Learning Semantic Representations for Video Summarization via MLLMs

Sumin Kim Minjun Kim Wonsik Shin Yebonn Han
Seoul National University
{sumink, minjun.kim, wonsikshin, bonnida}@snu.ac.kr

Abstract

We propose a novel approach to video summarization that leverages Multimodal Large Language Models(MLLMs) to extract high-dimensional semantic representations encompassing both visual and temporal context. Unlike prior caption-based methods, our framework encodes sequences of consecutive frames as contextual input to an MLLM, guided by an instruction prompt explicitly designed to estimate frame-level importance. The resulting embeddings can be directly integrated into existing summarization models without architectural modifications, significantly enhancing their expressive capacity. To address the computational overhead of MLLMs during inference, we further introduce a lightweight mapping function $F(x)$ that aligns low-level visual features from a pretrained GoogLeNet with the MLLM embedding space. Experiments on the SumMe and TVSum benchmarks demonstrate that our method outperforms existing state-of-the-art approaches in rank-based metrics such as Kendall's Tau and Spearman's Rho. Ablation studies validate the importance of instruction design in guiding MLLM reasoning. Our results highlight the potential of MLLM-driven representations as a scalable and expressive foundation for generalizable video summarization.

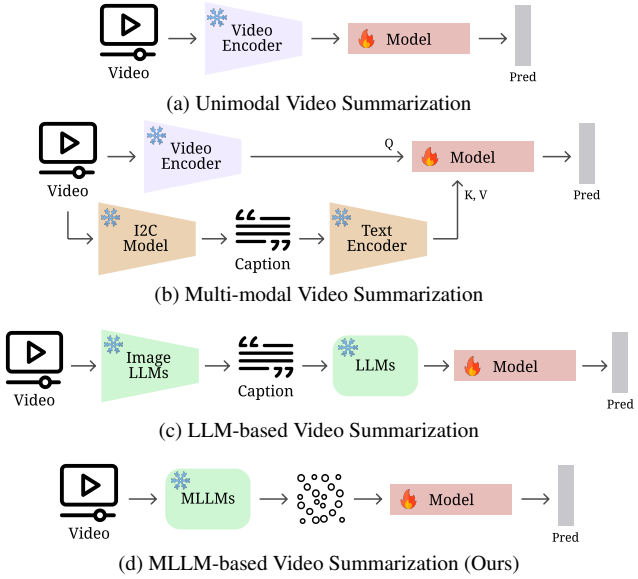


Figure 1. **Comparison of video summarization paradigms.** This figure presents the evolution of video summarization techniques from unimodal encoding, to visual-text fusion, followed by LLM-guided captioning, and culminating in our integrated MLLM-based framework.

1. Introduction

With the explosive growth in the production and consumption of video content in modern society, the demand for efficiently identifying key information from long videos has been steadily increasing. As a result, video summarization has emerged as a crucial technology across various applications such as search, navigation, recommendation, and personalized services by concisely compressing the essential content of original videos. In particular, the mobile-centric content consumption environment and the decreasing attention span of users have further highlighted the importance of summary quality in enhancing user experience.

Deep learning-based video summarization is typically defined as the task of predicting the importance of each frame and generating a summary that condenses the original video to approximately 15% of its length. Early approaches, as illustrated in Fig. 1a, focused primarily on visual information using CNN- [33] or LSTM-based [18] architectures. More recent models have adopted structures such as Transformers [4, 8, 15, 33, 45] and Diffusion models [32], which are capable of processing information more precisely, thereby improving performance. To address the limitations of vision-only approaches, cross-attention-based multimodal models [5, 11, 20, 28, 29] that incorporate textual information, as shown in Fig. 1b, have also been proposed. Recently, the importance of generalization capabilities—particularly those that can capture diverse user perspectives—has drawn attention to the language reasoning abilities of large language models(LLMs) [1–3, 39, 40]. For example, LLMVS [19], illustrated in Fig. 1c, gen-

erates captions for each frame, encodes them using an LLM [23, 25] to obtain embeddings, and predicts frame importance through a self-attention mechanism.

However, approaches like LLMVS [19] face inherent limitations. Captions generated on a per-frame basis often fail to capture critical semantic cues such as fine-grained visual details, scene dynamics, or subtle changes in human behavior. Moreover, the multi-stage pipeline that combines a multimodal LLM with a standard LLM introduces significant computational complexity and latency, making it unsuitable for real-time applications or large-scale video processing. These issues of semantic loss and computational inefficiency pose major obstacles to the practical deployment of such methods in real-world services and applications [22, 23, 38, 42, 43].

