automation

Article

Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems

Huthaifa I. Ashqar [1,2,*], Taqwa I. Alhadidi [3], Mohammed Elhenawy [4] and Nour O. Khanfar [5]

1 Civil Engineering Department, Arab American University, Jenin P.O. Box 240, Palestine
2 Artificial Intelligence Program, Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY 10027, USA
3 Civil Engineering Department, Al-Ahliyya Amman University, Amman 19328, Jordan; t.alhadidi@ammanu.edu.jo
4 CARRS-Q, Queensland University of Technology, Brisbane, QLD 4001, Australia; mohammed.elhenawy@qut.edu.au
5 Natural, Engineering and Technology Sciences Department, Arab American University, Jenin P.O. Box 240, Palestine; n.khanfar1@student.aaup.edu
* Correspondence: huthaifa.ashqar@aaup.edu

Abstract: The integration of thermal imaging data with multimodal large language models (MLLMs) offers promising advancements for enhancing the safety and functionality of autonomous driving systems (ADS) and intelligent transportation systems (ITS). This study investigates the potential of MLLMs, specifically GPT-4 Vision Preview and Gemini 1.0 Pro Vision, for interpreting thermal images for applications in ADS and ITS. Two primary research questions are addressed: the capacity of these models to detect and enumerate objects within thermal images, and to determine whether pairs of image sources represent the same scene. Furthermore, we propose a framework for object detection and classification by integrating infrared (IR) and RGB images of the same scene without requiring localization data. This framework is particularly valuable for enhancing the detection and classification accuracy in environments where both IR and RGB cameras are essential. By employing zero-shot in-context learning for object detection and the chain-of-thought technique for scene discernment, this study demonstrates that MLLMs can recognize objects such as vehicles and individuals with promising results, even in the challenging domain of thermal imaging. The results indicate a high true positive rate for larger objects and moderate success in scene discernment, with a recall of 0.91 and a precision of 0.79 for similar scenes. The integration of IR and RGB images further enhances detection capabilities, achieving an average precision of 0.93 and an average recall of 0.56. This approach leverages the complementary strengths of each modality to compensate for individual limitations. This study highlights the potential of combining advanced AI methodologies with thermal imaging to enhance the accuracy and reliability of ADS, while identifying areas for improvement in model performance.

Keywords: multimodal large language models (MLLMs); thermal images; RGB; object detection; autonomous driving systems

Citation: Ashqar, H.I.; Alhadidi, T.I.; Elhenawy, M.; Khanfar, N.O. Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems. *Automation* **2024**, *5*, 508–526. https://doi.org/10.3390/automation5040029

Academic Editor: Duc Truong Pham

Received: 19 August 2024
Revised: 19 September 2024
Accepted: 8 October 2024
Published: 10 October 2024

1. Introduction

Thermal imaging plays a pivotal role in the advancement of autonomous driving systems due to its ability to detect and interpret heat signatures, enhancing the perception capabilities of autonomous vehicles [1–3]. One of the key benefits of thermal imaging is its ability to provide enhanced visibility in adverse conditions, both in urban and highway environments. Thermal cameras can detect heat emitted by objects and living beings, offering clear images even in complete darkness. This capability is crucial for detecting pedestrians, animals, and vehicles at night [1].

Deep learning methods, such as convolutional neural networks (CNNs), have proven effective in automating object detection in thermal images [2,4,5]. These methods enable autonomous vehicles to detect and react to various objects and obstacles in their environment, which directly impacts the safety and efficiency of transportation systems. Vehicles equipped with sensor technologies like LiDAR, cameras, and thermal imaging can integrate their data into a unified sensor framework [6]. By combining these technologies with machine learning, thermal imaging facilitates object detection and lane-keeping, contributing to a safer and more comfortable driving experience [7]. Thermal imaging also plays a role in advanced driver assistance systems (ADAS) for automated driving [8].

Integrating thermal imaging with autonomous driving systems provides critical redundancy and enhances safety. Combining thermal imaging with sensors such as LiDAR, radar, and visible light cameras creates a more reliable perception system. Thermal imaging allows for early hazard detection, enabling vehicles to take preventive action more quickly than they would using visible light cameras alone, significantly reducing accident risks [6,7]. However, challenges such as the high cost of thermal cameras and the need for sophisticated algorithms for seamless integration with other sensors remain [7]. Ongoing research aims to develop algorithms that can effectively interpret thermal images and integrate this data with other sensor inputs. Machine learning and AI are crucial in improving the accuracy and reliability of these systems.

Multimodal large language models (MLLMs) are designed to handle data from multiple modalities, such as images, videos, and text, which enhances their functionality and safety in computer-driven systems. MLLMs have shown great potential in supporting tasks that require collaborative learning across various domains [9]. Their application spans from educational tools to clinical technologies, highlighting their versatility [10,11]. Moreover, integrating language models with image understanding is an emerging trend that promises significant advancements [12]. MLLMs have also demonstrated promise in computerized driving, where vision-language pre-training (VLP) has proven useful for tasks like image captioning and visual question answering [13,14].

MLLMs can overcome limitations associated with traditional thermal image object detection methods by leveraging their advanced feature extraction capabilities, contextual understanding, and multimodal data integration [15,16]. Unlike traditional techniques that often rely on manual feature engineering, MLLMs can automatically learn rich, hierarchical representations of thermal images through deep learning architectures, resulting in more accurate object detection [17,18]. Furthermore, MLLMs can integrate thermal data with other sensor inputs, such as RGB or LiDAR, providing a comprehensive understanding of the different scenes [19,20]. This multimodal approach improves the handling of variable object appearances and environmental conditions, leading to more reliable and adaptable autonomous systems.

LLMs enhance the multimodal capabilities of thermal imaging systems by processing complex information, understanding context, and reasoning logically [21]. They also bridge thermal perception with natural language processing (NLP), enabling the interpretation of thermal image data in textual form, and making it more accessible to non-experts. Recently, researchers have explored multimodal transformers for cross-modal representation learning, which is critical for tasks like image captioning and visual question answering [22]. These techniques improve the interpretability of thermal imaging data, facilitating better decision-making. While unimodal systems are limited in providing a complete picture, multimodal approaches offer significant benefits. Moreover, the integration of thermal imaging with traditional RGB data enhances object detection algorithms' robustness, especially under low-visibility or adverse weather conditions. Thermal images improve the detection of objects invisible in standard RGB images, thereby enhancing the situational awareness of autonomous systems [23,24].

This study explores the capabilities of MLLMs, specifically GPT-4 Vision Preview and Gemini 1.0 Pro Vision, in understanding thermal images for autonomous driving systems (ADS) and intelligent transportation systems (ITS). By addressing two key research questions, we assess the MLLMs' ability to detect and enumerate objects in thermal images and to discern whether pairs of images depict the same scene. Utilizing zero-shot in-context learning for object detection and employing the chain-of-thought technique for scene comparison, the study demonstrates how advanced AI methodologies can push the boundaries of thermal image analysis. Additionally, we propose a novel framework that integrates IR and RGB images to detect and classify objects within the same scene. This framework is particularly useful in environments requiring the use of both cameras. This research not only highlights the potential of MLLMs to enhance object detection accuracy and environmental understanding but also sets the stage for safer, more reliable autonomous driving systems, illustrating the transformative impact of integrating generative AI into transportation technologies.

2. Literature Review

Traditional methodologies for detecting objects in thermal images have heavily relied on image processing techniques and classical machine learning algorithms [2,7,8]. These methods include thresholding, where thermal images are converted into binary images based on predefined temperature thresholds, effectively distinguishing objects from the background. Another common approach is edge detection, using techniques like the Sobel, Canny, and Laplacian operators to identify object boundaries by analyzing temperature gradients. Segmentation methods, such as region growing, clustering, and watershed segmentation, partition thermal images into segments or regions that correspond to different objects or areas of interest. While these methods can isolate objects from the background and improve accuracy through further refinement, they are often computationally intensive and sensitive to noise [25,26]. These algorithms rely on extracting relevant features like texture, shape, and intensity from thermal images to train classifiers capable of identifying and categorizing objects [27]. Although promising, these methods require extensive feature engineering and may not generalize well to diverse or unseen scenarios. Template matching, another traditional approach, involves matching predefined templates of objects against thermal images to detect similar patterns [2,4,28]. While straightforward and effective for detecting objects with consistent shapes and sizes, template matching can struggle with variations in object appearance or scale.

