30 January 2025, Preprint v1 · CC-BY 4.0

Research Article

Are Vision-Language Models Truly Understanding Multi-vision Sensor?

Sangyun Chung[1], **Youngjoon Yu**[1], **Youngchae Chee**[1], **Se Yeon Kim**[1], **Byung-Kwan Lee**[1], **Yong Man Ro**[1]

1. Integrated Vision Language Lab, Korea Advanced Institute of Science and Technology, Korea, Republic of

Large-scale Vision-Language Models (VLMs) have advanced by aligning vision inputs with text, significantly improving performance in computer vision tasks. Moreover, for VLMs to be effectively utilized in real-world applications, an understanding of diverse multi-vision sensor data, such as thermal, depth, and X-ray information, is essential. However, we find that current VLMs process multi-vision sensor images without deep understanding of sensor information, disregarding each sensor's unique physical properties. This limitation restricts their capacity to interpret and respond to complex questions requiring multi-vision sensor reasoning. To address this, we propose a novel Multi-vision Sensor Perception and Reasoning (MS-PR) benchmark, assessing VLMs on their capacity for sensor-specific reasoning. Moreover, we introduce Diverse Negative Attributes (DNA) optimization to enable VLMs to perform deep reasoning on multi-vision sensor tasks, helping to bridge the core information gap between images and sensor data. Extensive experimental results validate that the proposed DNA method can significantly improve the multi-vision sensor reasoning for VLMs. Codes and data are available at https://github.com/top-yun/MS-PR

Corresponding author: Yong Man Ro, ymro@kaist.ac.kr

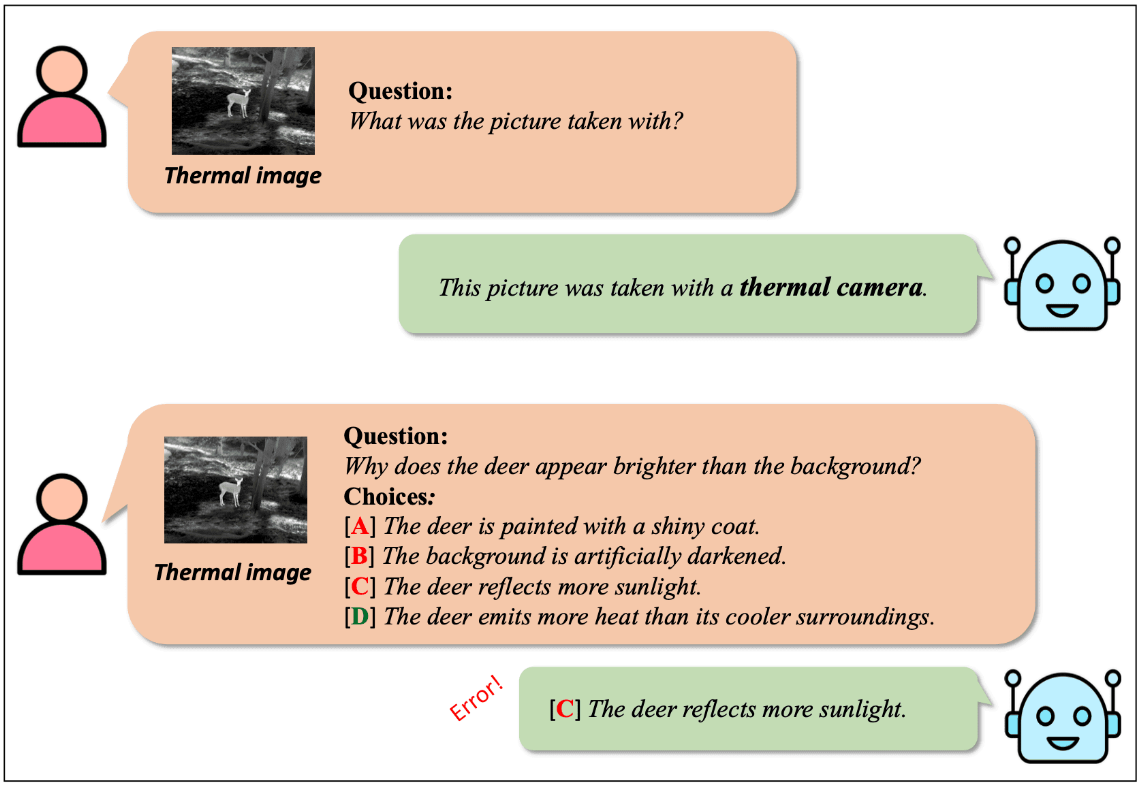
1. Introduction

In recent days, large-scale Vision-Language Models (VLMs) have made strides in areas like visual dialogue[1], video analysis[2], and document understanding[3], establishing themselves as valuable tools in the pursuit of artificial general intelligence (AGI). These models, similar to the human brain, process multi-sensor information to generate complex inferences. For instance, VLMs like OpenAI's GPT-4o[4] exhibit reasoning abilities that not only rival but sometimes even exceed human performance.

