Video Summarization with Large Language Models
– Supplementary Material –

Min Jung Lee[1,2] Dayoung Gong[1] Minsu Cho[1]

[1]Pohang University of Science and Technology (POSTECH) [2]GenGenAI

In this appendix, we provide additional experiments in Section A, additional qualitative results in Section B, and full prompts for LLM π in Section C,.

A. Additional Experiments

Model choice of (M-)LLM, ϕ and π. While LLaVA [3] generates captions based on single-frame inputs, Video-LLaVA [2] processes multiple frames simultaneously, enabling video-level understanding. Given the flexibility of our framework in M-LLM selection, we incorporate both models to examine how the design of the captioning model influences downstream performance.

We exploit Video-LLaVA as both the text generator ϕ and the local importance scorer π in the last row of Table A1. In this setup, we first obtain captions from Video-LLaVA and then feed both the generated captions and the corresponding video frames back into Video-LLaVA for local window-based scoring, where the model supports a maximum of 8 frames per window. Table A1 presents a comparison between two configurations, showing similar performance. This suggests that once textual descriptions are well aligned with the video content, additional visual input at the scoring stage offers limited benefit. Moreover, modeling local temporal context via π over sequential captions appears sufficient to compensate for the limitations of frame-level captioning.

Effects of the window size w. To determine the optimal window size w, we examine the impact of varying w on summarization performance. An appropriate window size should provide enough contextual information without introducing excess detail, balancing local context for effective summarization. As shown in Table A2, the LLMVS model achieves its best performance at $w = 7$, suggesting that this window size offers the ideal balance between capturing local context and maintaining coherence for summarization.

Effects of the number of self-attention blocks. To evaluate the impact of self-attention depth in the global context aggregator, we examine configurations with 2, 3, and 4 self-attention blocks. As shown in Table A3, using three self-attention blocks improves performance compared to two self-attention block, likely due to enhanced contextual integration. However, performance decreases when increasing to four self-attention blocks, possibly due to added complexity. Thus, three self-attention blocks provides the best balance for video summarization.

ϕ	π	ψ	τ	ρ
LLaVA	Llama	SAB*	**0.253**	**0.282**
Video-LLaVA	Video-LLaVA	SAB*	0.252	0.281

Table A1. **Model choice of (M-)LLM, ϕ and π.** ϕ: text generator, π: local importance scorer, ψ: global context aggregator, *:fine-tuned, SAB: self-attention blocks.

Method	w	τ	ρ
LLMVS (ours)	5	0.236	0.263
LLMVS (ours)	7	**0.253**	**0.282**
LLMVS (ours)	9	0.245	0.274

Table A2. **Effects of Window Size w.** Evaluation performed on the SumMe dataset [1] with 3 self-attention blocks and 2 multi-head attention heads.

N	τ	ρ
2	0.243	0.271
3	**0.253**	**0.282**
4	0.244	0.272

Table A3. **Effects of the Number of Self-Attention Blocks (N).** Evaluation performed on the SumMe dataset [1] with 2 multi-head attention heads.

B. Additional Qualitative Results

Figure A1 presents additional qualitative results, comparing the summaries generated by LLMVS with the ground truth on (a) the SumMe[1] and (b) the TVSum [4] datasets. In SumMe, the x-axis denotes the time step t, while the y-axis