To address the limitations of frame-wise captioning—namely, semantic loss and inefficiency—this work leverages the strong generalization and contextual reasoning capabilities of large language models(LLMs). We propose an embedding extraction method using a Multimodal Large Language Model(MLLM) [24] that processes a sequence of consecutive video frames as a single contextual unit. Specifically, we encode eight consecutive frames as visual tokens and input them into the MLLM [24] alongside a natural language prompt designed to assess the importance of the central frame. We then extract high-dimensional embeddings from the final hidden states. These embeddings go beyond simple textual representations by capturing temporal flow and visual context, forming a rich semantic representation space. Our approach can be integrated into existing video summarization models without structural changes, demonstrating not only enhanced representational power but also strong potential for generalization and scalability.

Most prior work in video summarization has relied on visual features extracted from a pretrained GoogLeNet [37] model to construct summarization models. This approach offers high computational efficiency and ease of implementation, but its representational capacity is inherently limited. In contrast, embeddings derived from MLLMs [24] provide rich, high-dimensional representations that incorporate both visual and linguistic context, enabling superior summarization performance. However, their high computational cost and latency make them impractical for real-time or large-scale batch applications. To resolve this trade-off, we propose to learn a mapping function $F(x)$ that aligns visual features extracted from GoogLeNet [37] with the embedding space of MLLMs [24]. Inspired by Platonic representational space theory [13] and the success of vec2vec-style embedding alignment [14], we optimize $F(x)$ using MLLM-derived embeddings as supervision during training. At inference time, the model can generate semantically rich representations using only lightweight GoogLeNet [37] features. This design achieves a balance between expressiveness and efficiency, while maintaining structural compatibility and practical usability within existing video summarization pipelines.

To validate the effectiveness of the proposed approach, we conducted experiments on two widely used video summarization benchmarks: SumMe [10] and TVSum [34]. The results demonstrate that our MLLM-based representation outperforms traditional visual feature-based methods, achieving superior summarization performance. Notably, our method surpasses state-of-the-art baselines on ranking-based evaluation metrics such as Kendall's Tau [17] and Spearman's Rho [35]. These findings suggest that the proposed unified embedding strategy can more precisely capture the semantic essence of videos, highlighting its potential to advance both the expressiveness and practical applicability of video summarization.

The main contributions of this work are as follows:

- **High-dimensional semantic embeddings via MLLMs:** We extract embeddings by processing sequences of frames as contextual units, capturing integrated visual and temporal information to reduce semantic loss.
- **Mapping lightweight visual features into the MLLM space:** We propose a mapping function that aligns GoogLeNet features with MLLM embeddings, achieving a balance between expressive power and computational efficiency.
- **Empirical validation on standard benchmarks:** Our method outperforms prior approaches on key ranking-based metrics using the SumMe and TVSum datasets, demonstrating its effectiveness and generalizability.

2. Related Work

Video Summarization. Video summarization typically involves predicting the importance of each frame, segmenting the video using Kernel Temporal Segmentation(KTS), and selecting about 15% of frames using the 0–1 Knapsack algorithm [30]. Early methods adopted encoder–decoder architectures with bidirectional LSTMs [18], later replaced by attention-based models [4, 8, 15, 33, 45] for improved long-range dependency modeling. More recent work, such as CSTA [33], employs a sliding-window CNN to learn spatio-temporal patterns. Diffusion-based [32] approaches have also emerged to incorporate user subjectivity through stochastic sampling.

To address the limitations of visual-only methods, multimodal summarization has been explored [5, 11, 20, 28, 29], leveraging subtitles, transcripts, and audio to improve informativeness and semantic alignment. Vision-language models like CLIP-It [28] and TL:DW? [29] use cross-modal attention and saliency to align visual and textual information, while A2Summ [11] and SSPVS [20] enhance temporal coherence through contrastive and self-supervised learning.

Recent advances incorporate large language models(LLMs) to improve multimodal reasoning. V2Xum-LLaMA [12] and LLMVS [19] adopt prompt-based instruction tuning and caption-conditioned attention to build controllable, language-driven summarization frameworks. Building on this trend, we propose a novel method that leverages high-dimensional semantic representations from Multimodal Large Language Models(MLLMs) [24] to capture fine-grained visual semantics often missed by language-centric approaches.