VLMs are currently reaching into applications in the real world, such as autonomous vehicles[5][6][7], Internet of Things (IoT) devices[8][9][10], and robotics[11][12][13][14][15]. Devices that connect to the real world often use multi-vision sensors, making it essential for VLMs to understand these kinds of information. Multi-vision sensors, such as thermal imaging, depth sensing, and X-ray detection, provide information that goes beyond human eyesight, enriching the understanding of real-world environments.

doi.org/10.32388/VN42C2

While humans can interpret multi-vision sensor images easily based on contextual knowledge of physical characteristics, we find that VLMs face significant challenges with multi-vision sensor data. Figure 1 demonstrates two different examples of interactions between humans and VLMs[16][17]. The first interaction shows that VLMs can easily recognize and correctly identify the type of sensor. However, in the second example, VLMs fail to select the correct answer when faced with the challenging question that requires deeper understanding and multi-vision sensor reasoning. As illustrated in Figure 1, even without direct experience, humans can understand a thermal image by integrating scientific and contextual knowledge, allowing them to interpret aspects like heat distribution. In contrast, VLMs confuse the brightness of thermals images for sunlight reflection, instead of heat emission.



Question:
What was the picture taken with?

This picture was taken with a ***thermal camera****.*

Question:
Why does the deer appear brighter than the background?
Choices:
[A] *The deer is painted with a shiny coat.*
[B] *The background is artificially darkened.*
[C] *The deer reflects more sunlight.*
[D] *The deer emits more heat than its cooler surroundings.*

[C] *The deer reflects more sunlight.*

Figure 1. Multi-vision sensor related question and response examples of recent VLMs[16][17]. Note that, this example underscores the difficulty that VLMs face in understanding physical properties unique to multi-vision sensors.

We hypothesize that it primarily arises because VLMs are predominantly trained on RGB images, making it difficult to effectively align each multi-vision sensor's unique physical properties with perceptual information. The scarcity of multi-vision sensor data further contributes to this significant problem. In other words, VLMs tend to make superficial judgments based on prior visible information derived from RGB image data. This limits their ability to understand sensor-specific details, leading to superficial RGB-bounded reasoning without genuine multi-vision sensor understanding. For instance, VLMs often confuse certain characteristics in multi-vision sensor data, such as confusing light scattering and glow effects in thermal images or mistaking fog and haze in depth images. VLMs might rely on the patterns learned from similar-looking RGB images rather than understanding the actual physical properties of the multi-vision sensors. This limitation severely affects applications where sensor-specific accuracy is crucial, such as autonomous driving[5][6], security systems[18], and medical image diagnosis[19][10].

In this paper, to handle the aforementioned challenge, we design a novel benchmark called the Multi-vision Sensor Perception and Reasoning (MS-PR) for evaluating multi-vision sensor reasoning in VLMs. MS-PR benchmark consists of multi-vision perception task and multi-vision reasoning task. Multi-vision perception refers to the task required to assess a VLM's effectiveness in meeting visual perception demands. Multi-vision reasoning measures the VLM's ability to base its responses on fundamental information from the provided sensor knowledge. To enhance the understanding capability of multi-vision sensor in VLMs, we also propose a novel Diverse Negative Attributes (DNA) optimization. By leveraging diverse negative examples, DNA improves learning in sensor-specific contexts, essential when data is scarce. A range of diverse negatives is incorporated into the optimization process, acting as stone bridges in the VLM's reasoning process and pushing it beyond the naive RGB-bounded assumptions. The evaluation of MS-PR benchmark demonstrates that most state-of-the-art VLMs display deficiencies in sensor reasoning to varying extents. Moreover, VLMs with the proposed DNA optimization show a significant increase of performance on multi-vision sensor reasoning task. In summary, the key contributions of this work are as follows:

- We first identify the limitations of current Vision-Language Models (VLMs) in multi-vision sensor reasoning. To address this issue, we propose new Multi-vision Sensor Perception and Reasoning (MS-PR) benchmark, providing a structured framework to rigorously assess VLMs' multi-vision sensor reasoning capabilities.
- We propose a novel training method, Diverse Negative Attributes (DNA) optimization, which enhances deep sensor understanding even with limited data. This approach can be applied to any large-scale VLM without altering the model architecture.
- We evaluate a total of 10 state-of-the-art VLMs using our MS-PR benchmark. Also, the extensive experimental results validates that the proposed DNA optimization can significantly improve the multi-vision sensor reasoning ability among VLMs.

2. Related work

Large-scale Vison Language Models.

Recently, there has been significant interest in visual language multimodal learning. Visual language models such as LLAVA[20][21], BLIP-2[22], InternVL2[16], VideoLLaMA2[23], MiniCPMv2.5[17], Qwen2-VL[24] have shown impressive performance in a variety of downstream tasks. In addition, to obtain richer contextual information, VLMs have developed the capability to handle multi-vision sensor inputs. For example, InternVL2[16] is an open-source multi-modal large language model that bridges the gap between open-source and commercial models by enhancing visual understanding, dynamic high-resolution processing, and bilingual dataset quality. Consequently,[25] presents ImageBind, which creates a joint embedding space across multi-vision sensors including depth and thermal sensor data. PandaGPT[26] is a VLM that integrates multi-modal encoders and large language models to enable multi-modal instruction-following capabilities, performing complex tasks. However, relatively less attention has been devoted to whether VLMs genuinely understand the physical meanings of multi-vision sensors.

