Agentic Video Summarization via Self-Reflecting Multimodal Understanding

Miaotian Guo[1,2] * Shuguang Dou[2] † Yin Li[2] Aidong Men[1]† Dongsheng Jiang[2]†

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Huawei Technologies Co., Ltd

{gmt, menad}@bupt.edu.cn, {doushuguang, liyin9, jiangdongsheng1}@huawei.com

Abstract

*The rise of AI agents powered by large language models (LLMs) has transformed intelligent systems by enabling autonomous tool utilizing, reasoning, and action across diverse tasks. Despite this rapid progress, existing video summarization approaches primarily focus on feature extraction or frame-level importance regression but lack the autonomous reasoning, self-correction, and decision-making capabilities that define true agent-based intelligence. To bridge this gap, we propose **AgenticVS**—the first agentic workflow for video summarization that leverages multimodal large language models (MLLMs) to complete the summarization–verify–reflection loop in a fully autonomous manner. Rather than designing new architectures for feature extraction or regression, we exploit the understanding and reflective reasoning abilities of MLLMs to build an adaptive summarization framework with a self-reflecting workflow. Experiments on SumMe and TVSum demonstrate that our agentic workflow outperforms state-of-the-art methods, enhancing interpretability, adaptability, and paving the way for agent-based multimodal video understanding.*

1. Introduction

With the rise of social media platforms, people frequently capture daily activities and special moments, generating vast amounts of video content. As mobile devices simplify video creation and sharing, the volume of online videos has surged. [33] In response to the fast pace of modern life and the short-video era, the demand for efficiently digesting and understanding long-form videos has grown. Consequently, video summarization has become essential for applications such as search, navigation, and recommendation, aiming to condense videos while preserving key semantics and narrative coherence. This growing need for rapid yet meaningful content consumption highlights the importance of intelligent, context-aware summarization frameworks that adapt to diverse domains and user preferences.

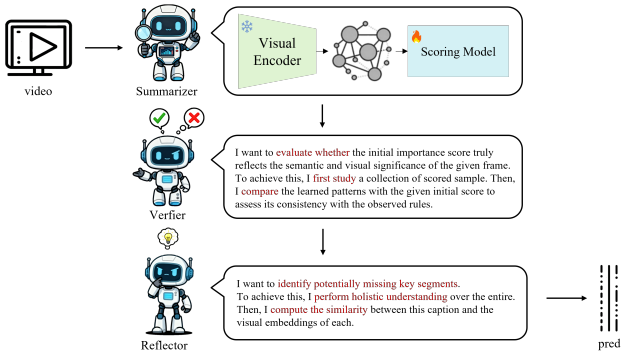


Figure 1. **An illustration of AgenticVS's self-reflecting strategy applied to video summarization**. The task is decomposed into three roles—Summarizer, Verifier, and Reflector—which systematically generate initial importance scores, perform verification, and identify missing keyframes. This role-based pipeline enables more human-like video reasoning compared to traditional methods.

Previous video summarization methodologies can be primarily categorized into two main categories: extractive and abstractive. Extractive methods often involve selecting keyframes based on labeled data and criteria like objects, events, and user perception. Among them, deep learning–based approaches primarily predict the importance score of each frame to generate a summary. Early approaches mainly relied on CNN- [28, 39, 62] or LSTM-based [6, 18, 31] architectures to extract visual features, based on which they regressed frame-level representations to predict importance scores for keyframe selection. However, these methods lacked high-level semantic understanding and global temporal reasoning across frames, limiting their ability to capture complex video dynamics.

In contrast to extractive approaches, recent studies have

*Work done during an internship at Huawei.
†Corresponding authors.

focused on the visual understanding and reasoning abilities of multimodal large language models (MLLMs), leading to an abstractive manner of video summarization. LLMVS [19] first introduced LLMs into video summarization by using LLaVA [24] for caption generation and Llama-2 [46] for caption embedding. However, this framework loses fine-grained visual cues during captioning, making embeddings less discriminative. To mitigate this, Kim et al. [36] employed VideoLLaMA [56] to directly extract visual embeddings, yielding richer and more expressive representations. Despite these advances, MLLMs still require manually crafted text prompts that clearly define frame-level importance criteria—an effort-intensive and subjective process that hinders reproducibility and scalability.

Unlike the above approaches that directly employ (M)LLMs for passive video processing, an agent is an autonomous (M)LLM-based system with active perception and decision-making capabilities. Researchers have been actively developing agents for a variety of tasks, especially in video reasoning. For example, VideoAgent [52] embraces the agent-based system to mimic human cognitive processes and establish a self-feedback mechanism for long-form video understanding. Agents offer key advantages such as temporal planning [11, 38, 44, 51], tool utilization [22, 32, 34, 42, 54], and iterative refinement [37, 53]. To fully exploit these strengths, we propose AgenticVS — the first agentic workflow for video summarization that incorporates self-reflection and automatic calibration to iteratively refine summaries.

In this study, instead of adopting the "one-time output" approach used by traditional models, we aim to more flexibly leverage the capabilities of MLLMs in deep reasoning and self-reflecting through an agent-based workflow. Within this framework, the MLLM can autonomously execute a closed-loop cycle of predicting and self-evolving.