MLLMs for Video Understanding. In recent studies on Multimodal Large Language Models(MLLMs) [6, 25, 26] for video understanding, early approaches primarily involved combining pre-trained video encoders with LLMs by aligning the extracted visual features to a language embedding space before feeding them into the LLM. Representative examples of this line of work include Video-ChatGPT [27] and Video-LLaMA [44]. Video-ChatGPT [27] incorporates frame-level spatiotemporal features obtained from a CLIP-based video encoder into the LLM, enabling detailed video descriptions. Meanwhile, Video-LLaMA [44] adopts a dual-branch architecture for vision-language and audio-language streams, utilizing modules such as Q-Former [21] and ImageBind [9] to integrate visual and auditory information, thereby extending its capacity to understand temporal dynamics and sound context. More recently, the field has shifted toward unified visual representation and joint multimodal training strategies that encompass both images and videos. For instance, Chat-UniVi [16] and Video-LLaVA [24] improve multimodal interaction learning by aligning image and video representations into a shared language feature space prior to LLM input, using mixed-modal training data. As such, the MLLM domain continues to advance its capability to comprehensively understand complex video content by focusing on modality alignment and integrated representation through joint training.

Semantic Embedding Alignment. The Platonic Representation Hypothesis [13] argues that as models become larger and more general-purpose, their representation spaces across heterogeneous modalities such as vision and language tend to converge into a shared semantic space. This perspective provides the conceptual basis for our design, in which lightweight visual features are aligned to the embedding space of an MLLM, enabling the two modalities to share a unified semantic representation. Meanwhile, vec2vec [14] demonstrates that the geometric structure of embedding spaces is sufficiently universal across models, allowing for unsupervised alignment between different embedding spaces without any paired data. This notion of unsupervised alignment offers both theoretical support and practical inspiration for our approach of learning a mapping function $F(x)$ that projects lightweight visual features into the MLLM semantic space—ultimately enabling real-time video summarization that balances expressiveness and efficiency.

3. Method

We describe the overall architecture of our model. Sec. 3.1 defines the video summarization task, and Sec. 3.2 explains how feature vectors are extracted using an MLLMs [24]. Sec. 3.3 introduces the method for aligning visual features from GoogLeNet [37] to the MLLMs embedding space. Sec. 3.4 describes how frame-level importance scores are predicted. Finally, Sec. 3.5 presents the training objective used for model optimization.

3.1. Problem Definition

Given a video $V = [v_1, v_2, \ldots, v_T] \in \mathbb{R}^{T \times H \times W \times 3}$, where T is the number of frames and H, W denote the height and width of each frame, respectively, the goal of video summarization is to predict a sequence of importance scores $s = [s_1, s_2, \ldots, s_T] \in \mathbb{R}^{T \times 1}$. A higher value of s_t indicates that the corresponding frame v_t is more likely to be included in the summary.

3.2. Multi-modal LLMs

In this section, we propose a method for extracting frame-level importance embeddings using a Multimodal Large Language Model(MLLM). We utilize a pre-trained MLLM [24] to process a sequence of 8 consecutive frames $I_0, \ldots, I_7$, where I_4 is considered the center frame and is provided as the visual input. Simultaneously, a dialogue-style instruction in an *instruction-following* format is given as the text input, guiding the model to assess how important the center frame is for summarizing the entire video. The final instruction is accompanied by the following examples:

> A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <image> <image> <image> <image> <image> <image> <image> <image>
>
> You are given 8 consecutive video frames represented as image tokens. The center frame(index 4) occurs at time t. Evaluate how important this center frame is for summarizing the video, considering its visual uniqueness and relevance to the overall narrative.
>
> Example 1) Score: 0.87 Explanation: The frame shows a key action different from others.
>
> Example 2) Score: 0.12 Explanation: The

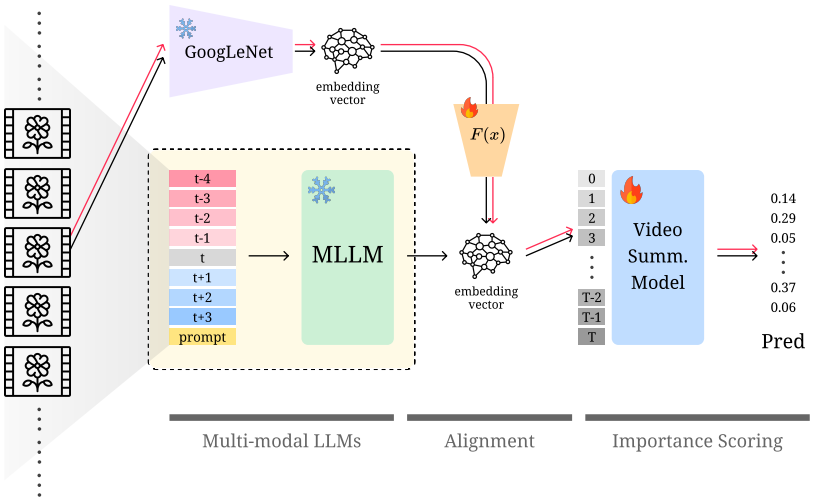


Figure 2. **Overview of the proposed architecture.** The model extracts frame-level embeddings using a pretrained GoogLeNet [37] and feeds them into a Multi-modal Large Language Model(MLLM) [24] with temporal prompts. The MLLM output is aligned with visual embeddings via a mapping function $F(x)$, and the resulting unified representations are used by the summarization model to assign importance scores for each video segment.

frame is nearly identical to adjacent ones.