Evaluation Benchmark for VLMs

Numerous studies have leveraged existing vision-language datasets to develop benchmarks for assessing the reliability of VLMs. MME[27] includes 14 sub-tasks based on publicly available images with manually created annotations, evaluating both the recognition and perception capabilities of VLMs through yes/no question answering. SEED-benchmark[28] designed to evaluate the generative comprehension capabilities of multimodal VLM through human-annotated multi-choice questions across 12 evaluation dimensions. Other comparable benchmarks include MMMU[29], Q-Bench[30], and MMBench[31]. Unlike those previous evaluation benchmarks, the proposed MS-PR is designed to rigorously test and evaluate the capabilities of understanding the physical meaning of multi-vision sensors.

3. Multi-Vision Sensor Perception and Reasoning (MS-PR) Benchmark

Model	Vision Sensors	Existence	Count	Position	General Description	Multi-vision Perception	Contextual Reasoning	Sensory Reasoning	Multi-vision Reasoning
Open Source Large-scale Vision-Language Models									
BLIP-2[22]	Thermal	59.2	32.2	57.8	65.3	53.6	74.8	42.7	58.7
	Depth	60.4	40.0	52.4	71.6	56.1	71.3	26.3	48.8
	X-ray	65.2	55.0	58.6	81.7	65.1	75.8	59.3	67.5
LLaVA-1.5-7B[32]	Thermal	60.7	27.6	65.6	60.7	53.7	74.4	41.1	57.8
	Depth	73.6	22.1	61.0	77.6	58.6	73.0	22.1	47.5
	X-ray	63.2	35.3	54.6	75.0	57.0	73.9	49.6	61.7
InternVL2-8B[16]	Thermal	66.7	47.7	70.3	73.0	64.4	74.8	50.4	60.6
	Depth	71.2	40.5	67.2	77.6	64.1	68.8	28.7	48.7
	X-ray	69.5	39.8	64.9	82.8	64.3	75.6	65.0	70.3
VideoLLaMA2-7B[23]	Thermal	82.4	49.8	69.5	81.7	70.8	83.8	76.2	80.0
	Depth	82.2	40.5	66.9	83.5	68.3	77.9	29.9	53.9
	X-ray	70.2	49.0	60.2	85.7	66.2	80.6	72.9	76.7
MiniCPM-V-2.5-8B[17]	Thermal	76.1	52.8	72.7	77.8	69.8	80.9	59.8	70.4
	Depth	76.8	43.7	71.6	84.7	69.2	77.4	51.3	64.3
	X-ray	75.2	51.0	72.1	85.3	70.9	85.7	81.6	83.7
Qwen2-VL-7B[24]	Thermal	76.1	47.7	72.7	77.6	68.5	70.6	62.8	66.7
	Depth	75.1	38.4	64.1	81.6	64.8	65.0	19.3	42.1
	X-ray	71.0	39.8	63.8	84.4	64.7	76.0	64.4	70.2
Phantom-7B[33]	Thermal	71.1	46.3	75.0	72.7	66.3	77.4	50.6	64.0
	Depth	67.8	36.3	68.1	76.6	62.2	66.9	29.6	48.2
	X-ray	69.9	44.6	64.1	82.4	65.3	76.8	67.6	72.2
Closed Source Large-scale Vision-Language Models									
Gemini-Pro[34]	Thermal	81.8	57.3	79.7	80.7	74.9	84.5	68.7	76.6

Model	Vision Sensors	Existence	Count	Position	General Description	Multi-vision Perception	Contextual Reasoning	Sensory Reasoning	Multi-vision Reasoning
	Depth	82.1	38.4	73.7	86.6	70.2	78.2	32.5	55.3
	X-ray	76.7	49.4	66.5	89.8	70.6	86.9	76.2	81.5
GPT-4o[4]	Thermal	79.3	55.3	78.9	84.4	74.5	90.6	69.7	80.2
	Depth	84.9	45.8	73.2	90.2	73.5	85.0	33.6	59.3
	X-ray	78.2	41.0	72.5	90.6	70.6	85.5	79.3	82.4
Claude-3.5-Sonnet[35]	Thermal	75.3	46.2	64.1	67.8	63.3	65.4	64.4	64.9
	Depth	63.3	30.5	52.3	73.0	54.8	53.8	44.5	49.1
	X-ray	66.8	33.1	68.1	82.4	62.6	76.9	72.9	74.9

Table 1. Evaluation results of different VLMs on the MS-PR benchmark are reported, using accuracy as the metric. "Multi-vision Perception" shows the average performance on four dimensions (Existence, Count, Position, and General Description) for evaluating visual perception, and "Multi-vision Reasoning" shows the average performance on two dimensions (Contextual Reasoning and Sensory Reasoning) for evaluating vision sensory understanding. VLMs are sorted in ascending order of release date.