To achieve this, we propose three atomic agents, the *Summarizer*, which aligns video- and image-level embeddings in V2I Alignment module and make the initial importance score prediction; the *Verifier*, which adopts a multi-round memory mechanism for initial score evaluation; and the *Reflector*, which refines the scores with a self-reflection correction mechanism for missing keyframes searching.

Specifically, the Summarizer utilize both a video encoder and an image encoder to extract visual embeddings and constructs a V2I Alignment module to align their embeddings in a shared feature space. Considering the Summarizer lacks a holistic understanding of the overall video content, we design a multi-round memory mechanism for the Verifier in order to re-evaluate the initial scores produced by the Summarizer and get prepared for the following calibration. The Reflector uses a self-reflection correction strategy, enabling the MLLM to generate a video text summary, and then leverages the CLIP [29] model to compute more precise calibration scores. This iterative process leverages the MLLM's memory and reasoning to decompose complex evaluation tasks, simulating human-like reasoning and avoiding the need for overly long textual prompts.

Our contributions can be summarized as follows:

1. We propose AgenticVS, the first approach to integrate agent-based workflows into video summarization tasks. AgenticVS emulates human cognitive processes, enabling MLLMs to autonomously execute a closed-loop cycle of prediction and self-evolution.
2. We designed a Summarizer that enhances the accuracy of initial importance score regression. The Verifier employs a multi-round dialogue strategy for iterative re-evaluation of scores, fully leveraging the memory and reasoning capabilities of MLLM. The Reflector adopts a self-reflection correction mechanism to refine the initial scores and search for missing keyframes.
3. Experimental results show that the proposed approach exhibits extraordinary superiority compared to existing extractive VS methods and surpasses state-of-the-art baselines on SumMe [10] and TVSum [40].

2. Related Work

2.1. Video Summarization

Video summarization is a technique to generate a concise representation of a video that captures the main events and ideas into a short skim. A diverse set of video summarization approaches can be categorized into two main categories: extractive and abstractive methods. [1] Extractive video summarization methods select importance segments from the original video without modifying the content. Most works focus on leveraging deep learning techniques to evaluate the importance score of each frame. A notable direction in this domain employs LSTMs [13, 47, 55, 57, 59, 60], which are adapted to capturing both short- and long-range dependencies in sequential frames. Abstractive video summarization aims to create concise summaries of videos by generating new content that captures the essence of the original video. Most generic methods [14, 15, 58] temporally aggregate the most informative moments of an input video to compose a summary video. Among these works, CSTA [39], one of the most representative extractive methods, which initially extracts and concatenates frame features, has been widely utilized as a foundation model in video summarization. The appearance of LLMs presents new opportunities for video summarization. LLMVS [19] is the first to introduce LLMs into video summarization. Kim et al. [36] utilize MLLMs to learn semantic representations. Building on previous work, we further leverage inter-frame information to obtain feature embeddings with contextual relationships.

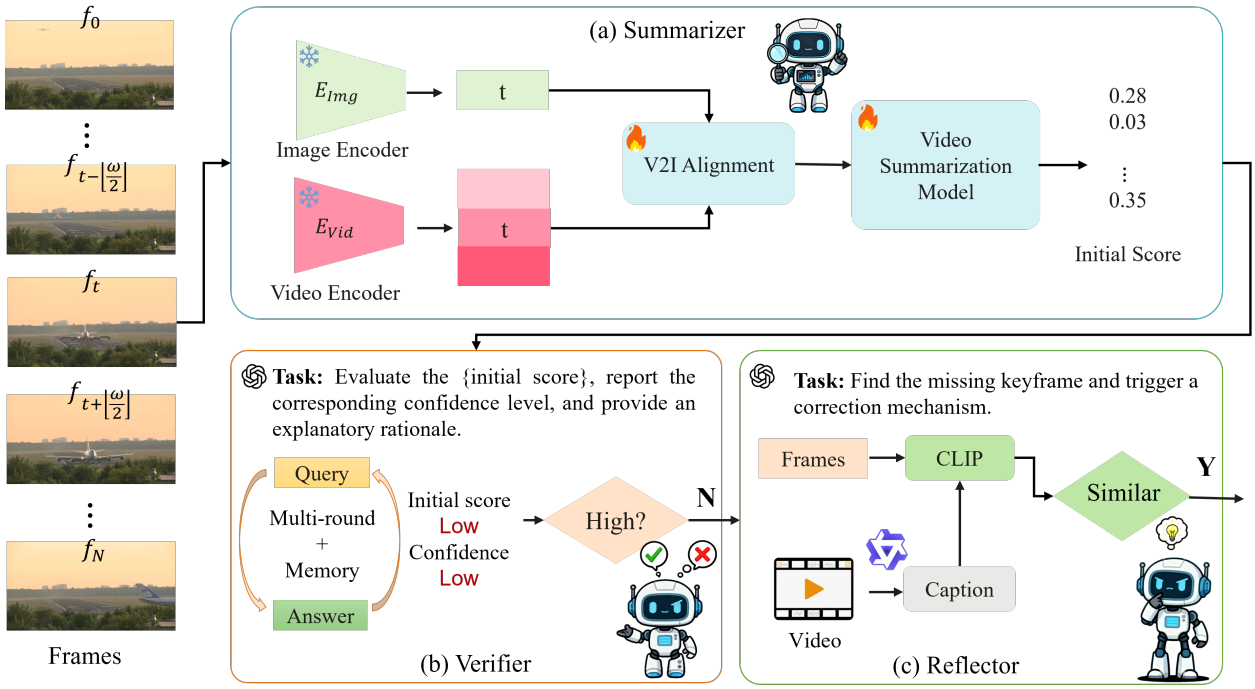


Figure 2. **The overall pipeline of our approach consists of three main atomic agents**. Given an input video, the Summarizer generates the initial predicted scores, after which the Verifier evaluates these scores and provides explanatory feedback. Finally, the Self-Reflector computes the similarity between frames and the overall caption to identify missing keyframes.