The integration of MLLM into detecting objects using thermal images is an emerging research direction with potential applications. Some multimodal models, such as CLIP, have demonstrated great potential in connecting imaging and natural language and fitting into thermal images, demonstrating that such models could reach new levels of downstream performance [29,30]. The multimodal guide also applies such information in the upstream stream to increase the performance of the visual downlink mission to enable the second stream to consider the image information contained in the acquisition of a multimodal machine learning model language that overcomes language variation depending on the classification of the object and the update of the online content [31,32]. The potential of large-scale dual-stream vision-language pre-training, such as CLIP and ALIGN, has proven to be useful for the overall performance and downstream goals of various multimodal alignment levels, including image text recovery and imaging [33]. Multimodal transformer networks have shown the best performance with excellent cross-functional modality, ideal for various vision and linguistic tasks such as image text recovery and image indication [33]. Additionally, to have contextually relevant crosses of speech with crossed-image labels, "tokenization" has three ideas because it supports the extrapolation of various multimodal alignment scopes to language-only scope data [34]. This idea would be suitable for incorporating thermal images into a large speech model integration, as it would facilitate a direct connection between visual and noun semantics. Finally, a large image dataset, such as GEM, is a viable option for researchers to align imaging artifacts at high- and multilingual labeling points to assign a perfect image for the first amendment to a multimodal model fit application [35].

In image interpretation and remote sensing, feature cameras achieve greater color accuracy than RGB technology, using a multispectral filter wheel [36]. Moreover, in drones, the combination of RGB-IR conservation aids in recognizing objects generated by transportation [37]. Moreover, RGB-D images can be used for segmentation to show reflections in an elevator environment, as the authors proved difficult to transport [38]. In geographic information sciences and computer vision, human orientation is estimated from the RGB photographs of users [39,40]. Encryption and image secret writing are often combined using RGB images to protect or distort the data [41].

While text-based language models such as GPT-3 [42], BERT [43], and RoBERTa [44] outperform humans in text production and encoding tasks, their comprehension and processing capabilities are almost non-existent, considering that understanding data encompasses a variety of other types. MLLMs solve this issue by working with various data types and by introducing opportunities to work with other types of data by transcending data models that only work with text. MLLMs demonstrated high performance on major league benchmarks for image and text input [45]. The ability to perceive and understand inputs from multiple sensory modalities is a crucial aspect of AI development. It is critical to be able to learn and navigate successfully in the physical world. The main contribution of this study is to introduce a novel MLLM framework that integrates the use of IR-RGB combination, which can be beneficial when the road environment requires the use of the two cameras such as at night-time and during adverse weather conditions.

3. Methodology

3.1. Visual Reasoning and Scene Understanding

The methodology for the first part of the study is illustrated in Figure 1. This flowchart outlines the methodology that aims to explore the visual reasoning and scene understanding capabilities of MLLMs (namely, GPT-4 Vision Preview and Gemini 1.0 Pro Vision) in understanding thermal images for applications in ADS and ITS. The methodology is divided into two primary research questions (RQs) and associated experiments. The first RQ provides a generalized understanding of MLLMs for object detection and enumeration using thermal training image subsets. This experiment focuses on assessing the MLLM's ability to detect and enumerate objects within thermal images. To address the first question, we designed a study utilizing zero-shot in-context learning. MLLM was used to identify objects within a thermal image and to enumerate the occurrence of each object. The second RQ deals with discerning the same scene images. This experiment evaluates the MLLM's capability to discern whether two images represent the same scene. It includes two iterations: one with pairs of images from the same scene and another with pairs from different scenes. Both iterations use test data that include mapping information to verify the model's accuracy. Specifically, to answer the second question, we experimented with thermal and RGB test images that have a one-to-one correspondence and have been captured in identical scenes. We employed the chain-of-thought technique in our prompt design, instructing the model to describe the RGB image followed by a thermal image. Subsequently, the model was used to estimate the likelihood of both images originating from the same scene. This experiment was conducted in two iterations: the first with image pairs from the same scene and the second with image pairs from separate scenes. Examples of the used prompt are shown in Table 1. The results of these experiments are used as a base for building the proposed MLLM framework for object detection using a combination of IR and RGB images.

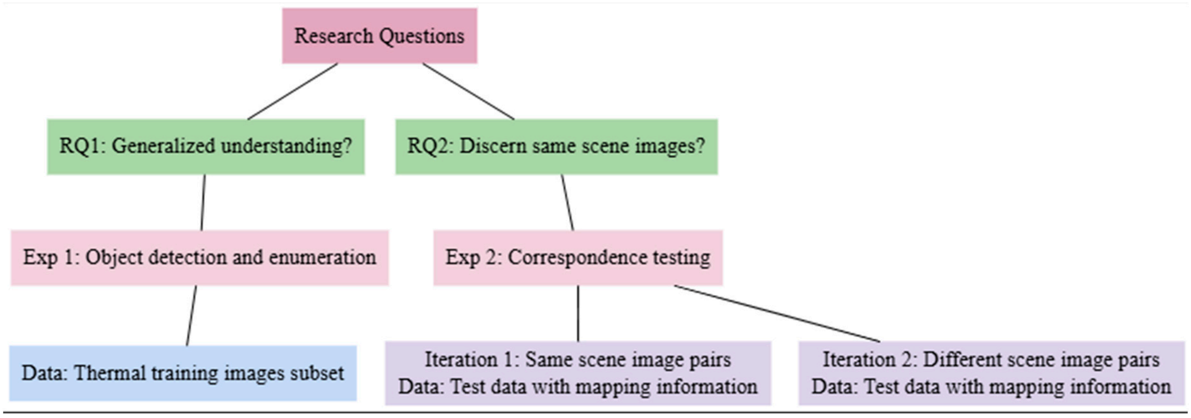

Figure 1. Proposed methodology for testing MLLMs' visual reasoning and scene understanding.

Table 1. Examples of used prompts.

Do the following/1—Describe the first image/2—describe the second image/3—compare the two descriptions then What is the probability that these two images were captured of the same scene? Please return a number between zero and one, where one means you are certain the two images depict the same scene.
Your role is systematically inspecting RGB image for ADAS and autonomous vehicles applications, analyze the provided RGB image. Your output should be formatted as a Python dictionary. Each key in the dictionary represents a category ID from the list of interest, and each corresponding value should be the frequency at which that category appears in the scene captured by the RGB image. The analysis must focus exclusively on the following categories:/Category Id 1: Person/Category Id 2: Bike/Category Id 3: Car (includes pickup trucks and vans)/Category Id 4: Motorcycle/Category Id 6: Bus/Category Id 7: Train/Category Id 8: Truck (specifically semi/freight trucks, not pickup trucks)/Category Id 10: Traffic light/Category Id 11: Fire hydrant/Category Id 12: Street sign/Category Id 17: Dog/Category Id 37: Skateboard/Category Id 73: Stroller (also known as a pram, a four-wheeled carriage for a child)/Category Id 77: Scooter/Category Id 79: Other vehicle (includes less common vehicles such as construction equipment and trailers)/Do not include descriptive text or explanations in your output. Only provide the dictionary with the category IDs and frequencies as values. The dictionary should look similar to this: {1: 0, 2: 1, 3: 3, . . .}. If a category is not present in the image, assign a frequency of 0 to it in the dictionary.
Your role is systematically inspecting thermal image for ADAS and autonomous vehicles applications, analyze the provided thermal image. Your output should be formatted as a Python dictionary. Each key in the dictionary represents a category ID from the list of interest, and each corresponding value should be the frequency at which that category appears in the scene captured by the RGB image. The analysis must focus exclusively on the following categories:/Category Id 1: Person/Category Id 2: Bike/Category Id 3: Car (includes pickup trucks and vans)/Category Id 4: Motorcycle/Category Id 6: Bus/Category Id 7: Train/Category Id 8: Truck (specifically semi/freight trucks, not pickup trucks)/Category Id 10: Traffic light/Category Id 11: Fire hydrant/Category Id 12: Street sign/Category Id 17: Dog/Category Id 37: Skateboard/Category Id 73: Stroller (also known as a pram, a four-wheeled carriage for a child)/Category Id 77: Scooter/Category Id 79: Other vehicle (includes less common vehicles such as construction equipment and trailers)/Do not include descriptive text or explanations in your output. Only provide the dictionary with the category IDs and frequencies as values. The dictionary should look similar to this: {1: 0, 2: 1, 3: 3, . . .}. If a category is not present in the image, assign a frequency of 0 to it in the dictionary.
Your role is systematically inspecting pair of thermal and RGB images for ADAS and autonomous vehicles applications, analyze the provided pair of images. Your output should be formatted as a Python dictionary. Each key in the dictionary represents a category ID from the list of interest, and each corresponding value should be the frequency at which that category appears in the scene captured by image pair. The analysis must focus exclusively on the following categories:/Category Id 1: Person/Category Id 2: Bike/Category Id 3: Car (includes pickup trucks and vans)/Category Id 4: Motorcycle/Category Id 6: Bus/Category Id 7: Train/Category Id 8: Truck (specifically semi/freight trucks, not pickup trucks)/Category Id 10: Traffic light/Category Id 11: Fire hydrant/Category Id 12: Street sign/Category Id 17: Dog/Category Id 37: Skateboard/Category Id 73: Stroller (also known as a pram, a four-wheeled carriage for a child)/Category Id 77: Scooter/Category Id 79: Other vehicle (includes less common vehicles such as construction equipment and trailers)/Do not include descriptive text or explanations in your output. Only provide the dictionary with the category IDs and frequencies as values. The dictionary should look similar to this: {1: 0, 2: 1, 3: 3, . . .}. If a category is not present in the image, assign a frequency of 0 to it in the dictionary.