3.1. Evaluation on Multi-vision Sensor Tasks

Our benchmark dataset was collected according to two multi-vision tasks: multi-vision perception and multi-vision reasoning. As illustrated in Figure 2, multi-vision perception focuses on the VLM's ability to accurately interpret and identify objects, scenes, and relationships from various multi-vision inputs. This involves tasks such as object detection, image classification, scene recognition, and relationship detection, where the model must process and understand the content of images from multiple vision sensors. The goal is to ensure that the model can consistently recognize and categorize visual elements across different contexts from different vision sensors. On the other hand, multi-vision reasoning requires the model to not only perceive but also make inferences based on the multi-vision sensor data. This involves higher-order cognitive tasks such as understanding relationships between objects, prediction of intent of sensor use, and understanding sensor knowledge. Multi-vision reasoning tests the VLM's capability to integrate multi-vision information with contextual sensor knowledge, making logical deductions that go beyond mere perception.

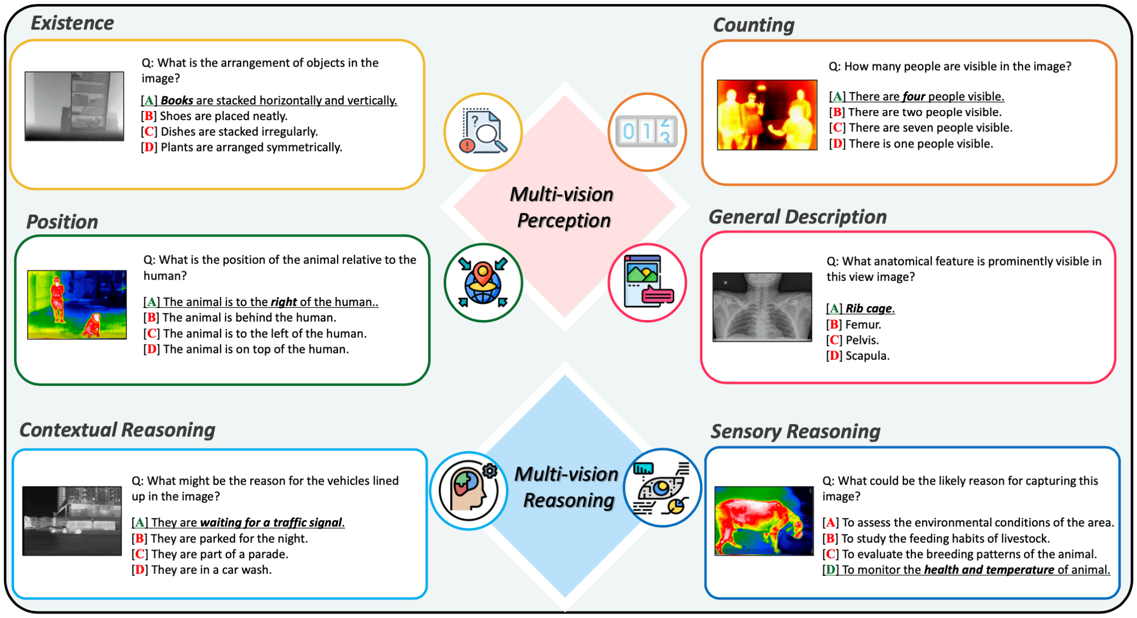

Existence
Q: What is the arrangement of objects in the image?
[A] Books are stacked horizontally and vertically.
[B] Shoes are placed neatly.
[C] Dishes are stacked irregularly.
[D] Plants are arranged symmetrically.
Counting
Q: How many people are visible in the image?
[A] There are four people visible.
[B] There are two people visible.
[C] There are seven people visible.
[D] There is one people visible.
Multi-vision Perception
Position
Q: What is the position of the animal relative to the human?
[A] The animal is to the right of the human..
[B] The animal is behind the human.
[C] The animal is to the left of the human.
[D] The animal is on top of the human.
General Description
Q: What anatomical feature is prominently visible in this view image?
[A] Rib cage.
[B] Femur.
[C] Pelvis.
[D] Scapula.
Contextual Reasoning
Q: What might be the reason for the vehicles lined up in the image?
[A] They are waiting for a traffic signal.
[B] They are parked for the night.
[C] They are part of a parade.
[D] They are in a car wash.
Multi-vision Reasoning
Sensory Reasoning
Q: What could be the likely reason for capturing this image?
[A] To assess the environmental conditions of the area.
[B] To study the feeding habits of livestock.
[C] To evaluate the breeding patterns of the animal.
[D] To monitor the health and temperature of animal.

Figure 2. Data samples of Multi-vision Sensor Perception and Reasoning (MS-PR) benchmark for evaluating the abilities of VLMs in multi-vision sensor understanding, which covers four types of multi-vision perception tasks (Existence, Counting, Position, and General Description) and two types of multi-vision reasoning tasks (Contextual Reasoning and Sensory Reasoning).

