Narrativity-Aware Video Summarization Based on Vision and Language Foundation Models

Shumpei Saito* Hiroyuki Ueda* Yosuke Ito* Kazuyoshi Yoshii*

*Graduate School of Engineering, Kyoto University, Japan
saito.shumpei.34c@st.kyoto-u.ac.jp, {ueda.hiroyuki.8f, ito.yosuke.6s, yoshii.kazuyoshi.3r}@kyoto-u.ac.jp

***Abstract*—This paper presents a novel video summarization approach that prioritizes the narrative quality of the summarized video to enhance its enjoyment and appeal. While most video summarization studies focus on extracting salient scenes using low-level visual features, they often neglect the storytelling aspect to optimize numerical performance on standard benchmarks. To address this, we propose a multifaceted video summarization method that leverages vision and language foundation models to assess shot-level importance (e.g., 2-sec intervals) based on both visual salience and textual narrativity. Specifically, our method employs a vision-language model (VLM) to generate objective captions for individual shots. These shot-wise textual descriptions are then fed into a large language model (LLM) with a prompt designed to produce a semantically-coherent text summary with strong narrativity. The narrativity-aware text embeddings obtained by the LLM, combined with visual embeddings from a vision foundation model, are processed by a recurrent neural network (RNN) to predict importance scores. The LLM and RNN are jointly fine-tuned to align with existing benchmarks. Experiments on the SumMe benchmark demonstrated the effectiveness of our multifaceted approach, highlighting significant performance improvements and the potential of text-domain video summarization.**

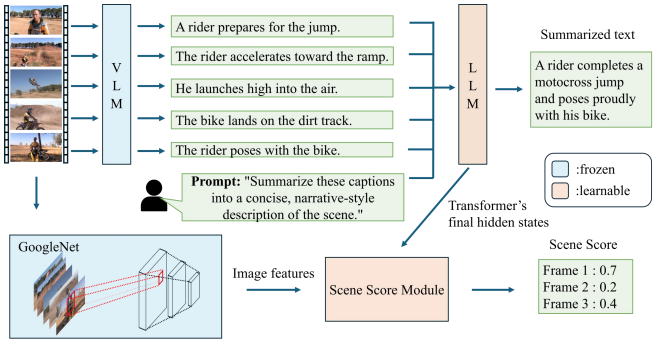

Fig. 1: The proposed multifaceted video summarization method. shot-wise captions generated by a VLM, along with a prompt, are fed to an LLM to generate a narrative-aware text summary. The final hidden states and GoogLeNet-derived visual features are jointly used to compute the importance scores.

I. Introduction

Today, video content has emerged as a primary medium for information dissemination, with its consumption growing rapidly. In this information-rich landscape, there is a growing demand for summarization techniques that not only condense videos but also preserve their narrative structure and coherence, moving beyond simple data reduction. These capabilities have wide-ranging applications in education, entertainment, news distribution, surveillance, and content management, highlighting the increasing societal and economic importance of robust summarization methods [1]–[3].

To meet this demand, early studies used deep learning models such as convolutional neural networks (CNNs) [4]–[6], long short-term memory (LSTMs) [7]–[9], and Transformers [10] to capture visual features and temporal dynamics for selecting key frames or shots. More recently, multimodal models like contrastive language-image pre-training (CLIP) [11] have been used to learn the correspondences between vision and language, supporting applications such as caption generation and retrieval. Specifically, vision-language models (VLMs) [11], [12] have emerged as powerful tools capable of transforming raw visual data into rich, semantically meaningful textual descriptions. Nevertheless, these approaches largely depend on low-level visual cues and temporal patterns. This makes it inherently difficult to capture higher-order semantic structures, such as overarching themes, emotional transitions, or causal relationships that span an entire video, thereby failing to generate summaries that truly reflect the abstract meaning and narrative flow of the content. The rise of large language models (LLMs), with their powerful contextual understanding and natural language generation abilities [13], has opened promising new directions for more semantically faithful and human-interpretable video summarization by enabling the creation of coherent narratives. In this paper, we define *narrativity* as the property of a summary that preserves temporal continuity and causal relations among successive events and maintains a coherent trajectory of the intentions of the main characters throughout the video, beyond low-level saliency.

