Summarizing Robot Records via LLM-Generated Key Frame Descriptions

***Abstract*—Recent advancements in robotics have positively impacted the adoption of autonomous systems supporting routine tasks. Autonomous robots are now present in vehicles, medical procedures, industrial plants, and even in our homes. However, large-scale adoption requires the storage and analysis of massive amounts of data to understand robot behavior and prevent potential failures in sensitive tasks. As a result, analytical tasks have become increasingly complex, often exceeding human capabilities due to the time required for manual analysis. In low-resource environments, large logs are typically reviewed only when an incident is reported. In the next generation of robotics, enhanced problem-solving reasoning will be essential to map and understand occurrences along a robot's path. To address these challenges, this article proposes a key frame selection method integrated with large language models to summarize video logs from autonomous robots. Our method selects the most relevant frames to generate meaningful summaries, producing a condensed archive that highlights the key events captured in the original footage. As a result, each video is reduced to a summary composed of descriptions based on approximately 0.5% of the total frames. These summaries are presented as paragraph-sized documents, effectively representing the essential content of the original videos.**

***Index Terms*—Video Summarization, Key Frame Selection, Autonomous Robots, Large Language Models**

I. Introduction

Autonomous robots are unsupervised systems capable of perceiving their environment, making decisions, and executing tasks without direct human intervention [1]. Instead, they rely on embedded sensors, algorithms, and adaptive control mechanisms to navigate dynamic and often unpredictable settings [2]. In this scenario, to understand how the system behaves or what happened while the robot performs a routine, we need to analyze and watch the logs produced, inspecting all the data to identify potential issues.

The data generated by these models in open-world testing scenarios is highly valuable for understanding events that occur in real-world conditions [3]. However, the video recordings produced during these tests often span several hours, making manual review by human operators both time-consuming and impractical [4]. For instance, Google has reported, via its Waymo Driver platform, a record of over 40 million miles of real-world driving experience [5]. In such cases, the volume of recorded data far exceeds what can be feasibly reviewed by humans. Therefore, intelligent filtering techniques are required to interpret, analyze, and optimize interactions with the environment.

Concerning the contrast between the vast amount of data and human interpretability, intelligent filtering is required to extract valuable information and analyze the recorded content efficiently [6]. In this regard, the method enables the structured organization of data, the extraction of key highlights for individual analysis, and the generation of focused analytical outputs.

Despite the opportunity to review robot-generated records, the latest state-of-the-art models still require increasingly large training datasets to enhance their performance. Consequently, data-driven applications call for in-depth analysis of the key factors influencing model effectiveness, particularly in real-world scenarios.

In light of these challenges, this paper proposes a method for organizing video data by selecting and describing key frames. The approach involves summarizing the video content through the identification of representative frames that capture its diversity and constitute the highlights. To enhance the analytical value, descriptive captions are generated for the selected frames, emphasizing the observed diversity and enriching the semantic interpretation of the results.

The remainder of this paper is structured as follows. Section 2 reviews related work concerning video summarization and frame selection in autonomous systems. Section 3 details the proposed method for selecting and describing key frames. Section 4 presents the dataset used for evaluation and its characteristics. Section 5 discusses the experimental results and the analytical value of the generated summaries. Finally, Section 6 concludes the paper and outlines directions for future work.

II. Related Works

Today, the semantic capabilities of language models have led to their application in a wide range of tasks. One important direction is the specialization of these models into Video Language Models (VLMs) [7], which aim to enhance video analysis by identifying the most relevant events within a video. Some strategies, such as those proposed by [8], [9], use these models to extract meaningful content by analyzing multiple modalities, including the environment, music, and voice, to infer contextual information. However, in the case of robotic recordings, audio is not always directly related to the tasks being performed. Instead, the visual stream is typically more informative, capturing key objects and movements. In this context, approaches that apply video analysis to identify relevant segments tend to focus on repetitive behaviors and the common patterns found in robotic routines [10].