Example 3) Score: 0.61 Explanation: The frame contains a common action but has some unique visual elements that contribute to the overall narrative.

Example 4) Score: 0.43 Explanation: The frame is similar to others but has a slight variation that makes it relevant to the narrative.

Now, respond for the current input.

Along with this textual prompt, each frame I_i is embedded into a D-dimensional visual feature vector $\mathbf{v}_i$ via the MLLM's built-in vision encoder, while the textual instruction is tokenized into a sequence of tokens $(w_1, \ldots, w_L)$ [24]. These visual tokens $(\mathbf{v}_0, \ldots, \mathbf{v}_7)$ and textual tokens $(w_1, \ldots, w_L)$ are concatenated to form the input sequence, which is then processed by the MLLM's transformer. The transformer outputs a hidden state vector $h_j^{(L)} \in \mathbb{R}^D$ for each input token at its final layer.

Finally, the average of these hidden state vectors is used as the semantic representation embedding of the center frame, defined as follows:

$$\mathbf{f}_{\text{MLLM}} = \frac{1}{N} \sum_{j=1}^{N} h_j^{(L)}, \tag{1}$$

Here, N denotes the total number of input tokens (both visual and textual). The resulting embedding $\mathbf{f}$MLLM serves as a contextual semantic representation of the center frame and is integrated into the input features of the downstream video summarization model. For example, by using $\mathbf{f}$MLLM as an additional input, the model's representational power can be enhanced without altering its architecture, ultimately leading to improved summarization performance.

3.3. Learning Semantic Representaions

To learn semantic representations from video frames, we designed a method to map the visual feature vectors extracted from a pre-trained GoogLeNet [37] model into the high-dimensional semantic embedding space of a large language model(MLLM) [24]. For each frame, we generate training data (x, y) by pairing the feature vector x obtained from GoogLeNet [37] with the embedding vector y produced by the MLLM for the same frame. The objective is to learn a transformation function $F : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_s}$ such that for any given visual feature x, $F(x) \approx y$.

Building on the idea from the vec2vec [14] paper, we implemented the transformation function $F(x)$ as a multilayer perceptron. Since an image feature vector is a fixed-length embedding with no spatial structure, we chose an MLP architecture over a CNN [14]. We incorporated residual connections, layer normalization, and the SiLU activation function into the network design to ensure stable training and rich representational capacity even for a deep network [14]. Furthermore, the output dimension of $F(x)$ was set to match the MLLM embedding dimension d_s, ensuring that $F(x)$ operates in the same representation space as the MLLM.

While the original vec2vec [14] work performed fully unsupervised training with additional loss terms (such as cycle consistency and vector space preservation), in our

work we leverage aligned data pairs and thus learn an effective mapping using only the core adversarial training objective, without those extra constraints. The theoretical justification for this embedding alignment approach is grounded in the Platonic representation hypothesis [13]. According to this hypothesis, embedding spaces learned by different models share a common underlying geometric structure, and a sufficiently expressive transformation can map vectors from one space to another while preserving their semantic content.

The learned mapping function $F(x)$ plays a crucial role in the subsequent video summarization stage. In the summarization process, we can obtain MLLM-level semantic embeddings for each frame simply by feeding the frame's GoogLeNet [37] feature through F, without needing to run the MLLM [24] for each frame. This means that visual information can be utilized as linguistic semantics almost in real time, greatly improving efficiency when processing large-scale video data.

In other words, the proposed semantic representation learning method effectively aligns the low-level features from a pre-trained vision model with a high-dimensional semantic space, enabling summarization based on the semantic importance of each frame. This can be seen as a form of knowledge distillation from the MLLM into the summarization model, integrating MLLM-level semantic understanding into the summarization process without directly relying on the MLLM at inference time.

3.4. Importance Scoring

The high-dimensional semantic embeddings extracted through our proposed method can be directly utilized as inputs to various video summarization models. Since the primary focus of this study is on enhancing representational power, the frame-level importance scoring must be integrable without modifying the architecture of existing models. To achieve this, we use the transformed embedding $F(x)$—obtained by mapping the GoogLeNet [37] features of each frame into the semantic space—as the input, allowing existing importance prediction models to process it without additional adaptation.