In response, several LLM-based video summarization methods have emerged [14], aiming to integrate the language-based thinking process into the summarization process. However, most existing approaches employ LLMs primarily as scoring mechanism to estimate the importance of frames or shots for content selection. For example, Lee et al. [15] uses local window-based frame captions to estimate frame-level importance with an LLM, aggregating these scores globally using self-attention. Barbara et al. [16] instead leverages user queries to evaluate scene-level descriptions via LLM and heuristically propagates these scores to frames for flexible zero-shot summarization.

979-8-3315-7206-8/25/$31.00

In both cases, the LLM output serves as an importance score, rather than generating a consistent, abstract natural language summary itself, and these methods provide limited mechanisms for maintaining precise alignment between generated text summaries and original video frames.

To address these limitations, we propose a mutifaceted video summarization method that explicitly leverages a LLM not merely as a scoring tool but as a generator of coherent, contextually aware natural language summaries. We first use a VLM to transform video shots into a sequence of natural language captions, capturing frame-level semantics. These caption sequences are then fed into an LLM to generate an abstract, semantically consistent summary of the entire video. Crucially, this semantically rich textual summary serves as a foundation for generating a concise, reconstructed video, enabling users to efficiently grasp the essence of long-form content through a dynamic and narrative-driven viewing experience.

To maintain precise alignment between the generated summary and original video content, we extract shot-level embeddings from the final hidden layer of the LLM and compute the importance scores using an LSTM. This design ensures semantic fidelity and narrative coherence, enabling the reconstruction of concise, human-aligned video summaries. For improved performance on existing benchmarks, we also use shot-level visual features obtained from a pre-trained vision foundation model along with textual representations. This allows the system to leverage complementary visual cues in addition to textual understanding, enabling the generation of video summaries that capture both textual and visual contexts and better align with human interpretation.

II. Related Work

This section reviews related work on video summarization and text summarization with or without LLMs.

A. Video Summarization

Video summarization research has typically leveraged deep learning techniques to capture temporal dynamics [3]. Primary approaches select key frames or shots based on visual features and temporal patterns. Among LSTM-based methods, Zhang et al. [7] conducted pioneering work modeling variable temporal dependencies among video frames, while Ji et al. and Yao et al. [9] proposed hierarchical frameworks and encoder-decoder structures. For effectively capturing visual and temporal features of videos, 3D convolutional neural networks (CNNs) have also been widely used, with C3D by Tran et al. [4] and I3D by Carreira and Zisserman [5] being adopted as powerful feature extractors in various video understanding tasks.

Transformer-based models with self-attention mechanisms have recently been used effectively [17]. While these models excel at capturing the dependency structure within a video, their summarization criteria tend to be limited to low-level features such as visual changes or auditory intensity. Consequently, it has still been an open problem to capture long-term semantic and causal structures such as overarching themes and emotional transitions. The narrative flow of the summary has been overlooked since early studies on structural summarization that aim to estimate the entity-level content (who, what, where, and when) for better semantic understanding [18].

B. Text Summarization

Text summarization is a major natural language processing (NLP) task that automatically generates concise, information-rich summaries from long documents. It is broadly categorized into the extractive approach aiming to select important sentences and the abstractive approach aiming to generate new, human-like expressions. Early studies used extractive methods such as TextRank [19], but the advent of Transformers triggered advanced abstractive models like BART [20] and Pegasus [21]. More recently, large language models (LLMs) such as the GPT [13] and LLaMA [22] have learned deep contextual understanding from massive text corpora, enabling the generation of high-quality summaries with logical structures. LLMs can adapt to various formats—including paragraphs, bullet points, and conversational styles—and flexibly generate *narrative* summaries based on given prompts. This study also leverages these LLM capabilities to apply them to video summarization.

C. Video Summarization with LLMs

The advent of large language models (LLMs) has opened new possibilities in video summarization through active attempts to integrate text-based thinking process into the summarization process. Among LLM-based methods, Lee et al. [15] proposed to use a multi-modal LLM to convert video frames into a sequence of captions, subsequently using an LLM to assess the frame importance. This method aims to achieve both detail and narrative coherence by refining local importance scores with a global attention mechanism, expecting that the knowledge learned by the LLM aligns with diverse semantics and human judgments. Barbara et al. [16] developed a zero-shot and text-queryable video summarizer that converts VidLM captions into user-guided skims based on LLM judgments, notably without requiring any training data. More recently, Argaw et al. [14] proposed a scalable pipeline that leverages an LLM as an oracle summarizer to generate large-scale video summarization datasets from long-form narrated videos. They extracted texts from these videos, summarizes them using an LLM with instructions, and then maps the summaries back to the original video segments to make pseudo-ground truth data. They subsequently trained a Transformer model using this large dataset to achieve autoregressive video summarization.