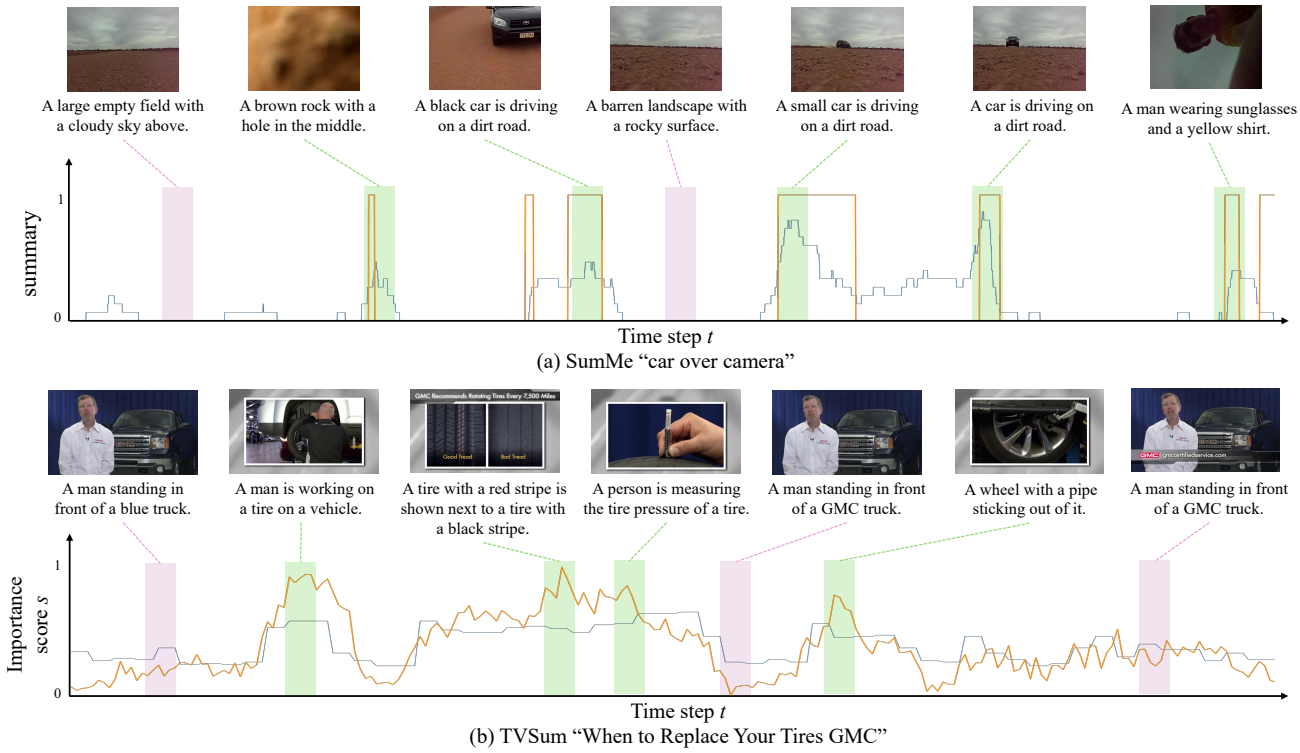

Figure A1. **Additional qualitative results.** Green regions highlight segments where importance scores are high, whereas pink regions indicate segments where importance scores are low. (a) Results from SumMe [1]. The x-axis and y-axis represent time step t and binarized summary, respectively. The blue line represents the average of binary user summaries in the ground truth, and the orange line is the predicted summary of our model, which is processed using the KTS and 0/1 knapsack algorithm on predicted frame score. (b) Results from TVSum [4]. The x-axis and y-axis represent time step t and importance score s, respectively. The blue line is the average of user scores ranging in $[0, 1]$, and the orange line is the normalized predicted importance score.

represents the binarized summary. The blue line shows the averaged binary summaries from multiple users, and the orange line represents our predicted summary, obtained by applying the KTS and 0/1 knapsack algorithm to the predicted frame scores. As shown in Figure A1(a), the predicted summaries closely align with the peaks in the ground truth. For example, LLMVS successfully identifies key transitions, such as when the camera falls to the ground or when a car drives over a ground-level camera. Figure A1(b) illustrates the results on TVSum. Here, the x-axis again represents the time step t, and the y-axis indicates importance scores. The blue line shows the average of user-provided scores ranging from 0 to 1, while the orange line represents normalized predicted scores of LLMVS. Both human annotations and our predictions exhibit similar trends—higher scores are assigned to action-oriented segments (e.g., working on or touching a tire), while lower scores are given to static or less informative scenes. By leveraging the local window of captions, LLMVS effectively captures the narrative context of shots and identifies critical contents, aligning closely with human perception of scene importance. These results further demonstrate the robustness and generalization capability of LLMVS across diverse user annotations and video summarization benchmarks.