2.2. MLLMs for Video Understanding

Recent studies have seen the emergence of many excellent MLLMs [5, 24, 25] for video understanding.Video-ChatGPT [26] incorporates frame-level spatiotemporal features obtained from a CLIP-based video encoder into the LLM, enabling detailed video descriptions. Video-LLaMA [56] 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 [23] 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. 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. Inspired by this work, our approach applies Qwen2.5-VL [4] to leverage its ability to incorporate semantic information and provide a richer contextual understanding of video content.

2.3. Video Agents

Nowadays, the idea of AI agents, systems that can autonomously plan and execute tasks, has prevailed in recent AI research. Within the video understanding field, VideoAgent [52] reconceptualizes video comprehension as a sequential decision-making process, emulating how humans progressively gather and interpret visual information. In this setting, a video is treated as a dynamic environment where the agent decides whether to acquire additional evidence or to finalize its judgment. PyVision [61] is an interactive framework in which the model automatically generates, executes, and interactively refines Python code in response to multimodal user queries. Considering that certain visual tasks follow relatively fixed workflows and cannot be organized as fully autonomous agents for VQA or reasoning tasks, researchers have proposed agentic workflows, where MLLMs are employed to invoke appropriate models and construct more flexible and adaptive solutions. VADAR [27] guides the LLM to invoke dynamic APIs for performing spatial reasoning. AoTD [35] automatically generates Chain-of-Thoughts (CoTs) to decompose complex questions into sub-tasks. Motivated by this perspective, we propose an early exploration of integrating an agent-style reasoning mechanism into video summarization, leading to the development of our agentic video summarization framework.

3. Methodology

AgenticVS is inspired by the human cognitive process of generating video summarization and existing Video Agents [27, 52]. Given a video, video summarization models typically generate importance score predictions based on visual features. However, these predicted results cannot be automatically refined through model training or regression alone. In contrast, humans can perceive event transitions and iteratively adjust their understanding across the entire video. Inspired by this, we aim to emulate the processes of verification and self-reflection to more effectively exploit high-level semantic information.

We establish three primary atomic agents to decompose the video summarization task into three stages: (a) initial importance score prediction (section 3.2), (b) verify the confidence of the initial score (section 3.3), and (c) subsequent refinement (section 3.4), as illustrated in Fig. 2.

The complete prompts used in both the Verifier and Reflector modules are provided in the Appendix.

3.1. Problem Formulation

The core idea of video summarization lies in encoding video frames into feature representations, which are subsequently input into a video summarization model to predict their importance scores. Given a video $F = [f_1, f_2, \cdots, f_N] \in \mathbb{R}^{N\times 3\times H\times W}$k, where N denotes the number of frames, and H and W represent the frame height and width, respectively. The input F is transformed into $F' \in \mathbb{R}^{N\times D}$ through the frozen visual encoder, where D is the dimension of frame features. Then, the VS model (e.g., CSTA [39]) maps the features of each frame F' into single values, and a sigmoid function computes the importance scores $S \in \mathbb{R}^N$.

3.2. Summarizer: Video Representation Alignment with Image Representation

Previous works [19, 29, 39] mainly focus on the visual features extracted from a single static frame by an image-level encoder. Since humans tend to determine key frames based on the content and event transitions when summarizing a video, we adopt both a video encoder and an image encoder, and align their embeddings in a *shared feature space*. The structure of V2I Alignment is detailed in Fig. 3.

The core concept of V2I Alignment is to leverage the rich semantic priors of static images to overcome the limitations of pure video models in fine-grained semantic discrimination, while maintaining their ability to capture temporal dynamics. Given a frame f_t that requires encoding and importance score prediction, we first use GoogLeNet [43] as the image encoder E_{img} to extract the image-level embedding vector $v_{img} \in \mathbb{R}^{d_i}$. To obtain inter-frame embeddings, we group ω frames as a shot and follow the same settings as VideoMAEv2 [49] for the video encoder E_{vid}, which adopts the joint space-time cube of size $2 \times 16 \times 16$ as one

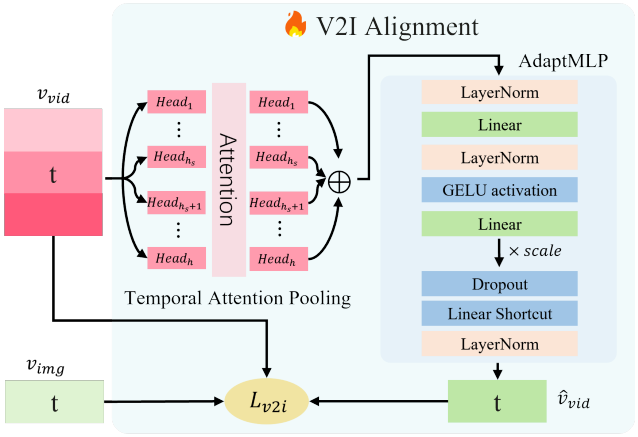


Figure 3. **The structure of V2I Alignment module.** Image features are used solely as alignment targets in the Summarizer during training, while only video-level features—after pooling and adaptMLP processing—are employed during inference.

token embedding to generate the video-level embeddings vector $v_{vid} \in \mathbb{R}^{\frac{\omega}{2}\times d_v}$.

