0.270
Naive modality addition						
✓	✗	✗	✗	✗	0.209	0.272
✗	✓	✗	✗	✗	0.209	0.272
✗	✗	✓	✗	✗	0.213	0.278
✓	✓	✓	✗	✗	0.211	0.275
SMART components						
✓	✓	✓	✓	✗	0.219	0.285
✓	✓	✓	✗	✓	0.219	0.284
✓	✓	✓	✓	✓	**0.225**	**0.292**

Compared Methods. We compare SMART with three groups of prior methods on TVSum: (i) visual-only methods, which estimate frame importance primarily from visual and temporal representations; (ii) visual-text methods, which incorporate textual information together with visual features; (iii) recent language-based methods, represented by LLMVS (Lee et al., 2025), which performs frame-importance estimation from visual captions using a pretrained language model.

3.2. Comparison with Prior Methods

Table 1 compares SMART with state-of-the-art video summarization methods on TVSum. SMART achieves the best performance among all methods, outperforming both visual-only methods (e.g., CSTA (Son et al., 2024)) and visual-text methods (e.g., SSPVS (Li et al., 2023)). Notably, our SMART outperforms the most recent LLMVS method by approximately 6.6% on Kendall's τ and 6.2% on Spearman's ρ. These results demonstrate the effectiveness of our SMART method in improving frame-importance estimation.

3.3. Ablation Study

Table 2 analyzes the contribution of SMART components on TVSum. The baseline in Table 2 only uses visual captions to represent frames. Experimental results show that adding multimodal cues improves over the baseline setting, demonstrating the effectiveness of additional semantic evidence from audio captions, transcripts, and facial emotions.

Visual-guided modality selection improves over the corresponding naive settings by conditioning auxiliary cue aggregation on the visual-caption representation. In particular, when audio captions, transcripts, and facial emotions are used together, applying Visual-guided modality selection improves performance from 0.211 to 0.219 in Kendall's τ and from 0.275 to 0.285 in Spearman's ρ. This suggests

that visual-guided selection helps exploit complementary information while suppressing redundant or weakly aligned signals. Progressive window attention further refines the selected representations over temporal context, leading to the best overall performance. These results show that SMART benefits from multimodal textual cues through selective aggregation rather than simple prompt-level fusion.

3.4. Discussion

SMART uses textualized multimodal cues as a unified interface to pretrained language models. This design improves interpretability and allows the framework to incorporate off-the-shelf captioning, transcription, and facial emotion models. At the same time, textualization may abstract away fine-grained visual or temporal details that are not captured by the generated descriptions. The quality of auxiliary cues can also depend on the external pretrained models. However, SMART is modular by design, so stronger cue extractors or domain-specific models can be incorporated without changing the overall aggregation and temporal refinement framework.

4. Conclusion

In this work, we propose SMART, a selective multimodal aggregation and temporal refinement framework that extends visual-caption-based video summarization with auxiliary textual cues from audio captions, transcripts, and facial emotions. Visual-guided modality selection emphasizes auxiliary cues relevant to the central visual context, while progressive window attention refines timestep-level representations over broader temporal contexts. Experiments show that SMART outperforms 13 state-of-the-art methods, demonstrating the effectiveness of our method for video summarization.

5. LLM/Agent Usage Disclosure

LLM-based tools were used to support manuscript preparation and implementation, including grammar correction, wording refinement, LaTeX polishing, code writing, debugging, and refactoring. The research idea, method design, experimental setup, and result interpretation were developed through discussions among the authors without LLM-based ideation. All final analyses and manuscript content were reviewed and verified by the authors.

Acknowledgements

This work was supported by the InnoCORE program of the Ministry of Science and ICT (N10250156, KAIST InnoCore LLM) (25%); the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00453301) (25%); and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project (25%); No. RS-2026-25518317, Development of AI memory mechanism that reflects human cognitive principles (25%)). This work was also partially supported by Toonation Co.

References

Apostolidis, E., Balaouras, G., Mezaris, V., and Patras, I. Combining global and local attention with positional encoding for video summarization. In *IEEE International Symposium on Multimedia (ISM)*, pp. 226–234. IEEE, 2021.

Argaw, D. M., Yoon, S., Heilbron, F. C., Deilamsalehy, H., Bui, T., Wang, Z., Dernoncourt, F., and Chung, J. S. Scaling up video summarization pretraining with large language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8332–8341, 2024.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Bain, M., Huh, J., Han, T., and Zisserman, A. Whisperx: Time-accurate speech transcription of long-form audio. *Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., and Chang, L. J. Py-feat: Python facial expression analysis toolbox. *Affective Science*, 4(4):781–796, 2023.

Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., and Remagnino, P. Summarizing videos with attention. In *Asia Conference on Computer Vision Workshops (ACCVW)*, pp. 39–54. Springer, 2019.

Ghauri, J. A., Hakimov, S., and Ewerth, R. Supervised video summarization via multiple feature sets with parallel attention. In *ICME*, pp. 1–6. IEEE, 2021.

Ghosh, S., Goel, A., Jayakumar, K., Koroshinadze, L., Anand, N., Kong, Z., Gururani, S., Lee, S.-g., Kim, J., Aljafari, A., et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., and Wang, Z. Align and attend: Multimodal summarization with dual contrastive losses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14867–14878, 2023.