3.1.1. Multi-vision Perception

Multi-vision perception is the foundational process by which large-scale Vision-Language Models (VLMs) analyze images captured by various multi-vision sensors, including thermal, depth, and X-ray images. This process involves recognizing and interpreting the fundamental elements within each visual input based on cognitive science[36][37]. In this context, multi-vision perception tasks include (1) Existence: the ability to identify and list common objects present in the image, such as people, vehicles, animals, and so on. (2) Count: the ability to count the number of identified objects or entities. (3) Position: the ability to determine the spatial arrangement of objects within the image, noting their positions relative to one another. (4) General Description: the ability to generate nuanced descriptions of the overall scene depicted in an image. VLMs can articulate what is happening, identify objects, and provide factual information that enhances the understanding of the image itself. At the perception stage, VLMs focus on extracting essential information directly from raw image data captured by multi-vision sensors. This foundational perception is critical for all subsequent reasoning tasks, serving as the foundation upon which more complex interpretations are built.

3.1.2. Multi-vision Reasoning

Multi-vision reasoning is where VLMs truly showcase their advanced capabilities. Beyond simply perceiving images, VLMs can engage in logical reasoning to derive deeper insights and make informed decisions. This distinguishes recent VLMs, which primarily focus on understanding and interacting with the real world, from traditional computer vision models. Multi-vision reasoning tasks include (1) Contextual reasoning: the ability to utilize fundamental knowledge and contextual clues to make judgments about a given scenario. This type of reasoning allows VLMs to ensure that the reasoning process remains consistent with the context provided by the image. (2) Sensory reasoning: a more complex reasoning ability to map 2D image data to the physical meanings associated with different multi-vision sensors. This process not only involves processing the raw data from images but also integrates it with a specific information about the underlying physical sensor knowledge in the real world. Sensory reasoning requires a deep understanding of the knowledge underlying the physical meaning of multi-vision sensor data. This approach goes beyond surface-level naive image recognition, demanding that VLMs make sense of the sensor data in a way that accurately reflects real-world environments.

3.2. Evaluation Benchmark Design

Our benchmark aims at evaluating the multi-vision sensor understanding capability of large-scale VLMs. We filtered images according to six tasks in Figure 3 to improve question quality, excluding low-resolution or sequentially captured images. According to Figure 4, we begin by curating a collection of detailed question set involving multi-vision sensor inputs that guide VLMs to interpret image information. To fully understand the multi-vision sensor, ChatGPT/GPT-4o is then used to generate challenging question and answer sets based on the sensor knowledge and task prompts. By doing this, each sensor type provides distinct knowledge relevant to specific sensor properties. Also, the model can produce challenging questions that require multi-hop reasoning and deep understanding based on the specific characteristics of each sensor type with targeted tasks. Human annotators thoroughly review and refine the question and answer set. Positive answer set provides accurate answers based on sensor-specific information. While negative answer set includes plausible but incorrect responses. Each question and answer pair is structured as a chain-of-thought instruction, simulating human reasoning and directing VLMs to focus on relevant details at each step.

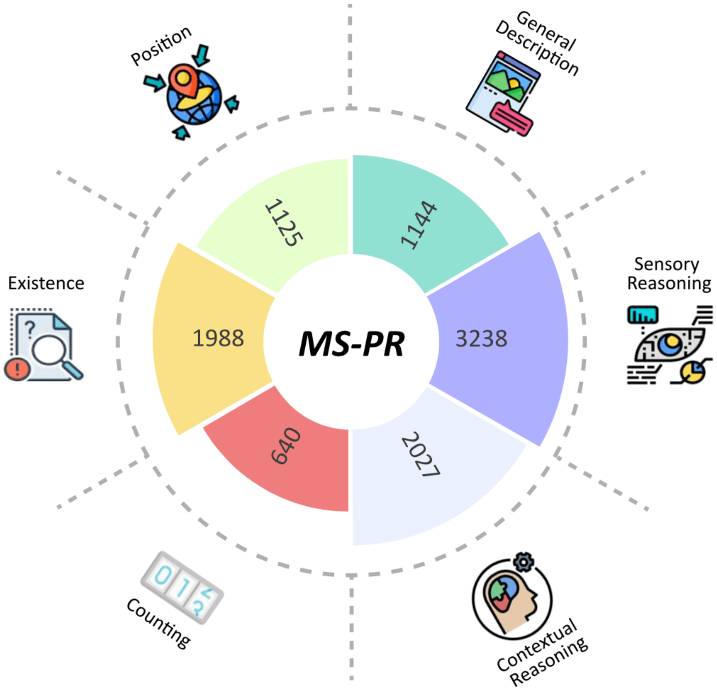
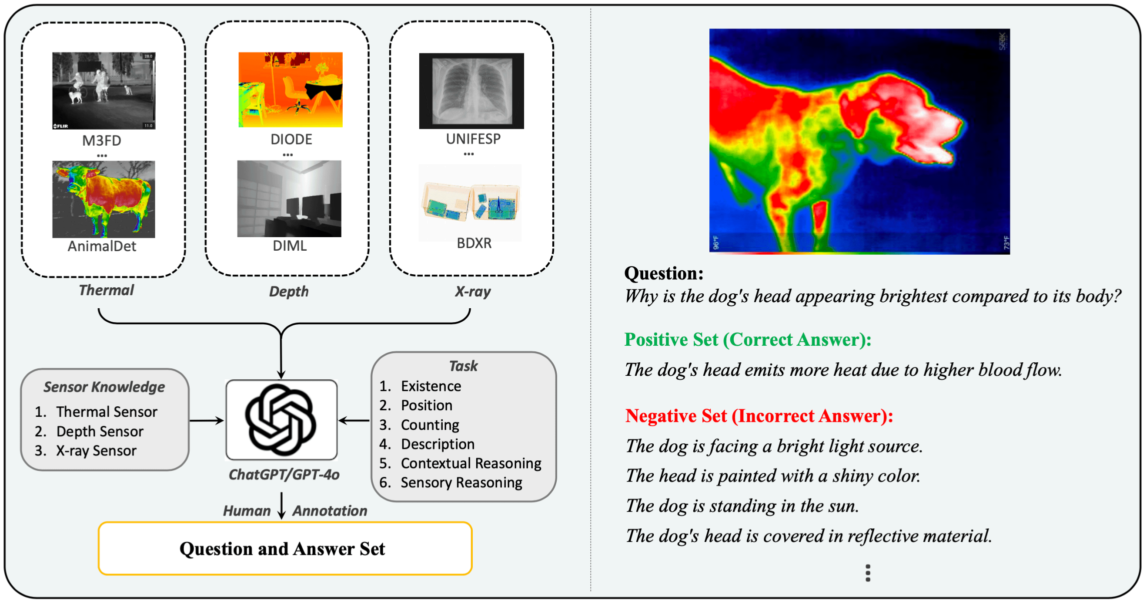