The prompt followed a well-structured and logical method of chain-of-thought, addressing key aspects of thermal image understanding in the context of MLLMs. The division into two research questions ensures a comprehensive assessment of the model's capabilities, covering both general object detection and specific scene recognition. This methodology also significantly contributes to the field of autonomous driving and ITS by addressing a crucial gap of the integration of thermal imaging with MLLMs. Thermal imaging is essential for night-time and adverse weather conditions, where traditional RGB cameras may fail. By assessing MLLMs' capabilities in understanding and detecting objects in thermal images, the study pushes the boundaries of current ITS technologies. This step will also lead to integrating IR with RGB data to create a reliable and robust MLLM framework for detecting and classifying objects, which will be discussed in the next section. This framework can be used for autonomous systems, enhancing safety and functionality.

3.2. Proposed MLLM Object Detection Framework Using IR-RGB Combination

This section presents a novel MLLM framework using Gemini 1.5 Pro that can integrate IR and RGB images of the same scene using a zero-shot in-context approach without providing localization information. This framework can be used for better object detection and can be deployed for ADS and ITS applications, especially when the environment requires the use of the two cameras (i.e., IR and RGB). Object detection using MLLMs can help overcome the limitations of traditional object detection methods by leveraging their advanced capabilities in feature extraction, contextual understanding, and integration of multiple data modalities. MLLMs can save time and resources if we were able to use its pre-trained capabilities for object detection. Moreover, MLLMs can also offer explainability and the opportunity to provide context and recommendations as well as understand direct feedback, which are considered major challenges when using deep learning.

To build the framework, we conducted three experiments that integrate IR and RGB images for the same scene utilizing the zero-shot in-context learning as shown in Figure 2. In the first experiment, we presented both IR and RGB images of the same scene to the MLLM model without specific instructions on how to combine the modalities for detecting and classifying objects. The objective is to perform object detection and classification directly from this combination. Let I_{IR} represent the IR images, I_{RGB} represent the RGB images, D represent the detected objects, and C represent the classifications of these objects. The MLLM's output can be represented as shown in Equation (1):

$$(D_1, C_1) = f(I_{IR},\ I_{RGB}) \tag{1}$$

where f is the MLLM's function that outputs the detected objects D_1 and their classifications C_1 based on the input images for the first experiment.

The second experiment involved a two-stage prompting process. Initially, we presented IR images and requested object detection and classification. Following this, we supplied this information obtained from the IR images along with an RGB image and a new prompt instructing the model to refine or reassess the object detection and classification based on the combined outputs. Let $P(D_{IR}, C_{IR}|I_{IR})$ represent the probability of detecting and classifying objects based on the IR image, and $P(D_2, C_2|D_{IR}, C_{IR},\ I_{RGB})$ represent the probability of refining the detection and classification using the RGB images, the process of the second experiment can be represented as Equation (2):

$$P(D_2, C_2|I_{IR}, I_{RGB}) = P(D_2, C_2|D_{IR}, C_{IR},\ I_{RGB}) \times P(D_{IR}, C_{IR}|I_{IR}) \tag{2}$$

In the third experiment, we reversed the order of image presentation compared to the second experiment. Initially, we presented the RGB images and asked for the object detection and classification. Subsequently, the IR images were shown alongside the results from the first prompt. Let $P(D_{RGB}, C_{RGB}|I_{RGB})$ represent the probability of detecting and classifying objects based on the RGB images, and $P(D_3, C_3|D_{RGB}, C_{RGB},\ I_{IR})$ represent the probability of refining the detection and classification using the IR images, the process of the third experiment can be represented as Equation (3):

$$P(D_3, C_3|I_{RGB}, I_{IR}) = P(D_3, C_3|D_{RGB}, C_{RGB},\ I_{IR}) \times P(D_{RGB}, C_{RGB}|I_{RGB}) \tag{3}$$

These equations capture the essence of each experimental setup, with the second and third experiments using Bayesian probability to model the refinement process. This approach enabled us to obtain varied responses for the same two images, allowing us to form a type of ensemble that could potentially yield more accurate results. By comparing different analytical sequences, we can assess how sequence variations influence the MLLM's performance and possibly enhance the reliability of its outcomes through ensemble techniques.

This approach involves generating structured prompts that guide the analysis process, with each prompt tailored to specific image modalities or combinations thereof. Example of the used prompts are shown in Table 2. Outputs are formatted as Python dictionaries, mapping category IDs to the frequency of their occurrence within the images. This structured approach ensures consistency across experiments and provides clarity on how each image type contributes to the detection and classification tasks.