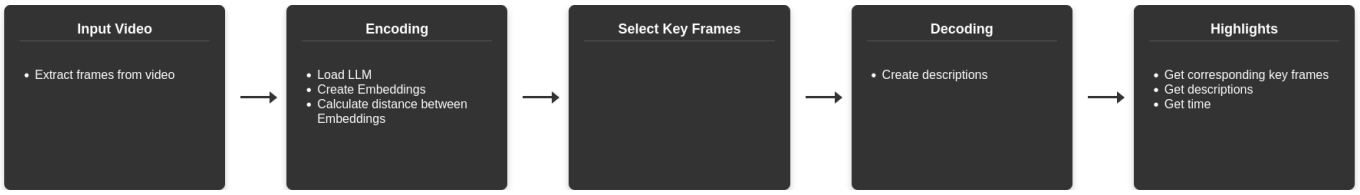

Fig. 1: Illustration of the five steps applied to generate the video summaries.

Other strategies aim to summarize the entire video by segmenting and analyzing the overall data contextually [11]. However, the computational cost, the need for human validation, and the time required for processing make such approaches overly ambitious for straightforward integration into robotic systems. The volume of data is typically too large to allow for exhaustive sequence-to-sequence verification without some form of filtering or prior information extraction [12].

To address these limitations, recent approaches have focused on the use of key frame selection combined with instruction-tuning techniques to reduce data complexity while preserving essential information [13]. By selecting representative frames that capture the core visual elements of each segment, these methods enable efficient processing without relying on deep contextual interpretation or human-in-the-loop supervision. This is particularly advantageous in large-scale robotic scenarios, such as autonomous driving, where hours of continuous video are generated and must be analyzed efficiently [12]. Key frame-based strategies make it possible to bypass redundant or low-relevance content, avoiding the need for time-consuming humanistic analysis, which often demands a rich understanding of context, intention, and semantics [14]. This, in turn, enables scalable summarization, indexing, and monitoring under constrained computational conditions.

III. Method

The video summarization task involves multiple steps to capture the environmental context and describe the behavior of moving objects [15]. Our proposed approach focuses on detecting key frames that effectively represent the diversity of the video, enabling a deeper understanding of the robot's behavior through its interactions with the environment. Figure 1 illustrates the five steps applied to generate the video summaries.

Figure 1 presents the five main steps of our proposed approach. The first step consists of preprocessing the input videos by converting them into sequences of frames. Once the frames are extracted, we encode the visual information using embedding models, which convert each frame into a numerical representation. These embeddings are then used to identify key frames that capture the robot's movement and the diversity of the environment. The selected key frames are subsequently interpreted using a Large Language Model (LLM) to generate descriptive captions. As a result, we produce a set of key frame–caption pairs that effectively summarize the video content.

To perform semantic interpretation and visual recognition, we employ LLaVA (Large Language and Vision Assistant) [16], a multimodal model that combines visual understanding with language generation. LLaVA is built upon a vision-language architecture that integrates a visual encoder with a Large Language Model (LLM). Given an image and an optional textual prompt, the model generates detailed and contextually grounded descriptions based on the visual input.

At the preprocessing stage, the input videos from the dataset are converted into frame sequences. To capture the essential visual elements that characterize the robot's operation, we reduce the frame rate to 10 frames per second. This adjustment allows us to preserve relevant visual information and object motion while minimizing redundancy, ensuring that each extracted frame contributes meaningfully to the overall scene understanding.

Afterwards, to enrich the processing with visual semantics, each image is transformed into an embedding vector. We use the *LLaVA 1.5 7B* model to extract these vectors from the hidden states of its internal representation. Each frame is thus represented by a 4096-dimensional feature vector, capturing its visual and contextual information. Since the video logs range from a few minutes to several hours in duration, this embedding process plays a key role in characterizing the visual context and dynamics across the entire frame sequence [17].

To determine the most relevant frames, we identify divergences between local and global visual features. Videos recorded by autonomous robots tend to exhibit repetitive scenarios and behaviors, resulting in consistent and predictable visual patterns [18]. Therefore, to summarize these videos effectively, we focus on detecting shifts in visual context and motion between frames.

Key frames are defined as those that exhibit significant divergence within their local temporal sequences. Specifically, any frame that presents at least a 15% divergence compared to others within its 8-second time window is considered a potential key frame. These sequences are designed to capture short-term movements and contextual transitions.

To quantify this divergence, we compute the cosine similarity between frame embeddings. Frames that deviate significantly from their local context are selected as representative, indicating a disruption in the visual flow. This method allows us to detect unexpected events and relevant changes by evaluating the coherence between local scenarios and global patterns [17].