Jiang, H. and Mu, Y. Joint video summarization and moment localization by cross-task sample transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16388–16398, 2022.

Lee, M. J., Gong, D., and Cho, M. Video summarization with large language models. *arXiv preprint arXiv:2504.11199*, 2025.

Li, H., Ke, Q., Gong, M., and Drummond, T. Progressive video summarization via multimodal self-supervised learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 5584–5593, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Narasimhan, M., Rohrbach, A., and Darrell, T. Clip-it! language-guided video summarization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13988–14000, 2021.

Otani, M., Nakashima, Y., Rahtu, E., and Heikkila, J. Rethinking the evaluation of video summaries. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7596–7604, 2019.

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

Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5179–5187, 2015.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, J., Bai, Y., Long, Y., Hu, B., Chai, Z., Guan, Y., and Wei, X. Query twice: Dual mixture attention meta learning for video summarization. In *ACM MM*, pp. 4023–4031, 2020.

Zhu, W., Lu, J., Li, J., and Zhou, J. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing (TIP)*, 30:948–962, 2020.

A. Implementation Details

Multimodal Cue Extraction. We extract multimodal textual cues from each video at one-second intervals. For visual cues, we use Qwen3-VL-8B-Instruct (Bai et al., 2025) to generate scene- and action-level captions. For audio cues, we use AF-Next (Ghosh et al., 2026) to generate captions that describe salient acoustic events and environmental sounds. We use WhisperX (Bain et al., 2023) to obtain speech transcripts and Py-Feat (Cheong et al., 2023) to extract facial emotion labels when visible faces appear. We convert these heterogeneous cues into text and construct the multimodal cue set for each timestamp.

LLM Encoding and Prompting. We use Llama-2-13B-chat-hf (Touvron et al., 2023) as the frozen language model. For each timestamp, we form a local temporal context with a window size of 7 and serialize the multimodal cues into a structured prompt. The prompt asks the language model to assess the importance of the central timestamp using the surrounding context. We then extract hidden states from the frozen language model and aggregate the token spans corresponding to each modality and the score field. Figure A1 shows the prompt template.

Temporal Alignment and Supervision. We extract multimodal cues at one-second intervals, while the preprocessed TVSum benchmark provides ground-truth importance scores at sampled frame indices. We therefore align the cue representations to the sampled timestep sequence before training and evaluation. Let D denote the video duration in seconds and T denote the length of the preprocessed ground-truth score sequence. For each cue timestamp, we map its temporal position to the score-sequence axis by scaling it with T/D. We then assign each target timestep to the nearest available cue representation.

We use the aligned ground-truth score sequence as supervision. During training, we keep the cue extraction models and the language model frozen. We optimize only the learnable modules after LLM encoding, including the projection modules, visual-guided modality selection, progressive window attention, and the MLP prediction head. We use mean squared error between the predicted and ground-truth timestep-level importance scores as the training loss.

Model Configuration. For progressive window attention, we use two attention heads. We set the attention window sizes to 5, 15, and N, where N denotes the full sequence length. This design lets the temporal receptive field expand from local neighborhoods to the full video context. The MLP prediction head maps the refined timestep representations to scalar importance scores.

Optimization. We optimize the learnable parameters using AdamW (Loshchilov & Hutter, 2017) with a learning rate of 7×10^{-5}. We set the batch size to 5 and train the model for 200 epochs. We apply modality dropout with probability 0.3 during training. For each auxiliary modality, modality dropout replaces its representation with a zero vector, while the visual caption and score-field representations remain available. This encourages the model to avoid over-reliance on a single auxiliary cue.

System Prompt

You are an intelligent chatbot designed to critically assess the importance of a central frame within a specific context. Each frame is described with multimodal information: visual scene description; per-second facial expression semantics summarizing the dominant visible facial emotion; a speech transcript describing what is being said around that second; an audio caption describing the most salient sounds around that second. Your task is to assign an importance score to the central frame based on its narrative contribution. Evaluate using the following criteria:

##INSTRUCTIONS:

1. **Narrative Significance**: Assign a high score if the frame captures pivotal plot developments, character milestones, or key conflicts/resolutions.
2. **Uniqueness and Novelty**: Score highly if the frame introduces new elements or showcases significant alterations in the story or setting.
3. **Action and Dynamics**: Give a high score if the frame depicts crucial actions or is characterized by high energy or movement.

##NOTE: Use all provided modalities to assess the importance of the central frame. Auxiliary modalities should complement, not override, the visual scene description.

User Prompt

[EXAMPLE]

{In-context examples}

[QUERY]

Please evaluate the importance score of the central frame #1 in following 4 frames. Be stingy with scores.

— —

#1: {Visual caption} | sound: {Audio caption} | face: {Facial emotion} | transcript: {Transcript}

#2: {Visual caption} | sound: {Audio caption} | face: {Facial emotion} | transcript: {Transcript}

...

#N: {Visual caption} | sound: {Audio caption} | face: {Facial emotion} | transcript: {Transcript}

— —

Provide your score as an integer between 0 and 10.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.

Answer

score:

Figure A1. Prompt template for local contextual encoding, where $\{\cdot\}$ indicates a placeholder for the target content, and N denotes the size of the local context.