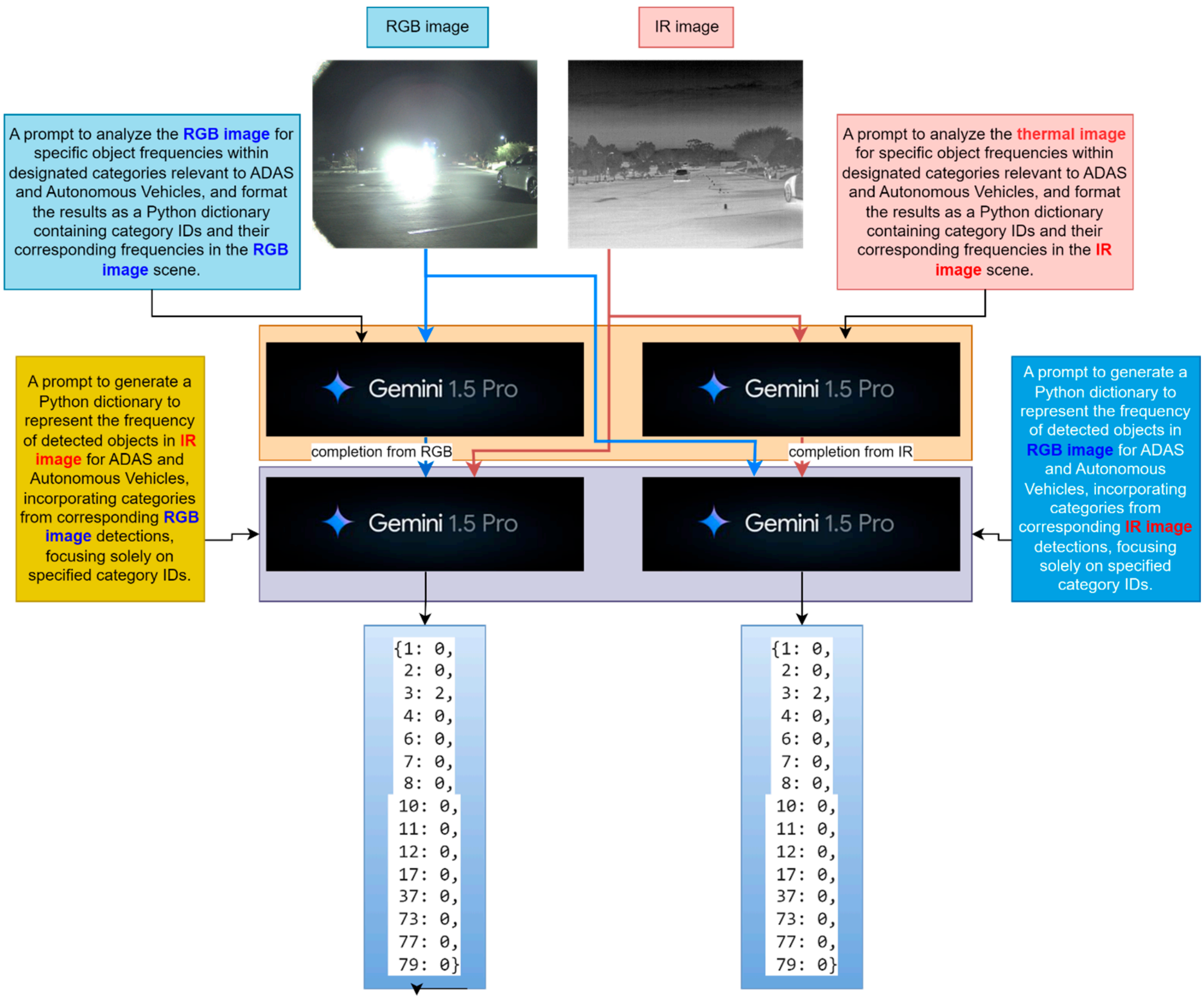

Figure 2. Proposed framework using two-stage zero-shot in-context learning to detect objects across two modalities.

The practical implications of this framework are substantial. For ADS, the ability to accurately detect objects and discern scenes using thermal imaging can enhance safety, particularly in low-visibility conditions. This can prevent accidents and improve navigation. In the broader context of ITS, such capabilities can aid in traffic management and monitoring, providing reliable data even in challenging environments. The framework also offers an opportunity for future research, enabling continuous improvement and innovation in the integration of MLLMs with various sensor modalities.

Table 2. Examples of used prompts for building the framework.

Experiment 1: IR and RGB combined	Your role is systematically inspecting pair of thermal and RGB images for ADAS and autonomous vehicles applications, analyse the provided pair of images. Your output should be formatted as a Python dictionary. Each key in the dictionary represents a category ID from the list of interest, and each corresponding value should be the frequency at which that category appears in the scene captured by image pair. The analysis must focus exclusively on the following categories:/Category Id 1: Person/Category Id 2: Bike/Category Id 3: Car (includes pickup trucks and vans)/Category Id 4: Motorcycle/Category Id 6: Bus/Category Id 7: Train/Category Id 8: Truck (specifically semi/freight trucks, not pickup trucks)/Category Id 10: Traffic light/Category Id 11: Fire hydrant/Category Id 12: Street sign/Category Id 17: Dog/Category Id 37: Skateboard/Category Id 73: Stroller (also known as a pram, a four-wheeled carriage for a child)/Category Id 77: Scooter/Category Id 79: Other vehicle (includes less common vehicles such as construction equipment and trailers)/Do not include descriptive text or explanations in your output. Only provide the dictionary with the category IDs and frequencies as values. The dictionary should look similar to this: {1: 0, 2: 1, 3: 3, . . .}. If a category is not present in the image, assign a frequency of 0 to it in the dictionary.

Experiment 2: RGB given IR	def generate_prompt(detected_objects):
	category_descriptions = {
	1: “Person”,
	2: “Bike”,
	3: “Car (includes pickup trucks and vans)”,
	4: “Motorcycle”,
	6: “Bus”,
	7: “Train”,
	8: “Truck (specifically semi/freight trucks, not pickup trucks)”,
	10: “Traffic light”,
	11: “Fire hydrant”,
	12: “Street sign”,
	17: “Dog”,
	37: “Skateboard”,
	73: “Stroller (also known as a pram)”,
	77: “Scooter”,
	79: “Other vehicle (includes construction equipment and trailers)”}
	# Start building the prompt
	prompt = (“Your task is to systematically inspect RGB images for ADAS and autonomous vehicles applications,”
	“taking into account prior detections from corresponding IR images. Detected objects in the IR image are provided in the format”
	f“‘{detected_objects}’. Note that certain categories might appear exclusively in the IR images and not in the RGB counterparts, and vice versa.\n\n”
	“Your findings should be structured as a Python dictionary, where each key corresponds to a category ID of interest and each value indicates”
	“the frequency of its appearance in both RGB and IR scenes. Concentrate your analysis on the following categories:\n\n”)
	# Add category list
	for category_id, description in category_descriptions.items():
	prompt += f“—Category Id {category_id}: {description}\n”
	# Finish the prompt
	prompt += (“\nDo not include descriptive text or explanations in your output. Provide only the dictionary with the category IDs and frequencies as values,”
	“such as ‘{1: 0, 2: 1, 3: 3, . . .}’. If a category is not present in either the RGB or IR image, assign a frequency of 0 to it in the dictionary.”)
	return prompt

Table 2. *Cont.*

Experiment 3: IR given RGB	def generate_prompt_IR_given_RGB(rgb_categories):
	    # Descriptions of categories detected in the RGB image
	    category_descriptions = {
	        1: "Person",
	        2: "Bike",
	        3: "Car (includes pickup trucks and vans)",
	        4: "Motorcycle",
	        6: "Bus",
	        7: "Train",
	        8: "Truck (specifically semi/freight trucks, not pickup trucks)",
	        10: "Traffic light",
	        11: "Fire hydrant",
	        12: "Street sign",
	        17: "Dog",
	        37: "Skateboard",
	        73: "Stroller (also known as a pram)",
	        77: "Scooter",
	        79: "Other vehicle (includes construction equipment and trailers)"}
	    # Start building the prompt
	    prompt = ("Your task is to systematically inspect IR images for ADAS and autonomous vehicles applications,"
	        "using the context of previously detected objects in corresponding RGB images. Categories detected in the RGB images are provided"
	        f"as '{rgb_categories}'. Be aware that some categories may appear in the RGB images but not in the IR images, and vice versa.\n\n"
	        "Your findings should be structured as a Python dictionary. Each key in the dictionary represents a category ID from the list of interest,"
	        "and each corresponding value should be the frequency at which that category appears in the scene captured by the IR image. Focus your analysis on the following categories:\n\n")
	    # Add category list
	    for category_id, description in category_descriptions.items():
	        prompt += f"- Category Id {category_id}: {description}\n"
	    # Finish the prompt
	    prompt += ("\nDo not include descriptive text or explanations in your output. Provide only the dictionary with the category IDs and frequencies as values,"
	        "such as '{1: 0, 2: 1, 3: 3, . . .}'. If a category is not present in the IR image, assign a frequency of 0 to it in the dictionary.")
	    return prompt

3.3. Dataset

We used the Teledyne FLIR Free ADAS Thermal Dataset V2, which is a comprehensive collection of annotated thermal and visible spectrum frames intended for the development of object detection neural networks [46]. This dataset aims to promote research on visible and thermal spectrum sensor fusion algorithms ("RGBT") to enhance the safety of autonomous vehicles. It comprises about 26,442 fully annotated frames covering 15 different object classes. The data were captured using a thermal and visible camera pair mounted on a vehicle, with the thermal camera operating in T-linear mode. Thermal images were acquired with a Teledyne FLIR Tau 2 13 mm f/1.0 camera, while visible images were captured with a Teledyne FLIR BlackFly S BFS-U3-51S5C (IMX250) camera [46]. Time-synced capture was facilitated by Teledyne FLIR's Guardian software, enabling frame rates of 30 frames per second in validation videos, which also include target IDs for tracking metrics computation. The dataset ensures diversity in training and validation by sampling frames from a wide range of footage, with some frames selected manually and others using a frame skip [46]. Redundant footage, such as identical frames during red light stops, was excluded by the curation team. Figure 3 shows an example of the images and corresponding annotations. Figure 4 shows a combination of IR-RGB images, which were sampled from diverse video sequences to be used in building the proposed framework. It showcases different traffic scenarios including urban, night-time, foggy, and bright sunlight conditions, supporting the comprehensive analysis of model performance. The images shown in Figure 4 capture diverse road scenarios, each of them extracted from video footage during sampling. Given that the videos are shot at 30 frames per second, and considering budget constraints, the framework was tested by sampling the video by selecting every IR-RBG 50th-frame image to be included in the dataset. This ensures that each snapshot represents a distinct moment, but also sufficiently representative of the overall video.