Figure 3. Distribution of data sources of the MS-PR benchmark. In MS-PR, we demonstrate six core multi-vision sensor tasks in the outer ring, and the inner ring displays the number of samples for each specific task.

doi.org/10.32388/VN42C2



Question:
Why is the dog's head appearing brightest compared to its body?

Positive Set (Correct Answer):
The dog's head emits more heat due to higher blood flow.

Negative Set (Incorrect Answer):
The dog is facing a bright light source.
The head is painted with a shiny color.
The dog is standing in the sun.
The dog's head is covered in reflective material.

Figure 4. Overview of the pipeline for generating the proposed benchmark dataset. Based on the prompts corresponding to knowledge on multi-vision sensors and tasks, ChatGPT/GPT-4o generates challenging question and answer set. We refine the dataset further by utilizing human annotators to construct positive and negative sets, allowing each pair to be classified into a specific evaluation dimension.

Model	Sensor Type	Existence	Count	Position	General Description	Multi-vision Perception	Contextual Reasoning	Sensory Reasoning	Multi-vision Reasoning
Phantom-7B	Thermal	71.1	46.3	75.0	72.7	66.3	77.4	50.6	64.0
	Depth	67.8	36.3	68.1	76.6	62.2	66.9	29.6	48.2
	X-ray	69.9	44.6	64.1	82.4	65.3	76.8	67.6	72.2
Phantom-7B + SFT	Thermal	82.8	46.2	73.4	81.7	71.0	79.3	78.5	78.9
	Depth	71.0	48.4	71.7	84.7	69.0	77.1	65.3	71.2
	X-ray	73.5	47.4	67.7	82.0	67.7	78.3	73.5	75.9
Phantom-7B + DNA	Thermal	**86.8**	**49.8**	**75.8**	**86.4**	**74.3**	**82.9**	**86.4**	**84.6**
	Depth	**79.1**	**49.0**	**74.5**	**87.9**	**72.6**	**81.2**	**86.1**	**83.7**
	X-ray	**78.2**	**49.4**	**73.3**	**84.8**	**71.4**	**85.8**	**82.1**	**84.0**
Qwen2-VL-7B	Thermal	76.1	47.7	72.7	77.6	68.5	70.6	62.8	66.7
	Depth	75.1	38.4	64.1	81.6	64.8	65.0	19.3	42.1
	X-ray	71.0	39.7	63.7	84.4	64.7	76.0	64.4	70.2
Qwen2-VL-7B + SFT	Thermal	85.7	50.8	80.5	82.6	74.9	85.8	80.6	83.2
	Depth	83.0	44.2	73.3	89.0	72.4	75.6	30.6	53.1
	X-ray	78.2	43.8	70.5	89.8	70.6	84.4	84.2	84.3
Qwen2-VL-7B + DNA	Thermal	**89.1**	**52.3**	**80.5**	**88.4**	**77.6**	**89.0**	**85.7**	**87.4**
	Depth	**84.4**	**44.2**	**74.8**	**90.0**	**73.3**	**80.5**	**59.8**	**70.2**
	X-ray	**79.8**	**45.0**	**73.7**	**91.4**	**72.5**	**86.4**	**86.0**	**86.2**
InternVL2-8B	Thermal	66.7	47.7	70.3	73.0	64.4	74.8	50.4	60.6
	Depth	71.2	40.5	67.2	77.6	64.1	68.8	28.7	48.7
	X-ray	69.5	39.8	64.9	82.8	64.3	75.6	65.0	70.3
InternVL2-8B + SFT	Thermal	80.8	48.8	70.3	78.7	69.6	77.0	69.4	73.2
	Depth	72.0	41.1	69.8	81.9	66.2	72.1	49.7	60.9
	X-ray	72.8	46.2	67.3	84.8	67.8	78.6	73.7	76.1
InternVL2-8B	Thermal	**84.0**	**50.3**	**75.0**	**84.5**	**73.4**	**82.9**	**82.0**	**82.4**
	Depth	**74.1**	**42.6**	**71.9**	**84.7**	**68.3**	**77.3**	**77.8**	**77.6**