While these pioneering works demonstrate the effectiveness of LLMs for video summarization, many methods primarily use LLMs for frame- or shot-wise importance estimation. Coherent, abstract text summaries are thus not guaranteed to be generated. Furthermore, these methods provide limited mechanisms for maintaining precise alignment between the generated text summaries and the original video content. In the promising work on the autoregressive summary generation [14], the primary focus lies in large-scale effective dataset construction. This differs from our emphasis on coherent text summary generation with LLMs and precise alignment between the gener-

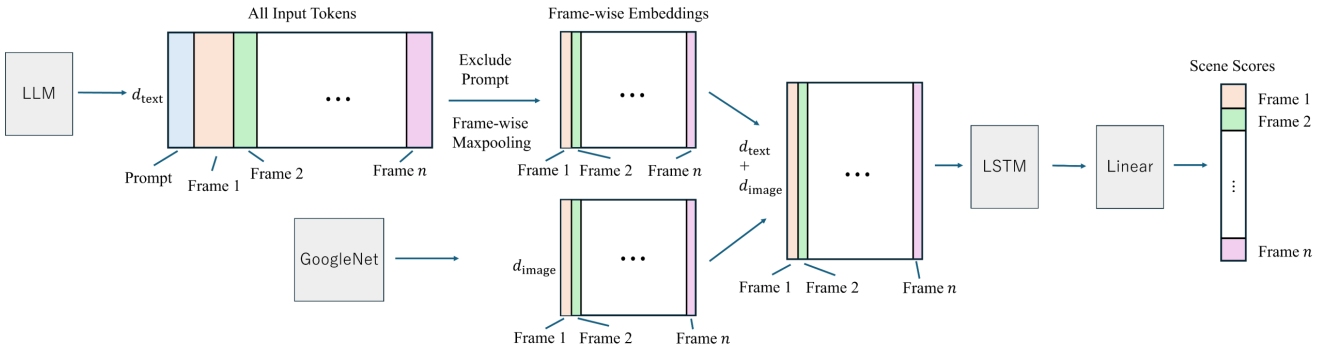

Fig. 2: Architecture of the proposed video summarization method. The original video is captioned with a VLM at a 2-s interval. The captions, combined with a instructive prompt, are input to an LLM to generate a coherent abstractive summary. The hidden text embeddings and GoogLeNet-derived visual features are fed into an LSTM to predict the importance scores aligned with the generated summary.

ated summary and the original content. We explicitly leverage an LLM not merely as a scoring tool but as a generator of coherent, contextually aware natural language summaries.

III. Proposed Method

This section outlines the multifaceted video summarization method that integrates a VLM-based caption generation and an LLM-based summarization for importance estimation.

Given an input video $\mathbf{V}$, our goal to generate a summary that preserves both semantic fidelity and narrative coherence, while also providing shot-level importance scores. We approach this task by converting video into shot-level textual descriptions, producing a coherent natural language summary, and selecting key shots that align with this summary.

A. Architecture

As shown in Fig. 1, the input video $\mathbf{V}$ is split into a sequence of shots $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]$ at a fixed interval to ensure uniform temporal coverage, where N is the number of shots. A sequence of shot-wise captions is generated with a VLM (Section III-A1). A coherent summary is then generated from all the captions with an LLM and shot-level importance scores using an LLM (Section III-A2). The importance scores are then used to identify key shots aligned with the summary (Section III-A3). The detailed pipeline is shown in Fig. 2.

1) VLM-Based Caption Generation: We use a pre-trained VLM with frozen parameters to generate shot-wise captions $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N]$ for the sequence of shots $\mathbf{F}$. Each caption $\mathbf{c}_n$ describes the visual content of shot $\mathbf{f}_n$. This step provides a semantically meaningful representation of the video content, forming the basis for the subsequent summarization process.

2) LLM-Based Text Summarization: We then use an LLM to generate an abstractive text summary that describes the overall content of $\mathbf{V}$ based on the shot-wise captions $\mathbf{C}$. We also use the internal representations of LLM to identify important shots.