To achieve alignment, we first compress the temporal dimension $\frac{\omega}{2}$ through temporal attention pooling based on the multi-head attention mechanism. In this way, v_{img} and v_{vid} share the same shape within their respective feature spaces. Subsequently, we introduce a mapping module equipped with an adapter to align the embeddings from both spaces, resulting in a unified visual embedding v_{v2i} that integrates both intra-frame and inter-frame information. Because the video-level embedding dimension d_v differs from the image-level dimension d_i, we choose an adapter architecture with layer normalization and the GELU activation function in the network design to ensure stable training and enhance representational capacity. For each frame f_t, we generate training data (v_{img}, v_{vid}). The objective of V2I Alignment is to learn a transformation function:

$$F_{v2i} : \mathbb{R}^{\frac{\omega}{2}\times d_v} \rightarrow \mathbb{R}^{d_i}. \tag{1}$$

The high-dimensional aligned embeddings extracted through our V2I Alignment method can be directly utilized as inputs to various video summarization models. Finally, the initial scores s_t are produced by the Summarizer.

3.3. Verifier: Multi-round memory mechanism for score confidence evaluation

Although the Summarizer considers information between adjacent frames, we observed that it still overlooks certain key segments in some visualizations (as shown in Fig. 5). This limitation arises because the Summarizer lacks a holistic understanding of the overall video content and event transitions. To address this issue, we leverage the video comprehension and reasoning capabilities of MLLMs to generate score confidence assessments through comparative

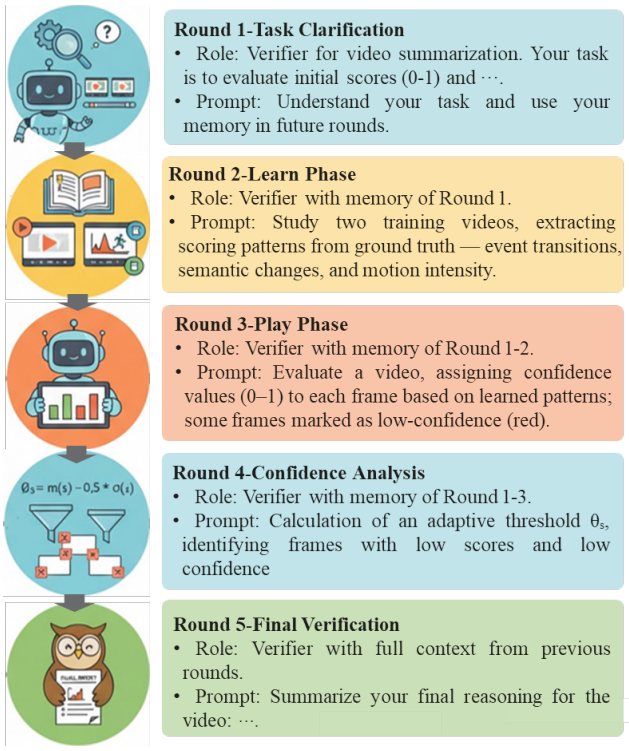


Figure 4. **Multi-round memory prompt for score confidence evaluation.** Only the simplified prompts are shown in the figure, while the complete versions are provided in the appendix.

analysis of video content.

For the Verifier, we design a multi-round memory prompt as shown in Fig. 4. First, Verifier learns human scoring criteria and underlying patterns during the learning phase. Subsequently, Verifier re-evaluates positions with initially low scores based on the scoring rules acquired in the previous phase, providing confidence assessments for these scores according to its learned judgments. Verifier then calculates an adaptive threshold $\theta_s = mean(s_t) - 0.5 * std(s_t)$ to identify frames with low scores and low confidence. After the evaluation, the Verifier determines the status of each initial score s_t based on the assessed confidence c_t and performs two possible actions:

- **Action 1: Search missing frames.** If s_t is low while c_t is also low, it indicates that the initially assigned low score is likely inaccurate, and the Reflector is applied to identify a potentially missing key frame.
- **Action 2: Pass to next video.** If s_t is low but c_t is high, it implies that the low score is reliable, and the corresponding frame can be confidently identified as a non-key frame.

3.4. Reflector: Self-Reflection correction mechanism for missing keyframes searching

After the Verifier makes an Action 1 decision, the Reflector performs self-reflection. In contrast to the Verifier, which directly queries the MLLM for inference and produces immediate results, the Reflector first leverages the MLLM to generate a video text summary and then utilizes the CLIP model to compute more precise calibration scores.