Figure 3. Example of the images and corresponding annotations [35]. Annotation of the classes in the image: {1: 1, #Person; 2: 0, # Bike; 3: 4, # Car; 4: 0, # Motorcycle; 6: 0, # Bus; 7: 0, # Train; 8: 0, # Truck; 10: 0, # Traffic light; 11: 0, # Fire hydrant; 12: 0, # Street sign; 17: 0, # Dog; 37: 0, # Skateboard; 73: 0, # Stroller; 77: 0, # Scooter; 79: 0 # Other vehicle}.

IR Image	RGB Image
	
	

Figure 4. *Cont.*

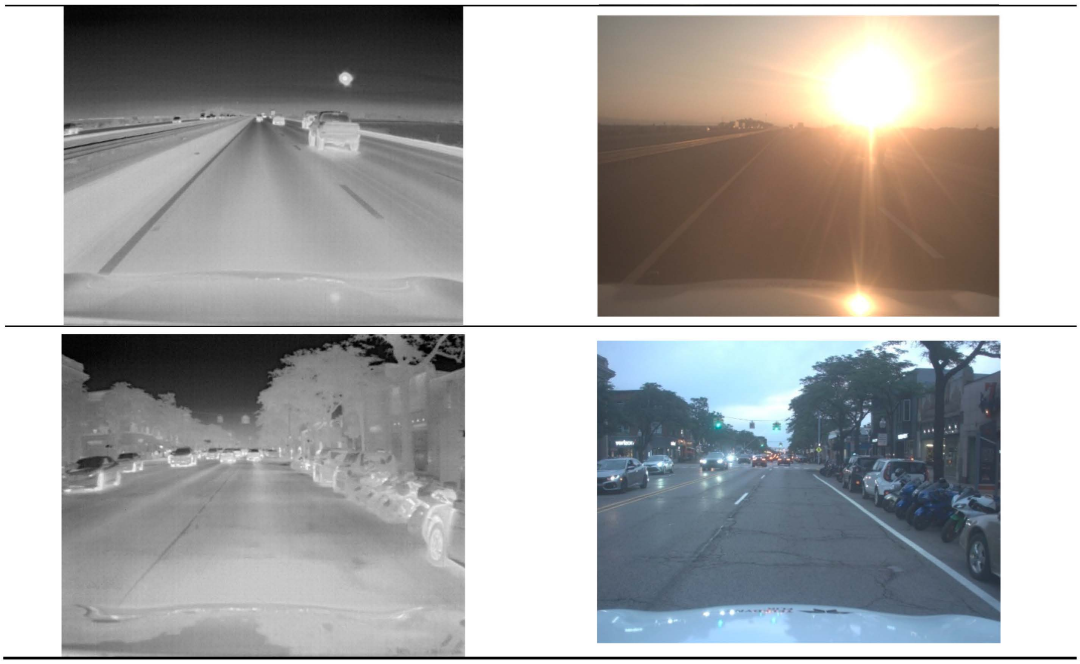

Figure 4. Combination of IR-RGB images, sampled from diverse video sequences, showcases different traffic scenarios including urban, night-time, foggy, and bright sunlight conditions, supporting comprehensive analysis of model performance [35].

4. Analysis and Results

4.1. Visual Reasoning and Scene Understanding Results

4.1.1. RQ1: Generalized Understanding

The main goal of the first experiment was to investigate the potential for generalizing MLLM knowledge across thermal and RGB imaging modalities. Zero-shot in-context learning methods indicate that the models can handle the process and analyze modalities. However, the accuracy of item detection varied among photos. Thermal images present a special challenge because they depend on thermal traces rather than visible light and contain less visual information. However, MLLM was able to accomplish moderate object recognition, including cars and people, demonstrating a solid starting point for future model development. The model also performed significantly well in recognizing and identifying objects. The confusion matrices for Gemini and GPT4 are shown in Figure 5 and Figure 6, respectively. The true positive rate for vehicle detection was 0.86, while that for motorcycle detection was 0.08. By contrast, RGB images contain more obvious visual hints, making them suitable for spotting smaller items.

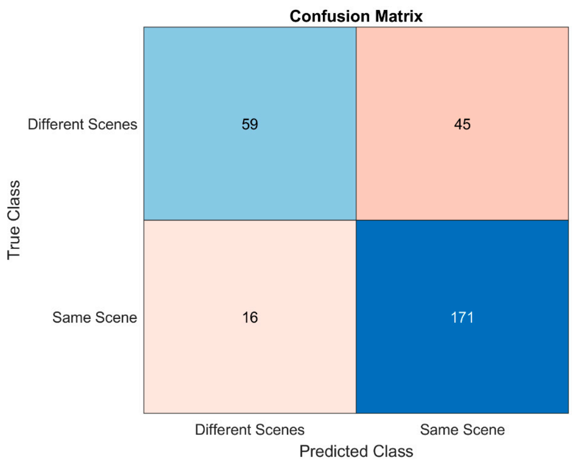

Figure 5. Gemini confusion matrix.

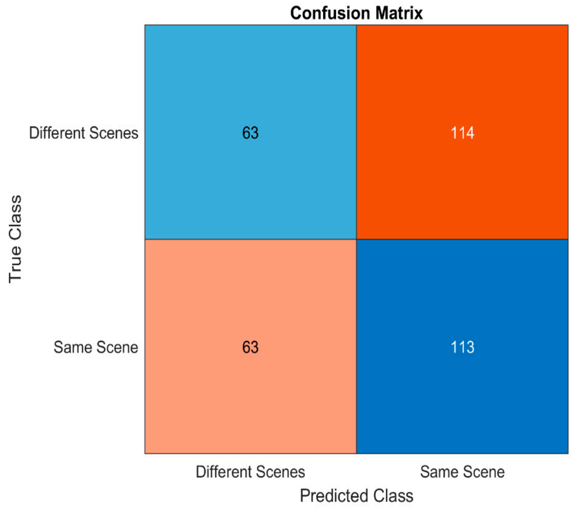

Figure 6. GPT4 confusion matrix.

4.1.2. RQ2: Discerning Same Scene Images

The model performed better when using two thermal images of the same scene and photos from two different image sensors, one with RGB and the other with infrared (IR). Using a chain-of-thought process, the models generated a comparison of photos from the same scenes and performed identical and diversified scenes. Accurate identification proved to be moderately adequate, with a recall of 0.91 and a precision of 0.79 for similar scenes. However, its performance deteriorated between different scenes, falling to a recall of 0.57 and a precision of 0.79. This variance illustrates the difficulty of the models in understanding their settings and distinguishing between identical scene structures using unique modality views. A comparison between the two models is presented in Table 3. On the other hand, Table 4 shows examples of MLLM answers to RQ2.

Table 3. Comparison between Gemini and GPT4.

Evaluation Metrics	Gemini 1.0 Pro Vision	GPT4 Vision Preview
Precision (same scene)	0.79	0.50
Recall (same scene)	0.91	0.64
F1 Score (same scene)	0.85	0.56
Precision (different scene)	0.79	0.50
Recall (different scene)	0.57	0.36
F1 Score (different scene)	0.66	0.42

4.1.3. Evaluation Results

Our evaluation methodology encompasses two primary aspects: accuracy of object category detection and precision of enumeration within these categories. For this analysis, the successful identification of at least one object within a given category is deemed sufficient for category detection without considering the total number of objects detected. To assess enumeration accuracy, we employed two statistical measures including the mean absolute error (MAE) and mean absolute percentage error (MAPE). These metrics were applied to each category of the annotations. In scenarios where a category is inferred by the MLLM but does not exist in the annotations, the instance or image is omitted from the MAE and MAPE calculations to avoid skewing the results. These instances were considered false positives.

Table 4. Examples of MLLM answers to RQ2.