Once the key frame sequences are generated and the high-

lights identified, the next step is to summarize and describe them. To generate these descriptions, we input the selected key frames into the LLaVA model, prompting it to recognize the robot's context and its surrounding environment. Prompt calibration is performed to guide the model toward more accurate and relevant descriptions, explicitly specifying the robot's task and the operational setting. In our case, a handcrafted and task-specific prompt was designed and fine-tuned for the self-driving car scenario, ensuring that the model focuses on elements relevant to autonomous navigation, such as road layout, traffic participants, and vehicle behavior. The prompt used is as follows:

Customized Prompt for Scene Description

You are an autonomous vehicle perception system. Analyze the scene and respond using the exact structure below:
1. **Ego-vehicle state:** [Lane position (e.g., "center of right lane", "between lanes", etc.), and most likely next action if inferable (e.g., "continue straight", "prepare to turn left"). Do not describe speed or current motion.]
2. **Traffic objects:** [List relevant vehicles, pedestrians, cyclists, or obstacles in or near the ego-vehicle path. Describe their location relative to the ego-vehicle and any observed behavior.]
3. **Signals/signs:** [Report any visible traffic lights or road signs that may affect the ego-vehicle.]
4. **Potential hazards:** [Mention any immediate risks, blocked lanes, conflicting objects, or unusual traffic conditions that could impact driving decisions.]

Formatting rules:

- Be strictly neutral and technical.
- Maximum 2 sentences per item.
- Focus only on actionable, traffic-relevant information.
- Ignore all non-traffic elements (buildings, scenery, weather, etc.).
- Do not mention uncertainty or limitations.

Describe the current traffic scene from the autonomous vehicle's perspective.

This tailored prompting approach improves consistency and contextual alignment across different video segments, enabling the model to produce concise and meaningful summaries of the robot's visual experience.

The results obtained from this process serve as input for the subsequent analytical stage, where the video logs are indexed based on the selected key frames and their corresponding descriptions. This index acts as a structured summary of the video content, allowing for efficient retrieval, navigation, and interpretation of the robot's behavior over time. By associating representative images with concise textual descriptions, the index highlights the most relevant events and contextual shifts within the video. This synthesis not only reduces the volume of data to be analyzed but also enhances the semantic understanding of long-duration recordings, enabling human operators or automated systems to quickly grasp the core activities and anomalies present in the logs.

IV. Datasets

To validate our approach, we utilize open-world video logs collected from autonomous driving platforms. This type of data was selected due to its richness in continuous and diverse environmental interactions, which are ideal for evaluating visual understanding systems. The primary objective is to identify key frames that capture meaningful changes in the environment, while filtering out repetitive or redundant scenes. Autonomous driving videos are particularly suitable for this task because they naturally involve continuous movement through dynamic and varied scenarios, such as urban, suburban, and highway environments. These characteristics make them an excellent proxy for real-world robotic perception challenges. Moreover, the use of autonomous driving data allows us to assess the proposed system's applicability in broader contexts, including surveillance, robotic task monitoring, and human task assessment. By leveraging these videos, we ensure that the evaluation of our system reflects realistic operational conditions that are critical for its intended applications.

The selected data is the *ROad event Awareness in Autonomous Driving (ROAD Challenge)* dataset [19], which was built upon the *Oxford RobotCar Dataset* [1]. The videos are relatively short (approximately 8 minutes each), but still provide rich and diverse driving scenarios from the perspective of the autonomous vehicle. Each video is annotated with location information, the actions being performed, and the moving agents present in the scene. Despite their limited duration, these videos are highly relevant for our purposes, as they contain sufficient event diversity to enable the detection of meaningful transitions and support the identification of key frames. This makes them particularly suitable for evaluating video summarization techniques aimed at capturing the most important moments in dynamic, real-world environments. Table I presents the duration and number of extracted frames for each video in the dataset.