First, the MLLM is explicitly instructed to understand and summarize the video, focusing on aspects relevant to video summarization, such as the overall content, event transitions, and scene changes. It then generates appropriate captions, which are subsequently used as input to CLIP for scoring. Next, the MLLM perceives and summarizes the entire segment of the video to produce captions representing the content that should be included in the summary according to the model's understanding. Then, based on the frames identified by the Verifier as requiring recalibration, we extract the corresponding frame images $\{\hat{f}_t\}$. Finally, each frame $\hat{f}_t$ is paired with the caption and fed into the CLIP model to compute the cosine similarity between the image $\{\hat{f}_t\}$ and caption, which is used as the calibrated score $\{\hat{s}_t\}$to replace the inaccurate initial scores $\{s_t\}$.

$$\hat{s}_t = cos(E_v(\hat{f}_t), E_t(F_{MLLM}(V))) \tag{2}$$

Considering that the initial scores $\{s_t\}$ and the cosine similarities $\{\hat{s}_t\}$ computed by CLIP may differ in scale and cannot be directly substituted, we first normalize both types of scores. The CLIP-based calibrated scores are then rescaled to match the scale of the initial scores, ensuring that the re-evaluated scores provided by the Reflector can be correctly used as calibration results to replace the previous scores. Finally, following the same decision procedure as before, the key frame positions are re-determined according to the new importance scores, yielding the final video summarization results of AgenticVS.

3.5. Training Objective

Since the Verified and Reflector are training-free, we only train the V2I alignment module and the VS model.

For the V2I alignment module, we adopt a composite loss function that combines an MSE term with a residual regularization term. Specially, given the video-level features v_{vid} and the corresponding image-level features v_{img} for each frame $t = 1, \cdots, T$, the mapping function $F_{V2I}(\cdot)$ produces the aligned features $\hat{v}^t_{vid} = F_{V2I}(v^t_{vid})$. The loss function is formulated as:

$$\mathcal{L}_{\text{v2i}} = \frac{1}{T}\sum_{t=1}^{T} \left\|\hat{v}^t_{vid} - v^t_{img}\right\|_2^2 + \lambda \frac{1}{T}\sum_{t=1}^{T} \left\|\hat{v}^t_{vid} - v^t_{vid}\right\|_2^2, \tag{3}$$

where λ is set to 10^{-3} to balance the regularization term. The first term encourages the mapped features to align with

the image-level space, while the second term limits deviation from the original video-level embeddings as a residual constraint. This design maintains both alignment and stability, yielding more robust cross-model feature adaptation.

The VS model is trained by minimizing the squared error between the predicted initial importance score s_t and the ground-truth s_t^* for each frame. The loss is formulated as follows:

$$\mathcal{L}_{\int} = \frac{1}{T}\sum_{t=1}^{T}(s_t^* - s_t)^2. \tag{4}$$

This loss encourages the aligned embeddings to accurately reflect the contextual and intrinsic importance of each frame.

4. Experiments and Results

4.1. Experimental Settings

Datasets. We evaluate our approach and several state-of-the-art video summarization models [8, 19, 39] on two standard benchmarks, SumMe [10] and TVSum [40]. SumMe consists of 25 YouTube videos with 15–18 human-generated summaries per video, each limited to under 15% of the original length. TVSum includes 50 YouTube videos with title and category metadata, along with human-annotated frame-level importance scores every two seconds.

Evaluation Metrics. We evaluate AgenticVS using Kendall's τ [17] and Spearman's ρ [41] coefficients. Both are rank-based correlation measures used to assess the similarity between model-predicted and ground-truth scores. Although the F1-score has been widely adopted in video summarization, it often favors shorter clips over actual key shots due to length constraints [30, 45], making it an unreliable performance metric. Therefore, we base our evaluation primarily on τ and ρ.

Implementation Details. In our implementation, VideoMAEv2 processes a sliding window of eight consecutive frames to encode video-level embeddings, resulting in a space–time cube embedding of $\omega = 8$. For the video summarization model, we adopt the CSTA [39] architecture as a representative baseline to validate the general applicability of our Summarizer. For constructing the Verifier-vs and Reflector, we employ Qwen2.5-VL-7B-Instruct [50] as the MLLM to facilitate multi-turn, dialogue-based reasoning and decision-making. In the Reflector module, CLIP-ViT-B/32 [29] is used to compute cosine similarity for generating modified importance scores. All experiments are conducted on NVIDIA V100 GPUs with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, and a batch size of 1 across all training stages.

4.2. Performance Comparison

We compare our proposed AgenticVS model with three types of methods: (1) random and human baselines, (2) models that utilize visual features, and (3) models that incorporate both visual and textual features. The random and human performance metrics are obtained from [30].

Table 1. Comparison with Visual-Based and Visual–Text-Based Video Summarization Methods.