To ensure good alignment between the generated summary and the original video content, we leverage the hidden representations $\bar{\mathbf{h}}_t \in \mathbb{R}^d$ given by the last layer of LLM, where t is the token index and d is the vector size. For each shot $\mathbf{f}_n$, we first identify the corresponding token range $[t_n^{\text{start}}, t_n^{\text{end}}]$ within the LLM input to form $\bar{\mathbf{H}}_n = [\bar{\mathbf{h}}_{t_n^{\text{start}}}, \bar{\mathbf{h}}_{t_n^{\text{start}}+1}, \ldots, \bar{\mathbf{h}}_{t_n^{\text{end}}}] \in \mathbb{R}^{(t_n^{\text{end}} - t_n^{\text{start}}) \times d}$. We then apply max pooling over these token representations, excluding prompt tokens, to obtain a fixed-dimensional embedding $\mathbf{h}_n \in \mathbb{R}^d$ as follows:

$$\mathbf{h}_n = \max\left(\bar{\mathbf{H}}_n\right). \tag{1}$$

This max pooling operation reduces the token-level representations with a total size of (total number of tokens $\times\, d$) into shot-level embeddings with a total size of $(N \times d)$, providing one fixed-dimensional vector per shot.

3) Importance Estimation: We gather these embeddings into $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N] \in \mathbb{R}^{N \times d}$ and process it with an LSTM network to capture temporal dependencies:

$$\mathbf{Z} = \text{LSTM}(\mathbf{H}), \tag{2}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times d'}$ and d' is the LSTM output size. Finally, we apply the following linear projection for importance prediction:

$$\hat{\mathbf{s}} = \text{Linear}(\mathbf{Z}), \tag{3}$$

where $\hat{\mathbf{s}} = [\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N] \in \mathbb{R}^N$ contains scores for selecting key shots that semantically align with the generated summary. These scores can be subsequently normalized (e.g., via min-max scaling) during evaluation to obtain values in $[0, 1]$.

We can incorporate shot-level visual features to enhance the importance scoring process. For each shot, we extract visual embedding (image features) $\mathbf{v}_n \in \mathbb{R}^p$ using GoogLeNet [23], a popular vision foundation model, and concatenate it with the text embedding $\mathbf{h}_n$ to form a combined representation: $\tilde{\mathbf{h}}_n = [\mathbf{h}_n^\top, \mathbf{v}_n^\top]^\top \in \mathbb{R}^{d+p}$. In the same way as the computation of $\mathbf{H}$, we construct $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \ldots, \tilde{\mathbf{h}}_N] \in \mathbb{R}^{N \times (d+p)}$ and process it with an LSTM and a linear projection as in Eqs. (2) and (3). This approach allows the model to leverage visual saliency as complementary information alongside textual narrativity for improved shot-level importance estimation.

B. Optimization

Our model is trained to predict shot-level importance scores that align with human-annotated ground truth for an existing benchmark on video summarization. Shots are sampled uniformly from the video at a fixed interval of τ sec producing N shots, where N varies across different videos with different durations. The ground-truth importance scores are originally provided at a different temporal resolution, which may vary across videos. To align these annotations with the extracted shots, the ground-truth score s_n for shot $\mathbf{f}_n$ is given by averaging the ground-truth score over the corresponding interval from $\tau(n-1)$ to τn sec.

Given the predicted importance scores $\hat{\mathbf{s}}$ and the aligned ground-truth scores $\mathbf{s} = [s_1, s_2, \ldots, s_N] \in \mathbb{R}^N$, we minimize the mean squared error (MSE) loss defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (s_n - \hat{s}_n)^2. \tag{4}$$

This loss encourages the model to be consistent with human-annotated importance distributions while maintaining semantic alignment with the generated summary.

IV. Evaluation

This section presents an experiment conducted to evaluate the effectiveness of the proposed visual-textual multifaceted approach to video summarization.

A. Dataset and Evaluation Metric

We used SumMe [1], a common dataset for video summarization. It consists of 25 video clips ranging from 1 to 6 min in length with an average duration of approximately 2 min and 40 sec. Those clips were captured with egocentric, moving, and static cameras in diverse scenes and cover topics such as holidays, events, and sports. Each video clip is given frame-level importance scores by 15 to 18 human subjects. As ground truth, we used the official frame-level importance scores in SumMe, computed as the mean across all annotators. No additional annotations were collected. To obtain 2-sec shot-level labels, we averaged the frame-level scores within each shot interval.