C. Full Prompts for LLM π

In this section, we provide full prompts given to LLM π for in-context learning in Table A4. As illustrated in Figure [3], our prompts consist of three parts: instruction $\boldsymbol{i}$, examples $\boldsymbol{e}$, and queries $\boldsymbol{q}$. The instructions guide LLM regarding the video summarization task, followed by three examples. Each example includes a question-answer pair, where the question requests score evaluations with frame captions, and the answers, ranging from one to ten, are derived from the dataset. The queries are direct questions given to LLM, requiring the desired actual answers $\boldsymbol{a}$.

Instruction	You are an intelligent chatbot designed to critically assess the importance of a central frame within a specific context. Given a set of consecutive frame descriptions from a video with narrative changes, your task is to assign an importance score to the central frame based on its narrative contribution. Evaluate the frame using the following criteria: —— ##INSTRUCTIONS: 1. **Narrative Significance**: Assign a high score if the frame captures pivotal plot developments, character milestones, or key conflicts/resolutions. This measures the frame's impact on the overall story. 2. **Uniqueness and Novelty**: Score highly if the frame introduces new elements or showcases significant alterations in the story or setting. This reflects the frame's contribution to refreshing the narrative. 3. **Action and Dynamics**: Give a high score if the frame depicts crucial actions, events, or is characterized by high energy or movement. This assesses the intensity and momentum conveyed by the frame. ##NOTE: Keep in mind that the descriptions provided may not fully capture the essence of the corresponding image. Therefore, it's crucial to consider the overall context when determining the importance of the central frame. Assess its significance not only based on the explicit details given but also in the context of the narrative progression and thematic development.
Example 1	Please evaluate the importance score of the central frame #7 in following 13 frames. Be stingy with scores. —— #1: A man is standing on a ramp next to a car. #2: A man is standing on a flatbed truck. #3: A man is standing on a ramp next to a car. #4: A man is standing on a ramp with a blue car on it. #5: A man is standing in front of a crowd of people. #6: A blue shirt with a white collar. #7: A close up of a piece of cloth. #8: A purple wall with a blue stripe. #9: A person's arm with a white shirt on. #10: A person is wearing a purple shirt. #11: A man is holding a rock in his hand. #12: A man is sitting on a chair and holding a car hood. #13: A man is holding a car door open while another man is holding a piece of paper. —— Provide your score where the score is an integer value between 0 and 10, with 10 indicating the highest important frame in a context. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANARION. Only provide the Python dictionary string.
Answer	score: 1
Example 2	Please evaluate the importance score of the central frame #4 in following 7 frames. Be stingy with scores. —— #1: A group of people are standing on a roadway near a railroad crossing. #2: A group of people are standing on a street corner. #3: A group of people are standing on a ramp in the middle of a street. #4: A group of people are standing on a road that is blocked off. #5: A group of people are standing around a car that is stuck in a puddle. #6: A group of people are standing around a car that is on its side. #7: A group of people are standing around a car that is on its side. —— Provide your score where the score is an integer value between 0 and 10, with 10 indicating the highest important frame in a context. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANARION. Only provide the Python dictionary string.

Answer	score: 5
Example 3	Please evaluate the importance score of the central frame #6 in following 11 frames. Be stingy with scores. —— #1: A group of people are standing in the middle of a street. #2: A group of people are standing in front of a traffic light. #3: A group of people are standing on a roadway near a railroad crossing. #4: A man is standing on a railroad crossing. #5: A man is standing on a railroad crossing. #6: A car is driving on a street with a red light. #7: A car is driving on a road with a man standing next to a railroad crossing. #8: A man is pushing a large metal object in front of a train. #9: A man is sitting on a couch in the middle of a street. #10: A car is driving through a red light. #11: A man is standing on a railroad crossing. —— Provide your score where the score is an integer value between 0 and 10, with 10 indicating the highest important frame in a context. # DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANARION. Only provide the Python dictionary string.
Answer	score: 9

Table A4. Full prompts for LLM π

References

[1] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer, 2014. 1, 2

[2] 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. 1

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[4] 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. 1, 2