Method	SumMe		TVSum	
	τ	ρ	τ	ρ
Random [30]	0.000	0.000	0.000	0.000
Human [30]	0.205	0.213	0.177	0.204
Visual				
VASNet [7]	0.160	0.170	0.160	0.170
DSNet-AB [62]	0.051	0.059	0.108	0.129
DSNet-AF [62]	0.037	0.046	0.113	0.138
DMASum [48]	0.063	0.089	0.203	0.267
PGL-SUM [2]	-	-	0.206	0.157
MSVA [8]	0.200	0.230	0.190	0.210
iPTNet [15]	0.101	0.119	0.134	0.163
CSTA [39]	0.246	0.274	0.194	0.255
Visual+Text				
CLIP-It [29]	-	-	0.108	0.147
A2Summ [12]	0.108	0.129	0.137	0.165
SSPVS [20]	0.192	0.257	0.181	0.238
Argaw et al. [3]	0.130	0.152	0.155	0.186
LLMVS [19]	0.253	0.282	0.211	0.275
Agentic				
AgenticVS (ours)	**0.274**	**0.308**	**0.220**	**0.290**

Table 1 shows that AgenticVS outperforms all non-agentic methods—including both visual-only and visual–text approaches—across all four evaluation metrics on the SumMe and TVSum datasets, establishing itself as the top-performing model. Its performance substantially surpasses that of all purely visual methods. For instance, on the SumMe dataset, AgenticVS achieves a correlation of $\rho = 0.308$, notably higher than the best visual method, CSTA ($\rho = 0.274$). AgenticVS also outperforms all visual–text methods; for example, on TVSum, it attains $\rho = 0.290$, exceeding the best visual–text approach, LLMVS ($\rho = 0.282$). These results indicate that the agentic framework more effectively integrates visual and textual information through advanced reasoning, planning, and multimodal information fusion, thereby achieving a significant performance improvement.

4.3. Ablation Study

More ablation studies are provided in the appendix.

Workflow Establishment. We conduct ablation studies to evaluate the contribution of each component in our AgenticVS workflow. As shown in Table 2, integrating the V2I Alignment module into the Summarizer yields notable improvements: both τ and ρ increase by 15.2% on SumMe and by 20.8% and 19.8%, respectively, on TVSum, demonstrating that aligning the two feature spaces effectively en-

Table 2. Ablation for atomic agents in AgenticVS. In the baseline setting, the image-level and video-level embeddings are directly concatenated and reshaped before being fed into the CSTA model for training, without the agentic workflow. V and R represent Verifier and Reflector, respectively.

Components				SumMe		TVSum	
Baseline	V2I	V	R	τ	ρ	τ	ρ
✓	✗	✗	✗	0.230	0.257	0.178	0.232
✓	✓	✗	✗	0.265	0.296	0.215	0.278
✓	✓	✓	✓	**0.274**	**0.308**	**0.220**	**0.290**

Table 3. Ablation study of pooling strategies and alignment methods in the V2I Alignment module on SumMe. The study focuses on the V2I module without using the Verifier and the Reflector.

Method	τ	ρ
Without Alignment		
Only E_{img}	0.228	0.254
Only E_{vid}	0.220	0.245
Concat E_{img} and E_{vid}	0.230	0.257
With Alignment		
I2V alignment	0.238	0.265
V2I alignment	0.265	0.296
With V2I Alignment		
Mean pooling	0.243	0.270
Temporal Attention Pooling	0.265	0.296

riches inter- and intra-frame representations. The Verifier serves as a bridge between the Summarizer and the Reflector and is meaningful only when used in conjunction with the Reflector. When jointly applied, these two components reassess initial scores, provide a global perspective of video content, recover missing keyframes, and refine predictions. This results in further performance gains, with τ increasing from 0.265 to 0.274 on SumMe and from 0.215 to 0.220 on TVSum.

V2I Alignment Establishment. We conduct extensive experiments to analyze the pooling strategy and mapping design in the V2I Alignment module, as shown in Table 3. We ultimately adopt temporal attention pooling for video-level pooling and use AdaptMLP to align video- and image-level embeddings, producing single-frame features enriched with inter-frame context. Experiments using only the image-level or only the video-level encoder show that either alone falls short of the aligned fusion strategy. We also compare pooling and mapping variants: temporal attention pooling outperforms mean pooling by assigning adaptive importance weights to frames or regions, yielding richer semantics and stronger representations. Additionally, V2I alignment proves more effective than I2V.

Reflector Establishment. For train-free Reflector, we explore two distinct construction methods: one in which Qwen-VL directly generates importance scores, and the other described in Section 3.4. From the results in Table 4, we observe that under the same train-free setting, the scores produced by Reflector align more closely with the human annotations, whereas Qwen-VL exhibits less satisfactory performance. Moreover, when using CLIP-generated scores, the prediction accuracy under the train-free condition already surpasses several early fully supervised methods [12, 15, 48]. The inferior performance of Qwen-VL compared to CLIP stems from their scoring mechanisms. As an MLLM, Qwen-VL generates human-like responses based on queries and multi-step reasoning, resulting in less precise scores with limited inter-frame variation, making them unsuitable for metrics like τ and ρ. In contrast, Reflector uses CLIP to produce stable, fine-grained scores with clear frame distinctions, making it more suitable for video summarization—consistent with findings in CLIP-It [29]. These findings demonstrate that in the unsupervised setting, our Reflector—combining Qwen-VL for video understanding with CLIP for score generation—achieves practical and competitive performance. This strategy requires no model training and can be readily adapted to a wide range of task scenarios.

Table 4. Ablation study of model selection within the Reflector module. The study focuses on train-free Reflector without using Summarizer and Verifier.