We followed the standard evaluation metric with adjustments to suit our method. Shots were uniformly extracted from the video at fixed 2-sec intervals to generate shot-level inputs. For evaluation, we ranked the predicted shot-level importance scores and selected the top 15% of frames with the highest scores from both the predictions and the ground truth annotations. The performance was then assessed using the F1 score computed between these selected sets.

Note that the conventional way of evaluation on SumMe [1] was disadvantageous for the proposed method because existing benchmarks on video summarization do not consider the textual narrativity in making the ground-truth annotations. Nonetheless, we aimed to indirectly evaluate the effectiveness of the text-domain narrativity-aware thinking for video summarization through the systematic evaluation of the importance score estimation performance.

TABLE I: F1-scores (%) on SumMe dataset.

Configuration	Mean (%)	Std (%)	Min (%)	Max (%)
With visual features	33.2	5.4	28.2	43.0
Without visual features	33.8	3.4	29.1	38.4

B. Configurations

We used **CogVLM2** [24] as a frozen VLM without any additional fine-tuning to generate captions for the shots uniformly sampled from a target video clip. The maximum length of each caption was limited to 128 tokens using the default generation settings with a generic captioning prompt. The textual description of the whole clip was obtained by concatenating all the shot-wise captions. To generate the semantically-coherent narrativity-aware text summary, we used **Llama-3.1-8B-Instruct** [25] as an optimizable LLM with the instructive prompt *"Summarize these captions into a concise narrative description of the scene."* followed by the video description. The total input length was limited to 16384 tokens to fit within the context window of the LLM.

To fine-tune the LLM, the low-rank adaptation (LoRA) [26] technique was applied to the attention layers of the LLM with a rank of $r = 128$, a dropout of 0.1, and $\alpha = 32$. We froze all the non-LoRA parameters Training was conducted using the 8-bit AdamW optimizer from bitsandbytes [27], with a learning rate of 10^{-4} and a weight decay of 10^{-2}. We used a batch size of 1 and an early stopping with a patience of three epochs based on the validation F1 score. The experiment was performed on a mixed GPU setup comprising two NVIDIA A100 (80GB) and six RTX A6000 (48GB) GPUs.

C. Quantitative Results and Ablation Study

We evaluated the proposed method via 5-fold cross-validation on the SumMe dataset, reporting the average F1-score across folds. Table I shows our main results along with an ablation comparing the use of visual features versus text-only inputs. These results indicate that incorporating visual features introduces higher variance across folds, with the maximum F1-score exceeding that of the text-only setting. This suggests that using complementary visual cues can improve scene-level alignment in some cases, although the gains are not consistent across data splits, as shown by a higher across-fold standard deviation (5.4 vs. 3.4) and a wider F1 range (28.2–43.0 vs. 29.1–38.4) in the visual-feature setting.

D. Qualitative Analysis and Discussion

Tables II, III, and IV listed the generated textual summaries for video clips #1–#3 under two conditions: without fine-tuning and with fine-tuning using visual features. Figures 3, 4, and 5 showed the corresponding importance-score curve of the fine-tuned model only (one curve per clip). Overall, the non-fine-tuned model tended to generate short, objective summaries that focus on static attributes like colors, settings, or general activities without mentioning the temporal video structure or evolving events. In contrast, the fine-tuned model seemed to

TABLE II: Generated textual summaries (video clip #1).

Condition	Generated text summary
Without fine-tuning	*The video captures a group of children playing on an inflatable water slide in an outdoor setting. The slide is multi-colored, with sections in blue, yellow, and red, and the children are seen sliding down, climbing up, and splashing in the water at the bottom. The video is captioned with humorous and relatable observations about the children's behavior, such as their excitement and joy while playing on the slide.*
fine-tuning with visual features	*The video captures a young boy's exciting encounter with a large inflatable object in his backyard. Initially, the boy is seen walking towards the structure, seemingly unaware of its purpose. However, as the video progresses, it becomes clear that the object is a bouncy castle or an inflatable play area, judging by its size, color scheme, and the children's joyful reactions. The scene unfolds with the boy and other kids climbing, sliding, and playing on the inflatable, showcasing their carefree and energetic spirit.*