MLLM Model	Input Images	MLLM Answer
GPT4	 	1. The first image depicts a dimly lit scene, possibly at night, with heavy fog or mist that reduces visibility. A single person is visible and walks away from the viewpoint of a bright light source that appears to be a street lamp. The ground shows parking lot markings, and the overall atmosphere is quite eerie owing to low light and fog. 2. The second image shows an urban environment with clearer visibility, which is likely captured during the day or evening in a well-lit area. There are multiple vehicles, both parked and in motion, and the architecture suggests an overpass or a bridge structure. The scene is black and white, and reflections on wet ground indicate recent rain or water presence. Comparing the two descriptions, the scenes depicted in the images were quite different. The first image is characterized by a solitary figure in a foggy, dimly lit parking lot, whereas the second image is an urban setting with multiple vehicles and clearer conditions. The architectural elements, lighting conditions, and presence of multiple subjects in the second image contrast with the solitary figure and the atmospheric conditions of the first image. Given these observations, the probability that these two images were captured in the same scene is very low. The differences in lighting, weather conditions, environment, and activity levels suggest that they represent different scenes. Therefore, the probability was estimated to be close to 0.
Gemini	 	1. The first image is a thermal image of a street at night. There are cars on the street, and people walk sidewalks. The image is blurry, and it is difficult to see many details. 2. The second image is a photograph of a parking lot at night. There is one person walking in the parking lot, and there are cars parked in the background. The image is also blurry, and it is difficult to see many details. 3. The two images are similar in that they are both thermal images of a street at night. However, there are also some differences between the two images. The first image is a wider shot, and it shows more of the street. The second image is a closer shot, and it shows more detail of the person walking. Overall, I would say that the probability that these two images were captured of the same scene is 0.5.
Gemini	 	1. The first image is a thermal image of a highway at night. The image is in black and white, and shows the highway from a driver's perspective. The highway is straight and there are cars driving in both directions. The image is clear and the details of the cars and the highway are visible. 2. The second image is a color image of a street intersection during the day. The image is in color and shows the intersection from a driver's perspective. The street is straight and there are cars driving in both directions. The image is clear and the details of the cars and the street are visible. 3. The two images are similar in that they are both images of roads from a driver's perspective. However, the images are different in that the first image is a thermal image and the second image is a color image. The first image is taken at night and the second image is taken during the day. The first image is of a highway and the second image is of a street intersection. Based on the differences between the two images, it is unlikely that they were captured of the same scene. I would give it a probability of 0.1.

Given the possibility of false positives, we adopted a binary approach to quantifying object presence within categories: a category was assigned a value of one if it contained at least one object and zero otherwise. This binary system facilitates direct comparison between the inferred data and annotations, enabling us to accurately calculate the true positive rate (TPR) and false positive rate (FPR) for each category. This methodological framework ensures a balanced evaluation of the detection efficacy and enumeration accuracy.

Table 5 provides a performance breakdown of two MLLMs; GPT-4 Vision Preview and Gemini 1.0 Pro Vision, when analyzing thermal images. These results were not intended to serve as a direct comparison between the two models; rather, the goal was to conduct a comprehensive analysis covering a wide range of images. This was to ensure that the models performed similarly and to demonstrate the potential of utilizing MLLM models with thermal imaging data. Given that these models are not freely available, a systematic selection process involving random sampling of images is required for zero-shot in-context learning for each model, thereby optimizing the utility derived from their application. This approach is used to maximize the value received from their use.

Table 5. Performance breakdown of GPT4 and Gemini analyzing thermal images.

MLLM Model	**GPT4 Vision Preview**				**Gemini 1.0 Pro Vision**			
Evaluation Metrics per Category	*TPR*	*FPR*	*MAE*	*MAPE*	*TPR*	*FPR*	*MAE*	*MAPE*
Category ID 1: Person	0.57	0	4.76	70.39	0.39	0	4.47	81.48
Category ID 2: Bike	0.31	0.01	1.89	78.40	0.41	0.02	1.37	66.53
Category ID 3: Car (includes pickup trucks and vans)	0.86	0.09	4.35	55.81	0.90	0.08	5.04	59.35
Category ID 4: Motorcycle	0.08	0	1.38	96.15	0.24	0.01	1.06	78.18

From Table 5, we can observe consistency in the performance across both models. This suggests that regardless of the model used, the ability to interpret the thermal image can be promising. Interestingly, there appears to be a trend where the larger the object, the better the models' enumeration and detection capabilities, with the car category showing high TPR for both models. Moreover, both models achieved zero FPR in detecting motorcycles, and the MAE and MAPE scores across categories indicated promising precision and recall, especially for larger objects such as cars.

According to the findings, the capacity to identify larger items such as cars and buses was far higher, demonstrating a substantial gap between categories. Because thermal and RGB sensors identify distinct characteristics, the performance disparity varies depending on the item category and situation. These results indicate that the models face challenges in accurately determining whether two images are captured in the same scene, primarily because of scene similarities, suboptimal lighting conditions, glare, or adverse weather conditions.

4.2. MLLM Framework of IR-RGB

For the three experiments that we tested for the MLLM framework, we used a sampled combination of IR-RGB images from the Teledyne FLIR Free ADAS Thermal Dataset V2. We investigated the detection and classification of eight different road objects (i.e., annotated classes) including person, bike, car, motorcycle, truck, traffic light, and street sign. To evaluate the framework, we performed a comparative analysis between the three experiments using mean absolute error (MAE) and symmetric mean absolute percentage error (SMAPE) across different object categories. SMAPE is preferred over MAPE in this case because SMAPE addresses the issue of asymmetry in MAPE. MAPE sometimes produces biased results if the actual values are small, leading to disproportionately large percentage errors. SMAPE normalizes the error by the average of the actual and predicted values, making it symmetric and providing a more balanced measure of prediction accuracy. It is also preferable when dealing with datasets that have a wide range of values, which is the case in our framework.

Figure 7 shows a comparative analysis of MAE and SMAPE across different objects when integrating both IR and RGB (i.e., the first experiment) and the other two Bayesian experiments, where one modality informs the other (i.e., IR given RGB, and RGB given IR). The integration of both IR and RGB modalities generally demonstrates a notable improvement in MAE and SMAPE across most categories compared to the results where only one modality informs the analysis of the other. This indicates a robust enhancement in detection accuracy when both modalities are used in conjunction, leveraging the strengths of each to compensate for the respective weaknesses.

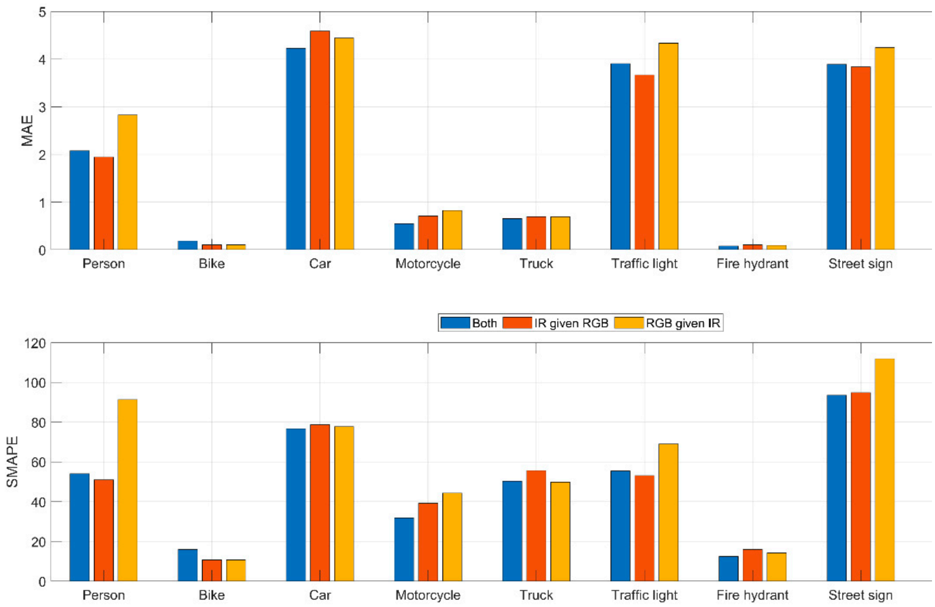

Figure 7. MAE and SMAPE for the three experiments in the MLLM framework across different objects.

Figure 8 shows a comparative analysis of precision and recall across different objects and for the three experiments. Integrating both modalities generally shows a higher performance across most categories, suggesting that the combination leverages the strengths of both modalities to reduce false positives. For instance, in complex categories like car, truck, and pedestrian which often involve challenging detection scenarios due to their variable appearances and sizes, the precision and recall are notably higher when both modalities are integrated. However, the bike category shows a surprisingly low performance in precision when using the integrated approach, and relatively low recall for all three experiments. This could be indicative of challenges in relatively smaller objects.