As shown in Table I, the dataset comprises a total of 2 hours, 21 minutes and 32 seconds of video records, amounting to 101,925 frames. Each of the 18 videos captures distinct traffic scenarios, including variations in traffic density, environmental conditions, and moving agents. This diversity is essential for

[1] *Oxford RobotCar Dataset* - https://robotcar-dataset.robots.ox.ac.uk/

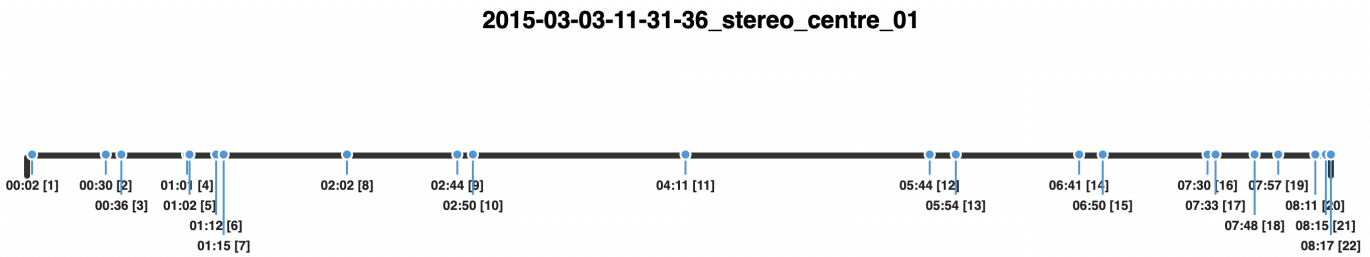

Fig. 2: Timeline showing the selected frames from the `2015-03-03-11-31-36_stereo_centre_01` recording

TABLE I: The 18 videos from the ROAD Challenge dataset used in the experiments

Sample	Video	
	Time (Minutes)	Frames
2014-07-14-14-49-50_stereo_centre_01	08:20	6002
2015-02-03-19-43-11_stereo_centre_04	08:20	6001
2014-11-21-16-07-03_stereo_centre_01	08:20	6002
2015-02-13-09-16-26_stereo_centre_02	08:20	6001
2015-02-13-09-16-26_stereo_centre_02	08:20	6001
2014-12-09-13-21-02_stereo_centre_01	08:20	6002
2014-08-08-13-15-11_stereo_centre_01	08:20	6002
2014-06-25-16-45-34_stereo_centre_02	06:35	4735
2014-08-11-10-59-18_stereo_centre_02	01:37	1169
2015-02-03-08-45-10_stereo_centre_02	08:20	6001
2014-06-26-09-53-12_stereo_centre_02	08:20	6001
2014-11-18-13-20-12_stereo_centre_05	08:20	6001
2015-03-03-11-31-36_stereo_centre_01	08:20	6002
2015-02-06-13-57-16_stereo_centre_02	08:20	6001
2014-11-25-09-18-32_stereo_centre_04	08:20	6001
2014-11-14-16-34-33_stereo_centre_06	08:20	6001
2014-07-14-15-42-55_stereo_centre_03	08:20	6001
2015-02-24-12-32-19_stereo_centre_04	08:20	6001
2015-02-13-09-16-26_stereo_centre_05	08:20	6001
ROAD Challenge	**141:32**	**101,925**

TABLE II: Results of the key frame selection process for each video in the ROAD dataset.

Sample	Frames		
	All	Selected	Ratio (%)
2014-07-14-14-49-50_stereo_centre_01	6002	23	0.38
2015-02-03-19-43-11_stereo_centre_04	6001	17	0.28
2014-11-21-16-07-03_stereo_centre_01	6002	23	0.38
2015-02-13-09-16-26_stereo_centre_02	6001	22	0.37
2015-02-13-09-16-26_stereo_centre_02	6001	22	0.37
2014-12-09-13-21-02_stereo_centre_01	6002	22	0.37
2014-08-08-13-15-11_stereo_centre_01	6002	26	0.43
2014-06-25-16-45-34_stereo_centre_02	4735	28	0.59
2014-08-11-10-59-18_stereo_centre_02	1169	5	0.43
2015-02-03-08-45-10_stereo_centre_02	6001	31	0.52
2014-06-26-09-53-12_stereo_centre_02	6001	26	0.43
2014-11-18-13-20-12_stereo_centre_05	6001	31	0.52
2015-03-03-11-31-36_stereo_centre_01	6002	22	0.37
2015-02-06-13-57-16_stereo_centre_02	6001	40	0.67
2014-11-25-09-18-32_stereo_centre_04	6001	26	0.43
2014-11-14-16-34-33_stereo_centre_06	6001	14	0.23
2014-07-14-15-42-55_stereo_centre_03	6001	35	0.58
2015-02-24-12-32-19_stereo_centre_04	6001	38	0.63
2015-02-13-09-16-26_stereo_centre_05	6001	28	0.47

evaluating the generalization capability of the proposed summarization approach. By covering a broad spectrum of real-world driving contexts, such as urban intersections, highway merges, pedestrian crossings, and low-visibility conditions. The dataset ensures that the system is exposed to representative and challenging situations. This variety enables a comprehensive assessment of how well the model can detect and summarize meaningful events across heterogeneous scenes.