Model	Sensor Type	Existence	Count	Position	General Description	Multi-vision Perception	Contextual Reasoning	Sensory Reasoning	Multi-vision Reasoning
+ DNA	X-ray	**75.2**	**47.0**	**70.9**	**85.7**	**69.7**	**83.0**	**78.1**	**80.6**

Table 2. Performance is increased by Diverse Negative Attributes (DNA) optimization in the sense of multi-vision reasoning. Highlighted columns show average performance for perception and reasoning capabilities. The best results are denoted in bold.

4. Enhancing Multi-vision Sensor Reasoning

4.1. Problems on Multi-vision Sensor Data

Through the MS-PR benchmark, we reveal that multi-vision sensor reasoning problems are widespread in current VLMs in Table 3. The primary reason is the scarcity of publicly available multi-vision sensor instruction-tuning datasets. Lacking sufficient opportunities to learn sensor knowledge, VLMs tend to misunderstand the image information. Due to this data constraint, VLMs often rely on RGB-bounded reasoning, causing them to confuse the unique characteristics of multi-vision sensor data. Considering these inherent issues in multi-vision sensor data, we propose an efficient, data-centric approach that enables effective learning even with a limited dataset. To demonstrate this, we design a method that achieves comparable performance by using only a small portion of data. We construct approximately 600 multi-vision sensor images, with 200 images for each sensor type, all of which are not included in the MS-PR evaluation benchmark.

4.2. Diverse Negative Attributes Optimization

In this paper, we propose a novel Diverse Negative Attributes (DNA) optimization. Unlike previous reinforcement learning-based methods such as Reinforcement Learning from Human Feedback (RLHF)[38], Direct Preference Optimization (DPO)[39], and Simple Preference Optimization (SimPO)[40], DNA optimization reduces the RGB-bounded reasoning during the training process by directly adding the designed loss to the supervised fine tuning process. It is the optimization process where the model identifies the correct answer while simultaneously learning to avoid being misled by confusing answers. By using various negative samples in a limited set of image-question pairs, richer knowledge can be filled. We jointly use the autoregressive supervised fine-tuning (SFT) loss as follows:

$$\min_{\theta} \mathcal{L}_{\mathrm{DNA}} = \mathcal{L}_{\mathrm{SFT}} + \mathcal{L} \tag{1}$$

$$\mathcal{L} = -\mathbb{E}_{\mathcal{D}}\mathbb{E}_{y^-}\left[\log\sigma\left(\beta\frac{\log\pi_\theta\left(y^+|x\right)}{|y^+|} - \beta\frac{\log\pi_\theta\left(y_j^-|x\right)}{\left|y_j^-\right|} - \gamma\right)\right] \tag{2}$$

Negative Sample k	Sensor Type	Existence	Count	Position	General Description	Multi-vision Perception	Contextual Reasoning	Sensory Reasoning	Multi-vision Reasoning
k = 1	Thermal	86.6	48.8	74.2	83.3	73.2	81.2	85.0	83.1
	Depth	**79.9**	48.4	73.5	87.6	72.3	80.8	80.2	80.5
	X-ray	76.8	48.2	69.7	82.8	69.4	84.2	80.5	82.3
k = 2	Thermal	**86.9**	49.3	73.4	84.2	73.4	81.6	**86.4**	84.0
	Depth	76.6	46.8	**74.9**	**88.5**	71.7	81.0	84.4	82.7
	X-ray	76.0	**49.4**	71.3	81.6	69.6	84.8	81.9	83.3
k = 3	Thermal	86.8	**49.8**	**75.8**	**86.4**	**74.3**	**82.9**	**86.4**	**84.6**
	Depth	79.1	**49.0**	74.5	87.9	**72.6**	**81.2**	**86.1**	**83.7**
	X-ray	**78.2**	**49.4**	**73.3**	**84.8**	**71.4**	**85.8**	**82.1**	**84.0**

Table 3. Ablation study on multi-vision sensor reasoning performance according to the number of negative sample k.

where θ represents the parameters to be trained, $\mathcal{L}_{\text{SFT}}$ denotes the supervised fine-tuning loss for question-answer pairs, and x stands for a specific input prompt. Here, y^+ denotes a correct answer, while $y^- = \{y_1^-, y_2^-, \cdots, y_k^-\}$ represents the set of confusing answers including k number of incorrect answers y_j^- such that $j = \{1, 2, \cdots, k\}$. The minimizing process effectively reinforce the model's tendency toward correct answers rather than confusing ones.