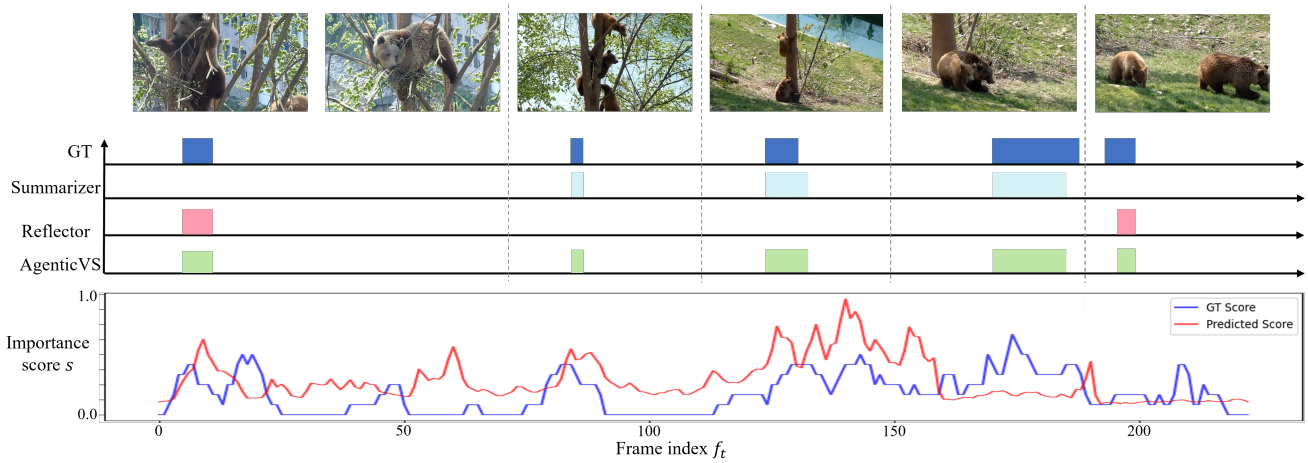
Method	Training Type	τ	ρ
DMASum [48]	Supervised	0.063	0.089
iPTNet [15]	Supervised	0.101	0.119
A2Summ [12]	Supervised	0.108	0.129
Qwen2.5-VL	Train-free	0.073	0.081
Reflector (Ours)	Train-free	0.116	0.128

4.4. Qualitative Results

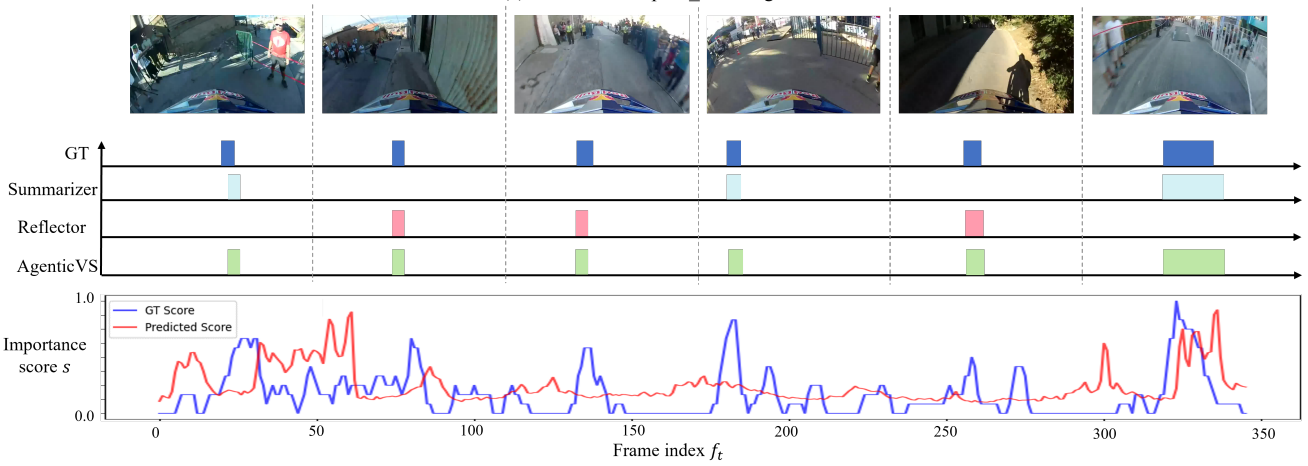
Fig. 5 presents qualitative comparisons on two example videos from the SumMe dataset. Each row illustrates the importance score and selected key segments over time by different methods, where the top row (GT) represents the human-annotated ground truth summaries aggregated from multiple users.

As shown in Fig. 5, our model produces a summary that closely aligns with the ground truth. The initial scores produced by Summerizer effectively capture key transitions and salient moments such as behavioral changes and interview scenes. Furthermore, the subsequent processing by the Verifier and Reflector effectively compensates for the missing segments from the Summarizer, resulting in improved summarization quality.