V. Results

To evaluate the effectiveness of our key frame selection method, we applied the system to the raw videos from the ROAD dataset. The objective was to measure how well the method reduces the total number of frames while retaining the essential visual content required for summarization and scene understanding. This assessment focuses on the compression ratio achieved, the distribution of selected frames, and the variation observed across different driving scenarios. The results of this evaluation are summarized in Table II.

Table II summarizes the results of the key frame selection process applied to each video sample from the ROAD dataset. Across the 18 video samples, a total of 101,925 frames were analyzed, from which 510 key frames were selected. The proportion of selected frames ranges from 0.23% to 0.67%, indicating a high level of reduction while still aiming to preserve the most relevant events. This selection rate reflects the system's ability to filter redundant content and focus on meaningful transitions or dynamic changes in the scene. The video with the highest key frame ratio (0.67%) is 2015-02-06-13-57-16_stereo_centre_02, suggesting a greater density of relevant activity or scene changes. Conversely, the lowest selection ratio (0.23%) was observed in 2014-11-14-16-34-33_stereo_centre_06, which may indicate a more static or repetitive scenario. These results validate the summarization system's sensitivity to visual complexity and its effectiveness in condensing video content with minimal loss of informative frames. Figure 2 presents the temporal distribution of events, while Figure 3 shows the selected frames according to the timeline labels.

From the complete set of selected frames, we further extracted a subset of 22 identified as the most contextually relevant. As shown in Figure 3, these frames are distributed along the timeline in segments corresponding to distinct events, interspersed with noticeable gaps indicating periods of low activity or visual redundancy. The selected frames reveal key nuances in traffic behavior, capturing unique movements and

Fig. 3: Visual references of the selected key frames from the `2015-03-03-11-31-36_stereo_centre_01` recording, each identified by ID and aligned with the timeline and descriptions.

the presence of diverse actors within the environment. This targeted selection aims to represent the complexity of the scenes while minimizing redundancy, providing a foundation for a concise and informative summary. Detailed information for each frame—including its identifier, timestamp, and a textual description generated from the ego-vehicle's perspective—is presented in Table III, highlighting the most significant elements observed in the recording.

Looking at Table III, the descriptions emphasize the main aspects of each key frame. Within self-driving car scenarios, these aspects include movement patterns, potential traffic issues, traffic signal statuses, and the presence of moving objects. The descriptions attentively cover this diversity by highlighting hazardous conditions and behavioral adaptations aimed at resolving traffic flow challenges. Furthermore, to demonstrate the summarizer's performance, we present representative scenarios, including the one shown in Figure 4 (key frame $\{11\}$).

Fig. 4: An example of LLM description with traffic details and robot behavior (frame $\{11\}$).

{11} *1.* ***Ego-vehicle state:*** *The car is driving down the road, in the middle of the right lane. 2.* ***Traffic objects:*** *There are multiple cars and a bus on the road, with some cars in front of the car and others behind it. 3.* ***Signals/signs:*** *There are traffic lights visible in the scene, but their positions and states are not provided. 4.* ***Potential hazards:*** *The car is driving near a bus, which could potentially obstruct its view of the road.*

TABLE III: Descriptions of the selected key frames from the `2015-03-03-11-31-36_stereo_centre_01` recording. Each entry corresponds to a marker in the timeline and summarizes the visual content of that frame.