Ultimately, our approach enhances performance by altering only the input representation while preserving the original output structure of existing models. This design choice ensures that the proposed embeddings can serve as a drop-in replacement, thereby enabling fair comparisons with prior methods. Moreover, by eliminating the need for structural changes, our method offers both practicality and general applicability, making it suitable for a wide range of model designs.

3.5. Training Objective

Our overall framework is trained with two main objectives: (1) frame-level importance prediction for video summarization, and (2) embedding transformation that aligns lightweight visual features with the MLLM semantic space. To this end, we employ dedicated loss functions tailored to each objective.

First, the loss for frame importance prediction is defined based on Mean Squared Error(MSE). The model minimizes the squared error between the predicted importance score $\hat{s}_t$ and the ground-truth score s_t for each frame. The loss is formulated as follows:

$$\mathcal{L}_{\text{s}} = \frac{1}{T}\sum_{t=1}^{T}(s_t - \hat{s}_t)^2 \tag{2}$$

This loss encourages the extracted embeddings (either $\mathbf{f}_{\text{MLLM}}$ or $F(x)$) to accurately reflect the semantic importance of each frame.

Second, the loss for aligning the low-dimensional visual features from GoogLeNet [37] to the high-dimensional semantic space of MLLM is constructed using an adversarial framework, inspired by the vec2vec [14] method. Specifically, the mapping function $F(x)$, acting as a generator, is trained to produce outputs indistinguishable from real MLLM embeddings y, while a discriminator $D(\cdot)$ is trained to differentiate between real and generated embeddings. The corresponding loss is defined as:

$$\mathcal{L}_{\text{v}} = \mathbb{E}_{y\sim P_y}[\log D(y)] + \mathbb{E}_{x\sim P_x}[\log(1 - D(F(x)))] \tag{3}$$

Through this adversarial training, the transformed vector $F(x)$ is aligned to reside in the same semantic space as y, the MLLM embedding for the same frame.

4. Experiments

We conduct a comprehensive set of experiments to evaluate the effectiveness of our proposed method. Specifically, Sec. 4.1 describes the experimental setup, including the datasets, baselines, evaluation metrics, and implementation details. Sec. 4.2 provides additional implementation details, including model architecture and training procedures. In Sec. 4.3, we present quantitative comparisons with state-of-the-art methods on standard benchmark datasets. Sec. 4.4 offers ablation studies that analyze the impact of different instruction types and design choices. Finally, Sec. 4.5 presents qualitative results on representative examples, highlighting the interpretability and semantic precision of our approach.

4.1. Experimental Setup

Datasets. We evaluate on two standard benchmarks: SumMe [10], and TVSum [34]. SumMe [10] comprises 25

Model	SumMe [10]		TVSum [34]	
	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$
Random [31]	0.000	0.000	0.000	0.000
Human [31]	0.205	0.213	0.177	0.204
Visual				
VASNet [7]	0.160	0.171	0.160	0.170
DSNet-AB [45]	0.051	0.059	0.108	0.128
DSNet-AF [45]	0.037	0.046	0.113	0.135
DMASum [41]	0.063	0.074	0.098	0.115
PGL-SUM [4]	0.065	0.072	0.206	0.157
MSVA [8]	0.200	0.230	0.204	0.233
iPTNet [15]	0.101	0.119	0.197	0.230
CSTA [33]	0.246	0.274	0.192	0.255
Visual + Text				
CLIP-It [28]	0.109	0.120	0.108	0.147
TL;DW? [29]	0.111	0.128	0.142	0.167
A2Summ [11]	0.108	0.129	0.137	0.198
SSPVS [20]	0.192	0.257	0.181	0.238
Argaw et al. [5]	0.165	0.231	0.220	0.268
LLMVS [19]	0.253	0.282	0.211	0.275
Ours	**0.267**	**0.298**	**0.225**	**0.296**

Table 1. **Comparison with SOTA Models on SumMe [10] and TVSum [34] dataset.** The table categorizes the compared methods into three groups: (1) random and human baselines, (2) models relying solely on visual features, and (3) models incorporating both visual and textual modalities. Our proposed approach achieves superior performance across both SumMe and TVSum datasets, setting a new benchmark among existing methods.

user-generated videos with multiple human summaries per video. TVSum [34] includes 50 YouTube videos, segmented into 2 second-long shots and scored by 20 annotators.