These qualitative results affirm the strength of our agentic workflow in identifying contextually significant segments with MLLM deep understanding and reasoning. For



(a) Video 3. Bearpark_climbing



(b) Video 22. Valparaiso_Downhill

Figure 5. **Qualitative results.** Videos are from the SumMe dataseet. The x-axis and y-axis represent frame index f_t and importance score s, respectively. The blue bars indicate the user summary from the ground truth annotations, while the green bars denotes the predicted key frames from our AgenticVS. Mint and pink bars represent that Reflector can effectively compensate for the missing segments from the Summarizer.

the adjustment of the initial scores, we adopt a unified instructional prompt for MLLM-based analysis rather than relying on human-provided explanations as prompts. This design not only reduces manual intervention and facilitates adaptation to various scenarios but also mitigates the subjective bias introduced by manually crafted prompts, thereby enhancing the interpretability of AgenticVS.

5. Conclusion

In this paper, we propose **AgenticVS**, a novel agentic framework for video summarization that leverages the reasoning, self-reflective, and tool utilization capabilities of MLLMs. Unlike traditional extractive or one-shot methods, AgenticVS reformulates video summarization as a closed-loop workflow of three agents—the *Summarizer*, *Verifier*, and *Reflector*. The Summarizer produces initial frame-level scores using aligned visual embeddings, and the Reflector performs multi-round reasoning to refine them based on global video context, recovering missed keyframes. V2I Alignment ensures effective fusion of video- and image-level features. Experiments on SumMe [10] and TVSum [40] show that AgenticVS surpasses state-of-the-art extractive and LLM-based baselines in ranking-based metrics (τ, ρ), confirming its ability to deliver more accurate predictions with fewer missed keyframes and without hand-crafted prompts. The framework is flexible and extendable to broader video understanding tasks (e.g., VQA, video reasoning, video tracking) and more complex settings such as long videos, multi-view data, and additional modalities. In summary, AgenticVS advances toward intelligent, adaptive, and human-like video summarization by integrating summarization, verification, and self-reflection to iteratively refine predictions and achieve robust global video understanding.

References

[1] Toqa Alaa, Ahmad Mongy, Assem Bakr, Mariam Diab, and Walid Gomaa. Video summarization techniques: A comprehensive review. *arXiv preprint arXiv:2410.04449*, 2024. 2

[2] 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. IEEE, 2021. 6

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8332–8341, 2024. 6

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 3

[6] Mohamed Elfeki, Liqiang Wang, and Ali Borji. Multi-stream dynamic video summarization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 339–349, 2022. 1

[7] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian conference on computer vision*, pages 39–54. Springer, 2018. 6

[8] Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. Supervised video summarization via multiple feature sets with parallel attention. *arXiv preprint arXiv:2104.11530*, 2021. 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 *European conference on computer vision*, pages 505–520. Springer, 2014. 2, 6, 8

[11] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022. 2

[12] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878, 2023. 6, 7

[13] Tanveer Hussain, Khan Muhammad, Amin Ullah, Zehong Cao, Sung Wook Baik, and Victor Hugo C De Albuquerque. Cloud-assisted multiview video summarization using cnn and bidirectional lstm. *IEEE Transactions on Industrial Informatics*, 16(1):77–86, 2019. 2

[14] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717, 2019. 2

[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. 2, 6, 7

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

[19] Min Jung Lee, Dayoung Gong, and Minsu Cho. Video summarization with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18981–18991, 2025. 2, 4, 6

[20] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5584–5593, 2023. 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[22] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing*, 3:0063, 2024. 2

[23] 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. 3

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 3

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 3

[26] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[27] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a

dynamic api. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19446–19455, 2025. 3, 4

[28] Ghulam Mujtaba, Adeel Malik, and Eun-Seok Ryu. Ltc-sum: Lightweight client-driven personalized video summarization framework using 2d cnn. *IEEE access*, 10:103041–103055, 2022. 1

[29] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021. 2, 4, 6, 7

[30] 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

[31] Anil Singh Parihar, Joyeeta Pal, and Ishita Sharma. Multi-view video summarization using video partitioning and clustering. *Journal of Visual Communication and Image Representation*, 74:102991, 2021. 1

[32] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40, 2024. 2

[33] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7902–7911, 2019. 1

[34] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023. 2

[35] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8523–8533, 2025. 3

[36] Sumin Kim Minjun Kim Wonsik Shin and Yebonn Han. Learning semantic representations for video summarization via mllms. 2

[37] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023. 2

[38] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. 2

[39] 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, 4, 6

[40] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 2, 6, 8

[41] Charles Spearman. The proof and measurement of association between two things. 1961. 6

[42] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 2

[43] 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. 4

[44] Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, 2023. 2

[45] Hacene Terbouche, Maryan Morel, Mariano Rodriguez, and Alice Othmani. Multi-annotation attention model for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2023. 6

[46] 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. 2

[47] Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. Stacked memory network for video summarization. In *Proceedings of the 27th ACM international conference on multimedia*, pages 836–844, 2019. 2

[48] 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, 7

[49] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 4

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[51] Xiyao Wang, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl. *arXiv preprint arXiv:2310.07220*, 2023. 2

[52] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 2, 3, 4

[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[54] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[55] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9143–9150, 2019. 2

[56] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2, 3

[57] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016. 2

[58] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7405–7414, 2018. 2

[59] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Tth-rnn: Tensor-train hierarchical recurrent neural network for video summarization. *IEEE Transactions on Industrial Electronics*, 68(4): 3629–3637, 2020. 2

[60] Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2793–2801, 2021. 2

[61] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025. 3

[62] 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, 6