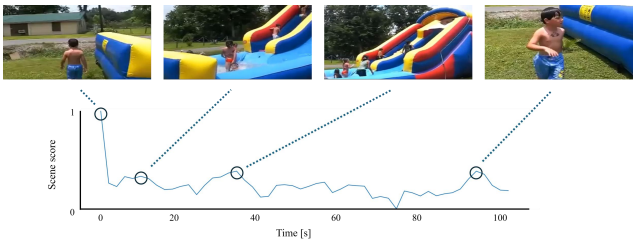

Fig. 3: Estimated importance scores (video clip #1).

TABLE III: Generated textual summaries (video clip #2).

Condition	Generated text summary
Without fine-tuning	*A car accident scene is captured on video, showing a vehicle partially wedged under a large piece of machinery, possibly a bulldozer or excavator. Several people are gathered around, observing the situation. The car is partially submerged in mud or dirt, and the surroundings suggest an urban or suburban setting with a paved road and some trees in the background.*
fine-tuning with visual features	*The video content captures a series of events at a level crossing, where a train is stopped, and a car is waiting to pass. The crossing gate is down, and people are gathered around the car, trying to assist or communicate with the driver. The scene changes to a car accident, where the vehicle has driven onto the railway tracks, causing it to derail. The video then shows a car being lifted by a crane, and a group of people are gathered around, observing the situation. The scene changes again to a car being stuck or wedged under a large piece of machinery, and people are trying to help or observe the situation.*

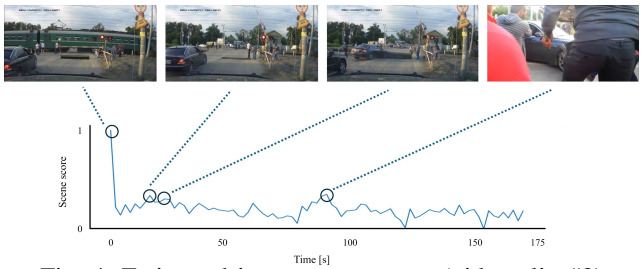

Fig. 4: Estimated importance scores (video clip #2).

have a better storytelling ability; it described a sequence of events occurring in the clip in a more context- and narrativity-aware manner. The textual descriptions were well-aligned with shots with high importance scores thanks to the joint training of the LLM and the score estimation head. We also found that visual cues help the model better align its output with the underlying narrative flow.

Interestingly, these improvements obtained with the fine-tuning came with trade-offs. The generated summaries tended to be longer and more exhaustive because details were listed without explicit prioritization. While describing sequential events could improve fidelity, it may also resulted in sprawling summaries that dilute the main message and reduce clarity. This highlights the challenge of balancing comprehensiveness and conciseness in the new task of narrativity-aware summarization.

We found that the beginning part of a video clip tended to be given higher importance scores. This bias likely stemmed from the autoregressive nature of text generation of the LLM that tends to attend to initial tokens, leading to underrepresentation of later context. To addressing this problem, we plan to investigate better attention control and temporal dependency modeling over the entire video clip. The current scoring module does not explicitly handle shot or scene boundaries; scores are computed uniformly over time. Incorporating scene-transition detection into the importance estimation is left for future work.

In our experiments, the performance gain from GoogLeNet features was modest. A broader ablation across visual backbones and fusion strategies is deferred to future work.

These findings illustrate both the promise and the challenges of integrating visual and textual cues for video summarization. Future work should focus on dynamically balancing modality contributions, mitigating position-dependent biases, and improving the ability to generate concise yet contextually rich summaries that remain robust across diverse content types.

V. Conclusion

This paper presents a multifaceted video summarization method that respects the narrative structure and evolution of generated textual summaries while incorporating visually-informed saliency. The key feature of the proposed method lies in its text-domain approach to narrativity-aware summarization, aligning the generated summary with human cognitive processes. From a technical perspective, the method combines a vision-language model (VLM) that generates shot-wise objective captions describing local semantics with a large language model (LLM) that produces a coherent, abstract summary of the entire video clip. Additionally, this model was extended to estimate the importance curve for an existing summarization benchmark.

We experimentally found that incorporating visual features enhances the ability to capture event sequences and narrative

TABLE IV: Generated textual summaries (video clip #3).