Evaluation Metrics. For evaluation, we compute Kendall's tau (τ) [17] and Spearman's rho (ρ) [35] on the SumMe [10] and TVSum [34] datasets. Although F1-score has been widely adopted in prior video-summarization work, [30] demonstrated its susceptibility to segmentation biases and poor semantic alignment. Accordingly, we restrict our evaluation to kTau, and sRho.

4.2. Implementation Details

In our implementation, we employ Video-LLaVA [24] as the Multimodal Large Language Model(MLLM) to extract high-level semantic embeddings, while visual features are obtained from a pre-trained GoogLeNet [37]. For the video summarization model, we adopt the CSTA [33] architecture, which serves as a representative baseline to validate the general applicability of our embedding method. Notably, our framework is model-agnostic and can be integrated with any existing importance prediction architecture without structural modifications.

Instruction Type	SumMe [10]		TVSum [34]	
	$\tau \uparrow$	$\rho \uparrow$	$\tau \uparrow$	$\rho \uparrow$
Summarize	0.205	0.223	0.209	0.233
Importance Score	**0.267**	**0.298**	**0.225**	**0.296**

Table 2. **Ablation studies on instruction design.** The table compares the effect of different instruction types given to the MLLM, showing that the "Importance Score" instruction leads to better performance.

The MLLM processes a sliding window of 8 consecutive frames, generating one semantic embedding per window [24]. This windowing operation is applied across the entire video in a sequential manner, resulting in per-frame representations aligned with the narrative context. Due to instability observed in early-stage training, we exclude the GAN-based embedding alignment from this study. This component will be revisited in Sec. 5.

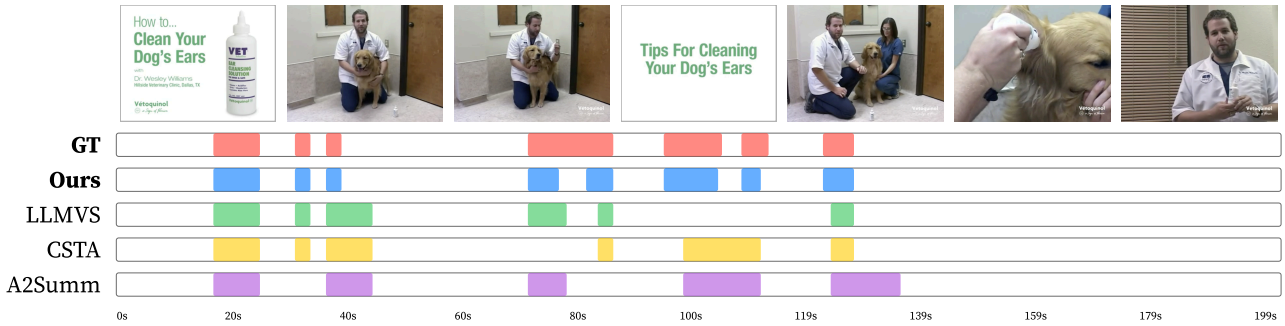
All experiments are conducted using NVIDIA A6000 GPUs. We set the learning rate to 1e-3, the weight decay to 1e-7, and train with a batch size of 1 throughout all stages.

4.3. Performance Comparison

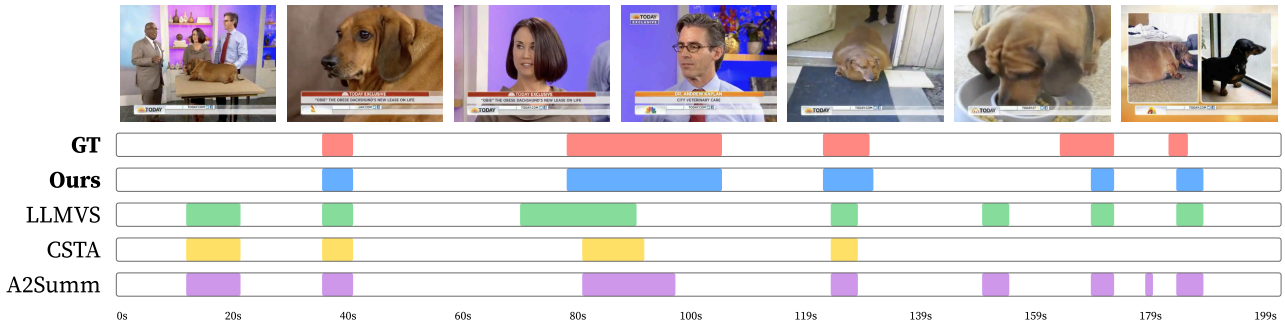
Tab. 1 reports the performance comparison on the SumMe [10] and TVSum [34] datasets in terms of Kendall's tau (τ) [17] and Spearman's rho (ρ) [35], which are rank-based correlation metrics. Our proposed method achieves the best performance across all models, recording $\tau = 0.267$, $\rho = 0.298$ on SumMe, and $\tau = 0.225$, $\rho = 0.296$ on TVSum. Compared to recent state-of-the-art models such as CSTA [33], MSVA [8], and LLMVS [19], our approach consistently outperforms them, demonstrating the effectiveness and generalization capability of MLLM-based semantic embeddings.