#	Time	Ego-vehicle behavior	Traffic objects	Signals/signs	Potential hazards
1	00:02	Driving down the street, approaching a tree and a brick wall.	Truck parked on side of road; chair near wall.	None visible.	Approaching tree and brick wall, possible hazards.
2	00:30	Parked in front of a house, possibly waiting for someone.	No visible traffic objects nearby.	None visible.	No immediate risks or hazards.
3	00:36	Parked in front of building, possibly waiting or preparing to leave.	No visible traffic objects nearby.	None visible.	Potential hazard if someone enters/exits vehicle.
4	01:01	Parked, waiting for pedestrians to cross.	Two pedestrians crossing street; truck parked nearby.	None visible.	Pedestrians crossing could delay vehicle.
5	01:02	Driving down street, approaching a building.	Man and woman walking on sidewalk; person near building.	None visible.	Pedestrians may step onto road unexpectedly.
6	01:12	Driving down street, approaching building.	No visible traffic objects.	None visible.	Building may obstruct view of oncoming traffic.
7	01:15	Driving down road, approaching intersection.	No visible traffic objects nearby.	None visible.	Needs caution at intersection.
8	02:02	Driving down street, approaching intersection.	Multiple cars, truck, pedestrian visible.	Traffic light present.	Pedestrian crossing at intersection.
9	02:44	Driving down street, approaching intersection.	No visible traffic objects nearby.	Traffic light visible.	Must check for traffic before proceeding.
10	02:50	Driving down road, approaching intersection.	No visible traffic objects.	Traffic light visible.	No immediate hazards.
11	04:11	Driving down road, in middle of right lane.	Multiple cars and bus nearby.	Traffic lights visible, state unknown.	Bus may obstruct view of road.
12	05:44	Driving, approaching crosswalk.	Couple crossing; two cars and a truck nearby.	Traffic light visible.	Must slow/stop for pedestrians.
13	05:54	Driving, approaching intersection.	Several cars and a truck nearby.	Traffic light visible.	Caution with other vehicles and pedestrians.
14	06:41	Driving, approaching intersection.	Cars, truck, pedestrians, cyclists.	Traffic light; "Yield to Traffic" sign visible.	Be cautious of signals and traffic.
15	06:50	Driving, approaching intersection.	Cars, truck, pedestrians, cyclists.	Traffic light; street sign visible.	Watch for other vehicles and pedestrians.
16	07:30	Driving, approaching intersection.	Cars, truck, pedestrians, cyclists.	Traffic light; "No right turn" sign visible.	Unclear if can proceed.
17	07:33	Driving in middle of right lane.	Several cars and a truck; pedestrian on sidewalk.	None visible.	Maintain safe distance from vehicles.
18	07:48	Driving, approaching intersection.	Two cars ahead; truck present.	Traffic light visible.	Caution needed.
19	07:57	Driving, approaching corner.	Two cars ahead; pedestrian crossing.	Traffic light; "No left turn" sign visible.	Pedestrian crossing is hazard.
20	08:11	Driving, approaching building with fence and sign.	No visible vehicles or obstacles.	Traffic light visible.	Fence and sign may require caution.
21	08:15	Driving, approaching curve.	Several cars and a truck, some opposing.	None visible.	Extra caution needed on curve.
22	08:17	Driving, approaching intersection.	Several cars and truck; some stopped.	Traffic light; "No left turn" sign visible.	Be cautious of signals and traffic.

As highlighted by the LLM, Figure 4 shows frame 11, where the ego-vehicle is driving down the road in the middle of the right lane, surrounded by multiple cars and a bus. Traffic lights are visible, and the proximity to the bus presents a potential hazard. Key frame 11 exemplifies how the summarizer captures the ego-vehicle's perspective in a nuanced and context-aware manner. The initial portion of the description establishes the precise position of the car, providing a clear spatial grounding for the vehicle.

The summary continues by noting the presence of "*multiple cars and a bus on the road, with some cars in front of the car and others behind it*", illustrating the complexity of the surrounding traffic environment. While the visual data includes "*traffic lights visible in the scene*", the LLM accurately reflects their uncertain status by indicating that their positions and states are not specified. Most critically, the summarizer identifies a possible risk: "*The car is driving near a bus, which could potentially obstruct its view of the road.*" This blend of detail and restraint shows the model's ability to distill relevant situational cues—offering a compact yet informative summary that approximates the type of awareness a human or robotic driver would require.

VI. Conclusion

To sum up, this paper proposes a method to identify and summarize video content through key frame selection. The generated summaries synthesize the most relevant aspects within each frame, focusing on elements aligned with task-specific objectives. A handcrafted prompt guides an LLM to describe both the scene context and the robot's behavior, leveraging the structured nature of the data and the clarity of the driving environment.