Condition	Generated text summary
Without fine-tuning	*A group of children are playing in a yard covered with fallen leaves, suggesting it's autumn. They are running, jumping, and laughing, enjoying the season. One child is lying on the ground, while the others are standing around. The video captures a candid moment of the children's playful interaction. The scene is set in a residential area with trees in the background, and a vehicle is parked on the side of the driveway.*
fine-tuning with visual features	*The video appears to capture a scene of children playing in a field or yard covered with fallen leaves, suggesting an autumn setting. The children are engaged in various activities, such as running, jumping, and playing with the leaves. As the video progresses, the scene shifts to a residential area with a house, driveway, and yard, where children are playing and interacting with each other and the environment. The video also shows a person, possibly a parent or adult, observing the children from a distance. The overall atmosphere of the video is one of joy and playfulness, capturing the innocence and wonder of childhood.*

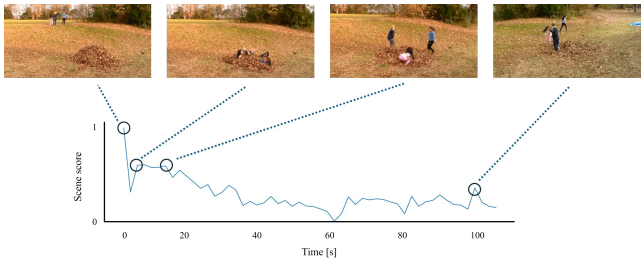

Fig. 5: Estimated importance scores (video clip #3).

flow in generated summaries. This approach mitigates the fundamental limitations of previous extractive or saliency-based methods. However, several challenges still remain unsolved. Fine-tuned outputs often become overly detailed, lacking clear event prioritization, and shot scoring exhibits positional biases, favoring early frames due to the autoregressive nature of the LLM. These findings highlight both the potential and complexity of shifting from saliency-based to viewer-centered, story-aware summarization. We plan to tackle these challenges through joint fine-tuning of the whole system including the VLM, LLM, and the importance estimator.

Acknowledgment

This work was partially supported by JST FOREST Grant No. JPMJFR2270 and JSPS KAKENHI Grant Nos. 24H00742, 24H00748, 25H01142, and 25K22841.

References

[1] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *ECCV*, 2014, pp. 505–520.
[2] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *CVPR*, 2015, pp. 5179–5187.
[3] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proc. IEEE*, vol. 109, no. 11, pp. 1838–1863, 2021.
[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *ICCV*, 2015, pp. 4489–4497.
[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 4724–4733.
[6] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018, pp. 318–335.
[7] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *ECCV*, 2016, pp. 766–782.
[8] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *CVPR*, 2017, pp. 2982–2991.
[9] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, 2020.
[10] A. Vaswani *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 1–11.
[11] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.
[12] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.
[13] T. Brown *et al.*, "Language models are few-shot learners," in *NIPS*, 2020, pp. 1877–1901.
[14] D. M. Argaw *et al.*, "Scaling up video summarization pretraining with large language models," in *CVPR*, 2024, pp. 8332–8341.
[15] M. J. Lee, D. Gong, and M. Cho, "Video summarization with large language models," 2025, arXiv:2504.11199.
[16] M. Barbara and A. Maalouf, "Prompts to summaries: Zero-shot language-guided video summarization," 2025, arXiv:2506.10807.
[17] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *ACCV*, 2019, pp. 39–54.
[18] B.-W. Chen, J.-C. Wang, and J.-F. Wang, "A novel video summarization based on mining the story-structure and semantic relations among concept entities," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 295–312, 2009.
[19] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *EMNLP*, 2004, pp. 404–411.
[20] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020, pp. 7871–7880.
[21] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *ICML*, 2020, pp. 11 328–11 339.
[22] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," 2023, arXiv:2302.13971.
[23] C. Szegedy *et al.*, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.
[24] W. Hong *et al.*, "CogVLM2: Visual language models for image and video understanding," 2024, arXiv:2408.16500.
[25] M. AI, "Meta Llama 3.1 models," https://github.com/meta-llama/llama-models, 2024, accessed: 2025-07-25.
[26] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685
[27] T. Dettmers, "bitsandbytes," https://github.com/TimDettmers/bitsandbytes, 2023.