Notably, our method attains strong performance without relying on any segment alignment procedure. While many existing methods incorporate temporal segmentation or post-processing to enhance summarization quality, our model solely leverages frame-level semantic embeddings to achieve high accuracy. This suggests that the learned embeddings are capable of capturing the underlying semantic flow of videos, even in the absence of explicit structural alignment. Furthermore, this also implies that incorporating segment-level alignment in future work could lead to additional performance gains.

Overall, these results validate the potential of our proposed semantic embedding strategy as a strong drop-in replacement for visual features in existing summarization models, and point to its promise as a core component for future multimodal video summarization frameworks.



(a) TVSum video 49. "TODAY- Obie the obese dog works toward weight loss"



(b) TVSum video 15. "How to Clean Your Dog's Ears - Vetoquinol USA"

Figure 3. **Qualitative results on the TVSum dataset.** The comparisons illustrate that our method yields more coherent and informative summaries by capturing key events and transitions, demonstrating superior alignment with human-generated summaries compared to existing baselines.

4.4. Ablation Studies

To evaluate the effect of instruction design on the quality of semantic embeddings, we conduct an ablation comparing two prompt types: Summarize, which asks for a short description of the center frame, and Importance Score, which explicitly requests a score reflecting the frame's relevance to the summary.

As shown in Tab. 2, the Importance Score instruction yields substantial relative improvements over Summarize across all metrics. On the SumMe dataset, τ improves by 30.2% (from 0.205 to 0.267) and ρ by 33.6% (from 0.223 to 0.298). On the TVSum dataset, τ improves by 7.7% (from 0.209 to 0.225), and ρ by 27.0% (from 0.233 to 0.296), demonstrating consistent gains.

These results indicate that prompting the model to evaluate importance rather than generate descriptions leads to more discriminative and semantically aligned embeddings. The numerical scoring format better aligns with the model's reasoning capabilities and the nature of the summarization task, resulting in more effective representation learning.

4.5. Qualitative Results

Fig. 3 presents qualitative comparisons on two example videos from the TVSum dataset. Each row illustrates the selected key segments over time by different methods, where the top row (GT) represents the human-annotated ground truth summaries aggregated from multiple users.

As shown in Fig. 3a, our model produces a summary that closely aligns with the ground truth, effectively capturing key transitions and salient moments such as behavioral changes and interview scenes. In contrast, baseline methods like LLMVS [19] and CSTA [33] include redundant or less informative segments, failing to fully reflect the semantic flow of the video. Similarly, as illustrated in Fig. 3b, our approach demonstrates superior alignment with the ground truth by selectively including meaningful steps in the dog ear-cleaning process while omitting visually repetitive or trivial content. Other methods either miss critical actions or generate overly fragmented summaries, which leads to reduced coherence and informativeness.

These qualitative results reaffirm the strength of our MLLM-based semantic representations in identifying contextually significant segments without explicit segment alignment. The improved alignment with human preferences highlights the model's potential for real-world summarization tasks where interpretability and relevance are crucial.

5. Future Work

While this study demonstrates the potential of MLLM-based representations for video summarization, several important directions remain for future exploration.

First, achieving stable training for semantic embedding alignment remains a challenging task. Although we initially adopted an adversarial training framework to learn the mapping function $F(x)$ that aligns GoogLeNet [37] features to the MLLM embedding space, our early experiments revealed significant convergence difficulties. This instability is primarily attributed to the high dimensionality and complex distribution of MLLM embeddings, which makes it difficult to maintain equilibrium between the generator and discriminator. In future work, we aim to integrate stabilization techniques such as gradient penalty, feature matching loss, or two-stage pretraining to improve convergence behavior and alignment quality. Moreover, we plan to explore geometry-aware alignment strategies—such as manifold regularization or geometric consistency loss—that leverage the intrinsic structure of the embedding space.