These summaries provide a compact yet informative representation of the video, supporting downstream tasks such as behavior analysis, anomaly detection, and performance assessment in self-driving scenarios. By enabling targeted browsing of specific environmental instances, the method facilitates the retrieval of meaningful samples, refinement of robotic behavior, and identification of problematic situations. Consequently, human effort can be redirected toward high-impact analysis and the exploration of critical or promising records.

As future work, we aim to improve the structure and accuracy of the key frame detection process by employing more contextualized scene filtering methods. In addition, we plan to evolve the LLM-based summarization to operate effectively in more complex environments and accommodate a wider range of robotic behaviors and tasks. This includes studying movement dynamics, scene rotation, and divergence across different scene structures, which are essential for capturing context shifts and behavioral adaptations in real-world scenarios.

References

[1] M. Misaros, O.-P. Stan, I.-C. Donca, and L.-C. Miclea, "Autonomous robots for services—state of the art, challenges, and research areas," *Sensors*, vol. 23, no. 10, 2023.

[2] J. Higgins and N. Bezzo, "Negotiating visibility for safe autonomous navigation in occluding and uncertain environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4409–4416, 2021.

[3] M. Gadd, D. de Martini, L. Marchegiani, P. Newman, and L. Kunze, "Sense–assess–explain (sax): Building trust in autonomous vehicles in challenging real-world driving scenarios," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, vol. 4, no. 1, 2020, pp. 150–155.

[4] J.-A. Bolte, A. Bar, D. Lipinski, and T. Fingscheidt, "Towards corner case detection for autonomous driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, vol. 4, no. 1, 2019, pp. 438–445.

[5] K. D. Kusano, J. M. Scanlon, Y.-H. Chen, T. L. McMurry, R. Chen, T. Gode, and T. Victor, "Comparison of waymo rider-only crash data to human benchmarks at 7.1 million miles," *Traffic Injury Prevention*, vol. 25, no. 1, pp. S66–S77, 2024.

[6] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, no. 2, pp. 690–712, 2021.

[7] R. Xu, Y. Shen, X. Li, R. Wu, and H. Dong, "Naturalvlm: Leveraging fine-grained natural language for affordance-guided visual manipulation," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 10 842–10 849, 2024.

[8] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, T. Jiang, S. Li, J. Xu, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2: Scaling foundation models for multimodal video understanding," in *Computer Vision ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15143, no. 1. Cham: Springer Nature Switzerland, 2025, pp. 396–416.

[9] D. M. Argaw, S. Yoon, F. C. Heilbron, H. Deilamsalehy, T. Bui, Z. Wang, F. Dernoncourt, and J. S. Chung, "Scaling up video summarization pretraining with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 8332–8341.

[10] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu, A. Vosoughi, C. Huang, Z. Zhang, P. Liu, M. Feng, F. Zheng, J. Zhang, P. Luo, J. Luo, and C. Xu, "Video understanding with large language models: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[11] H. Hua, Y. Tang, C. Xu, and J. Luo, "V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, pp. 3599–3607, Apr. 2025.

[12] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep 2018.

[13] H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, and P. Rota, "Deep unsupervised key frame extraction for efficient video classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 3, pp. 1–17, Feb 2023.

[14] R. Tan, X. Sun, P. Hu, J.-h. Wang, H. Deilamsalehy, B. A. Plummer, B. Russell, and K. Saenko, "Koala: Key frame-conditioned long video-llm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 581–13 591.

[15] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Event detection on roads using perceptual video summarization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 2944–2954, 2018.

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916.

[17] M. J. Lee, D. Gong, and M. Cho, "Video summarization with large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 18 981–18 991.

[18] Y. D. V. Yasuda, L. E. G. Martins, and F. A. M. Cappabianco, "Autonomous visual navigation for mobile robots: A systematic literature review," *ACM Comput. Surv.*, vol. 53, no. 1, feb 2020.

[19] G. Singh, S. Akrigg, M. D. Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson, J. Omokeowa, S. Grazioso, A. Bradley, G. D. Gironimo, and F. Cuzzolin, " ROAD: The Road Event Awareness Dataset for Autonomous Driving ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 01, pp. 1036–1054, Jan. 2023.