Results in Figure 8 show a significant enhancement in detection capabilities when both modalities were used in conjunction, with an average precision of 0.93 and an average recall of 0.56, leveraging the strengths of each IR and RGB to compensate for their respective weaknesses. The IR-RGB approach ensures a more comprehensive coverage, likely due to IR's strength in capturing objects in low-visibility conditions and RGB's high-resolution detail under good lighting. This suggests that integration increases overall detection capabilities. However, its effectiveness might vary based on the object type and the environmental conditions of the scenes being analyzed.

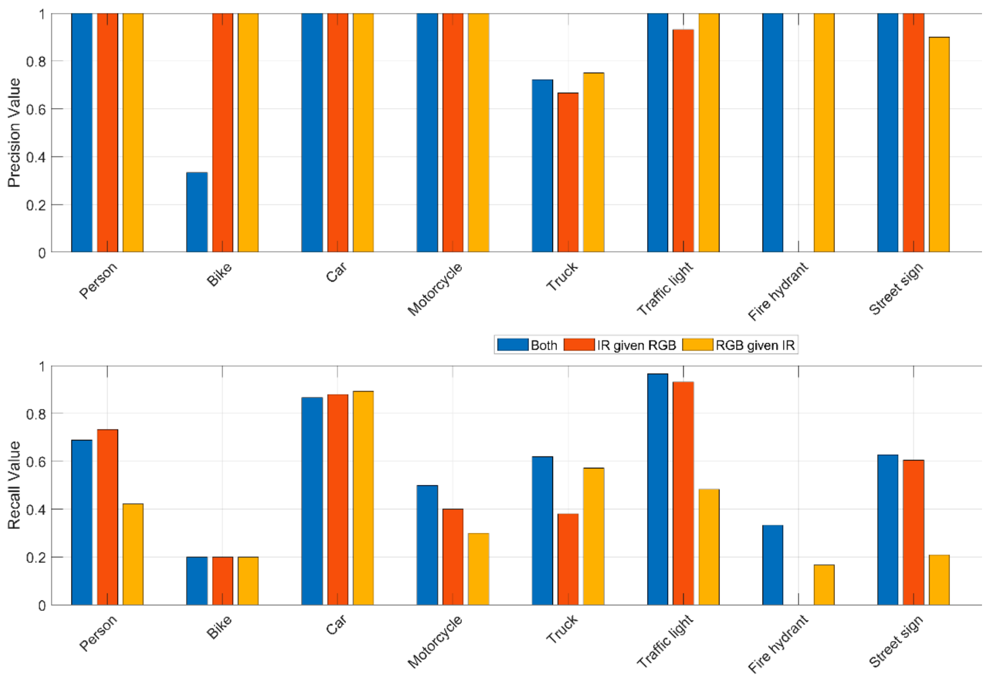

Figure 8. Precision and recall for the three experiments in the MLLM framework across different objects.

5. Conclusions

This study demonstrates the potential of MLLMs, specifically GPT-4 Vision Preview and Gemini 1.0 Pro Vision, in processing and analyzing thermal images for applications in ADS and ITS applications. We first assessed the models' generalized understanding and object detection capabilities using zero-shot in-context learning. Meanwhile, the models exhibited moderate success in recognizing and identifying objects within thermal images, especially in relatively larger object categories. However, the models struggled with smaller objects, as evidenced by the low detection rate of motorcycles at 0.08. This highlights the unique challenges posed by thermal imaging, which relies on heat signatures rather than visible light.

We then evaluated MLLM's ability to differentiate whether two images represent the same scene, using a combination of IR and RGB images. The results showed a recall of 0.91 and a precision of 0.79 for identifying similar scenes, indicating a moderately adequate performance. However, MLLM's ability to differentiate between different scenes was less robust, with a recall of 0.57 and a precision of 0.79. The consistency in performance across MLLM models suggests a promising potential for integrating MLLMs with thermal imaging data, despite the need for further refinement to enhance accuracy and adaptability in diverse scenarios. The ability of both models to generalize across RGB and thermal images is crucial to the robustness of autonomous systems under various lighting and weather conditions. The results indicate that, while generalization is possible, the differences in error rates across various object types and models imply that more calibration and training may be necessary. The safety of autonomous vehicles is highly dependent on accurate detection and identification of objects, particularly in complex environments. The existence of a somewhat high MAPE in specific categories implies that there might be instances where the models fail to accurately detect or misclassify objects, leading to hazardous driving decisions.

Finally, we built a novel MLLM framework using Gemini 1.5 Pro that can integrate IR and RGB images of the same scene using a zero-shot in-context approach without providing localization information. This framework was found promising and can be used for better object detection if deployed in ADS and ITS applications. This becomes more beneficial especially when the environment requires the use of the two cameras (i.e., IR and RGB).

Results showed that the MLLM framework achieved an average precision of 0.93 and an average recall of 0.56 when both modalities were used in conjunction. MLLM was able to leverage the strengths of each IR and RGB to compensate for their respective weaknesses. The IR-RGB MLLM model can ensure a more complete coverage in different scenarios including night-time and adverse weather conditions. This is likely due to IR's strength in capturing objects in low-visibility conditions and RGB's high-resolution detail under good lighting. This suggests that integration increases overall detection capabilities. However, its effectiveness might vary based on the object type and the environmental conditions of the scenes being analyzed.

The results suggest the need for ongoing enhancements to enhance the accuracy of the model, especially in diverse environmental conditions that autonomous vehicles may encounter. The findings endorse the ongoing use and enhancement of MLLMs for image-based processing in autonomous driving. The precision and reliability of the model training can be enhanced by integrating a broader array of images and environmental conditions. Moreover, these findings highlight the potential of employing these technologies in several domains that require robust and adaptable image recognition and processing skills, such as security and surveillance, environmental monitoring, and other related fields.

Author Contributions: Conceptualization, H.I.A., T.I.A., M.E. and N.O.K.; methodology, H.I.A. and M.E.; software, H.I.A. and M.E.; validation, H.I.A., T.I.A., M.E. and N.O.K.; formal analysis, M.E.; resources, M.E.; data curation, H.I.A., T.I.A. and M.E.; writing—original draft preparation, H.I.A. and T.I.A.; writing—review and editing, H.I.A., T.I.A., M.E. and N.O.K.; visualization, H.I.A. and M.E.; project administration, H.I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset is available through the FLIR website at: https://www.flir.com/oem/adas/adas-dataset-form/ (accessed on 10 March 2024).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Arnay, R.; Acosta, L.; Sigut, M.; Toledo, J. Asphalted road temperature variations due to wind turbine cast shadows. *Sensors* **2009**, *9*, 8863–8883. [CrossRef] [PubMed]
2. Ligocki, A.; Jelinek, A.; Zalud, L.; Rahtu, E. Fully automated dcnn-based thermal images annotation using neural network pretrained on rgb data. *Sensors* **2021**, *21*, 1552. [CrossRef] [PubMed]
3. Ashqar, H.I.; Jaber, A.; Alhadidi, T.I.; Elhenawy, M. Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive Review and Empirical Testing. *arXiv* **2024**, arXiv:2409.18286. [CrossRef]
4. Skladchykov, I.O.; Momot, A.S.; Galagan, R.M.; Bohdan, H.A.; Trotsiuk, K.M. Application of YOLOX deep learning model for automated object detection on thermograms. *Inf. Extr. Process.* **2022**, *2022*, 69–77. [CrossRef]
5. Zhang, C.; Okafuji, Y.; Wada, T. Reliability evaluation of visualization performance of convolutional neural network models for automated driving. *Int. J. Automot. Eng.* **2021**, *12*, 41–47. [CrossRef]
6. Hassouna, A.A.A.; Ismail, M.B.; Alqahtani, A.; Alqahtani, N.; Hassan, A.S.; Ashqar, H.I.; AlSobeh, A.M.; Hassan, A.A.; Elhenawy, M.A. Generic and Extendable Framework for Benchmarking and Assessing the Change Detection Models. *Preprints* **2024**, 2024031106. [CrossRef]
7. Muthalagu, R.; Bolimera, A.S.; Duseja, D.; Fernandes, S. Object and Lane detection technique for autonomous car using machine learning approach. *Transp. Telecommun. J.* **2021**, *22*, 383–391. [CrossRef]
8. Wang, L.; Sun, P.; Xie, M.; Ma, S.; Li, B.; Shi, Y.; Su, Q. Advanced driver-assistance system (ADAS) for intelligent transportation based on the recognition of traffic cones. *Adv. Civ. Eng.* **2020**, *2020*, 8883639. [CrossRef]
9. Lewis, A. Multimodal Large Language Models for Inclusive Collaboration Learning Tasks. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, Seattle, DC, USA, 10–15 July 2022. [CrossRef]
10. Radwan, A.; Amarneh, M.; Alawneh, H.; Ashqar, H.I.; AlSobeh, A.; Magableh, A.A.A.R. Predictive analytics in mental health leveraging llm embeddings and machine learning models for social media analysis. *Int. J. Web Serv. Res. (IJWSR)* **2024**, *21*, 1–22. [CrossRef]
11. Wang, S.; Zhao, Z.; Ouyang, X.; Wang, Q.; Shen, D. ChatCAD: Interactive computer-aided diagnosis on medical image using large language models. *arXiv* **2023**, arXiv:2302.07257. [CrossRef]