Second, to rigorously evaluate the generalization and scalability of our approach, it is essential to conduct experiments on large-scale datasets. The recently introduced Mr. HiSum [36] benchmark provides a highly diverse and realistic setting for video summarization, with a broad range of topics, video lengths, and human-annotated summaries. However, due to the substantial memory requirements and computational demands of MLLMs, we were unable to perform experiments on Mr. HiSum [36] in this study. In future work, we plan to revisit this direction with access to more powerful compute resources and optimized memory management strategies, in order to validate whether our method maintains strong performance at scale. Demonstrating efficacy on large-scale benchmarks will be critical for establishing the practical viability and robustness of MLLM-based summarization techniques.

Ultimately, our future efforts will focus on enhancing the theoretical grounding and empirical stability of semantic alignment, while validating the proposed framework across diverse datasets to ensure its extensibility and applicability to real-world multimodal video understanding scenarios.

6. Conclusion

This paper presents a novel approach to enhancing video summarization by leveraging Multimodal Large Language Models(MLLMs). We propose a semantic embedding extraction framework that processes consecutive video frames as contextual units, generating high-dimensional representations that integrate both visual and temporal cues. These embeddings can be seamlessly integrated into existing summarization models without architectural changes. To address the computational inefficiency of MLLMs, we introduce a lightweight mapping function $F(x)$ that aligns visual features from a pre-trained GoogLeNet [37] with the MLLM embedding space, enabling efficient inference. Experimental results on the SumMe [10] and TVSum [34] benchmarks demonstrate that our method consistently outperforms prior state-of-the-art techniques in rank-based metrics. Further ablation studies and qualitative analyses confirm the effectiveness and generalization capability of the proposed semantic embeddings. Overall, this work provides a practical framework for distilling MLLM knowledge into efficient representations, offering a scalable foundation for future multimodal video understanding and summarization research.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 1

[4] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 226–234, 2021. 1, 2, 6

[5] Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Dernoncourt, and Joon Son Chung. Scaling up video summarization pretraining with large language models. In *CVPR*, pages 8332–8341, 2024. 1, 2, 6

[6] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *CVPR*, 2024. 3

[7] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention, 2018. 6

[8] Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. Supervised video summarization via multiple feature sets with parallel attention. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6s. IEEE, 2021. 1, 2, 6

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 3

[10] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*. Springer International Publishing, 2014. 2, 5, 6, 8

[11] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *CVPR*, pages 14867–14878, 2023. 1, 2, 6

[12] Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *AAAI*, pages 3599–3607, 2025. 3

[13] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 2, 3, 5

[14] Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025. 2, 3, 4, 5

[15] Hao Jiang and Yadong Mu. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16388–16398, 2022. 1, 2, 6

[16] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 3

[17] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 2, 6

[18] Luis Lebron Casas and Eugenia Koblents. Video summarization with lstm and deep attention models. In *International conference on multimedia modeling*, pages 67–79. Springer, 2018. 1, 2

[19] Min Jung Lee, Dayoung Gong, and Minsu Cho. Video summarization with large language models. *arXiv preprint arXiv:2504.11199*, 2025. 1, 2, 3, 6, 7

[20] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. Progressive video summarization via multimodal self-supervised learning. In *CVPR*, pages 5584–5593, 2023. 1, 2, 6

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3

[22] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *ECCV*, 2024. 2

[23] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2

[24] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2, 3, 4, 5, 6

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 2, 3

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 3

[27] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 3

[28] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *NeurIPS*, 34: 13988–14000, 2021. 1, 2, 6

[29] Medhini Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *ECCV*, pages 540–557. Springer, 2022. 1, 2, 6

[30] Mayu Otani, Yuta Nakahima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6

[31] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7596–7604, 2019. 6

[32] Zirui Shang, Yubo Zhu, Hongxi Li, Shuo Yang, and Xinxiao Wu. Video summarization using denoising diffusion probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6776–6784, 2025. 1, 2

[33] Jaewon Son, Jaehun Park, and Kwangsu Kim. Csta: Cnn-based spatiotemporal attention for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18856, 2024. 1, 2, 6, 7

[34] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, pages 5179–5187, 2015. 2, 5, 6, 8

[35] C Spearman. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39 (5):1137–1150, 2010. 2, 6

[36] Jinhwan Sul, Jihoon Han, and Joonseok Lee. Mr. hisum: a large-scale dataset for video highlight detection and summarization. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2, 3, 4, 5, 6, 8

[38] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. *arXiv preprint arXiv:2502.21271*, 2025. 2

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[41] Junyan Wang, Yang Bai, Yang Long, Bingzhang Hu, Zhenhua Chai, Yu Guan, and Xiaolin Wei. Query twice: Dual mixture attention meta learning for video summarization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4023–4031, 2020. 6

[42] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 2

[43] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer, 2024. 2

[44] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 3

[45] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020. 1, 2, 6