12. Gao, P. Research on Grid Inspection Technology Based on General Knowledge Enhanced Multimodal Large Language Models. In Proceedings of the Twelfth International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2023), Wuhan, China, 10–12 November 2023. [CrossRef]
13. Jaradat, S.; Alhadidi, T.I.; Ashqar, H.I.; Hossain, A.; Elhenawy, M. Exploring Traffic Crash Narratives in Jordan Using Text Mining Analytics. *arXiv* **2024**, arXiv:2406.09438.
14. Tami, M.; Ashqar, H.I.; Elhenawy, M. Automated Question Generation for Science Tests in Arabic Language Using NLP Techniques. *arXiv* **2024**, arXiv:2406.08520.
15. Alhadidi, T.; Jaber, A.; Jaradat, S.; Ashqar, H.I.; Elhenawy, M. Object Detection using Oriented Window Learning Vi-sion Transformer: Roadway Assets Recognition. *arXiv* **2024**, arXiv:2406.10712.
16. Ren, Y.; Chen, Y.; Liu, S.; Wang, B.; Yu, H.; Cui, Z. TPLLM: A traffic prediction framework based on pretrained large language models. *arXiv* **2024**, arXiv:2403.02221.
17. Zhou, X.; Liu, M.; Zagar, B.L.; Yurtsever, E.; Knoll, A.C. Vision language models in autonomous driving and intelligent transportation systems. *arXiv* **2023**, arXiv:2310.14414.
18. Zhang, Z.; Sun, Y.; Wang, Z.; Nie, Y.; Ma, X.; Sun, P.; Li, R. Large language models for mobility in transportation systems: A survey on forecasting tasks. *arXiv* **2024**, arXiv:2405.02357.
19. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intell. Transp. Syst. Mag.* **2024**, *16*, 81–94. [CrossRef]
20. Sha, H.; Mu, Y.; Jiang, Y.; Chen, L.; Xu, C.; Luo, P.; Li, S.E.; Tomizuka, M.; Zhan, W.; Ding, M. Languagempc: Large language models as decision makers for autonomous driving. *arXiv* **2023**, arXiv:2310.03026.
21. Voronin, V.; Zhdanova, M.; Gapon, N.; Alepko, A.; Zelensky, A.A.; Semenishchev, E.A. Deep Visible and Thermal Image Fusion for Enhancement Visibility for Surveillance Application. In Proceedings of the SPIE Security + Defence, Berlin, Germany, 5–8 September 2022. [CrossRef]
22. Chen, W.; Hu, H.; Chen, X.; Verga, P.; Cohen, W.W. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering Over Images and Text. *arXiv* **2022**, arXiv:2210.02928. [CrossRef]
23. Li, G.; Wang, Y.; Li, Z.; Zhang, X.; Zeng, D. RGB-T Semantic Segmentation with Location, Activation, and Sharpening. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 1223–1235. [CrossRef]
24. Morales, J.; Vázquez-Martín, R.; Mandow, A.; Morilla-Cabello, D.; García-Cerezo, A. The UMA-SAR Dataset: Multimodal Data Collection from a Ground Vehicle During Outdoor Disaster Response Training Exercises. *Int. J. Rob. Res.* **2021**, *40*, 835–847. [CrossRef]
25. Kütük, Z.; Algan, G. Semantic Segmentation for Thermal Images: A Comparative Survey. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 19–20 June 2022.
26. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038. [CrossRef]
27. Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; Zhao, H. Futr3d: A unified sensor fusion framework for 3d detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–23 June 2023; pp. 172–181.
28. Selvia, N.; Ashour, K.; Mohamed, R.; Essam, H.; Emad, D.; Elhenawy, M.; Ashqar, H.I.; Hassan, A.A.; Alhadidi, T.I. Advancing roadway sign detection with yolo models and transfer learning. In Proceedings of the IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 13–14 April 2024; pp. 1–4.
29. Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y.K.; Li, Z.; Zhao, H. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8186–8193. [CrossRef]
30. Jaradat, S.; Nayak, R.; Paz, A.; Ashqar, H.I.; Elhenawy, M. Multitask Learning for Crash Analysis: A Fine-Tuned LLM Framework Using Twitter Data. *Smart Cities* **2024**, *7*, 2422–2465. [CrossRef]
31. Elhenawy, M.; Abutahoun, A.; Alhadidi, T.I.; Jaber, A.; Ashqar, H.I.; Jaradat, S.; Abdelhay, A.; Glaser, S.; Rakotonirainy, A. Visual Reasoning and Multi-Agent Approach in Multimodal Large Language Models (MLLMs): Solving TSP and mTSP Combinatorial Challenges. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1894–1920. [CrossRef]
32. Masri, S.; Ashqar, H.I.; Elhenawy, M. Leveraging Large Language Models (LLMs) for Traffic Management at Urban Intersections: The Case of Mixed Traffic Scenarios. *arXiv* **2024**, arXiv:2408.00948.
33. Tami, M.A.; Ashqar, H.I.; Elhenawy, M.; Glaser, S.; Rakotonirainy, A. Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. *Vehicles* **2024**, *6*, 1571–1590. [CrossRef]
34. Tan, H.; Bansal, M. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv* **2020**, arXiv:2010.06775.
35. Su, L.; Duan, N.; Cui, E.; Ji, L.; Wu, C.; Luo, H.; Liu, Y.; Zhong, M.; Bharti, T.; Sacheti, A. GEM: A general evaluation benchmark for multimodal tasks. *arXiv* **2021**, arXiv:2106.09889.
36. Brauers, J.; Schulte, N.; Aach, T. Multispectral filter-wheel cameras: Geometric distortion model and compensation algorithms. *IEEE Trans. Image Process.* **2008**, *17*, 2368–2380. [CrossRef]
37. Yang, L.; Ma, R.; Zakhor, A. Drone object detection using rgb/ir fusion. *arXiv* **2022**, arXiv:2201.03786. [CrossRef]
38. Reithmeier, L.; Krauss, O.; Zwettler, A.G. Transfer Learning and Hyperparameter Optimization for Instance Segmentation with RGB-D Images in Reflective Elevator Environments. In Proceedings of the WSCG'2021–29. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2021, Plzen, Czech Republic, 17–20 May 2021.

39. Shinmura, F.; Deguchi, D.; Ide, I.; Murase, H.; Fujiyoshi, H. Estimation of Human Orientation using Coaxial RGB-Depth Images. In Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISAPP-2015), Berlin, Germany, 11–14 March 2015; pp. 113–120.
40. Yamakawa, T.; Fukano, K.; Onodera, R.; Masuda, H. Refinement of colored mobile mapping data using intensity images. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 167–173. [CrossRef]
41. Rosalina, N.H. An approach of securing data using combined cryptography and steganography. *Int. J. Math. Sci. Comput. (IJMSC)* **2020**, *6*, 1–9.
42. Dale, R. GPT-3: What's it good for? *Nat. Lang. Eng.* **2021**, *27*, 113–118. [CrossRef]
43. Kenton, J.D.M.-W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the naacL-HLT, Minneapolis, MN, USA, 2–7 June 2019; p. 2.
44. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
45. Sanderson, K. GPT-4 is here: What scientists think. *Nature* **2023**, *615*, 773. [CrossRef]
46. Krišto, M.; Ivašić-Kos, M. Thermal imaging dataset for person detection. In Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 20–24 May 2019; pp. 1126–1131.