The proposed DNA optimization is designed for scenarios with limited training samples, effectively enhancing learning through a greater diversity of negative examples. This approach is particularly valuable for multi-vision sensor data, where data scarcity is a significant issue. By introducing diverse counterfactual negatives, VLMs gain more opportunities to learn from a small dataset. Furthermore, VLMs often misinterpret multi-vision sensor data due to an over-reliance on RGB-bounded reasoning. DNA counteracts this by creating diverse negatives that prevent the VLMs confuse with similar-looking images. Instead, it encourages deep understanding on the unique attributes of each sensor type. This approach aligns with multi-vision sensor reasoning tasks by allowing the VLMs to acquire a deeper understanding of sensor-specific contexts.

5. Experiment

5.1. Experimental Setup

5.1.1. Dataset Collection

To construct the MS-PR benchmark, we focused on assembling a dataset that captures diverse scenarios and sensor-specific information. We collected a total of 13 distinct datasets, which include 7k images that represent a wide variety of situations. From these images, we can generate approximately 10k unique questions for evaluation. The depth images include both indoor and outdoor environments and capture various objects in diverse settings. For thermal images, we collected datasets covering a broad range of different objects and scenarios, such as in-vehicle sensors, landscapes, people, animals, and thermal screening&scanning. The X-ray images include human body-part images and the security inspection of luggage in airport datasets. This collection offers a robust dataset where different multi-vision sensors are represented across a broad range of real-world scenarios. We described the overall distribution of data of the MS-PR benchmark in Figure 3. For the training dataset, we used 600 images (200 images per sensor) from the 13 datasets mentioned above that were not included in the MS-PR benchmark, generating 3,600 question-answer pairs. We focused on six problem task types requiring high level reasoning to compile the source dataset: Existence, Counting, Position, General Description, Contextual Reasoning, and Sensory Reasoning. More details about the task type and visual context of each source dataset are demonstrated in the supplementary materials.

5.1.2. Implementation Details

In our evaluation, we selected 10 state-of-the-art (SOTA) Vision-Language Models (VLMs) that represent the leading edge of the current research field. These models were chosen to provide a comprehensive assessment of the capabilities and performance of both open-source and closed-source VLMs across a variety of multi-vision sensor tasks on the MS-PR benchmark. Open source model include BLIP-2[22], LLAVA-v1.5-7B[20], InternVL2-8B[16], VideoLLaMA2-7B[23], MiniCPM-V-2.5-8B[17], Qwen2-VL-7B[24], and Phantom-7B[33]. While closed source model include GPT-4o[4], Claude 3.5 Sonnet[35], and Gemini-Pro[34]. In the DNA Optimization, we set the hyper-parameters to $\beta = 2$, $\gamma = 0.2$, and $k = 3$. Each VLM was trained using QLoRA[41], and during training, we used the AdamW optimizer[42]. For Phantom-7B[33], a learning rate of $2e - 5$ was applied, with one training epoch. All layers of the VLM utilized 256 rank and 256 alpha parameters. For Qwen2-VL-7B[24], a learning rate of $2e - 5$ was also applied, with one training epoch. All layers of the VLM utilized 64 rank and 64 alpha parameters.

Model	Number of Training Images per Sensor n	Sensor Type	Existence	Count	Position	General Description	Multi-vision Perception	Contextual Reasoning	Sensory Reasoning	Multi-vision Reasoning
Phantom-7B + SFT	n = 50	Thermal	75.7	**46.8**	73.4	74.9	67.7	77.8	56.6	67.1
		Depth	69.5	43.7	69.5	81.2	66.0	72.9	43.7	58.3
		X-ray	71.5	45.0	66.7	82.4	66.4	77.7	69.3	73.5
Phantom-7B + DNA	n = 50	Thermal	**85.7**	**46.8**	**75.0**	**81.4**	**72.2**	**81.6**	**77.5**	**79.5**
		Depth	**73.7**	**47.4**	**72.0**	**85.2**	**69.6**	**78.0**	**66.6**	**72.3**
		X-ray	**73.4**	**49.0**	**70.1**	**83.2**	**68.9**	**84.2**	**81.6**	**82.9**
Phantom-7B + SFT	n = 100	Thermal	80.9	46.3	73.4	78.5	69.8	79.3	70.1	74.7
		Depth	71.2	43.7	68.0	82.3	66.3	72.9	50.0	61.4
		X-ray	71.7	46.6	67.3	82.0	66.9	78.8	70.7	74.8
Phantom-7B + DNA	n = 100	Thermal	**86.5**	**48.2**	**75.8**	**83.0**	**73.4**	**82.2**	**83.8**	**83.0**
		Depth	**76.8**	**48.4**	**73.6**	**86.2**	**71.3**	**80.0**	**69.9**	**75.0**
		X-ray	**76.2**	**49.0**	**71.3**	**83.6**	**70.0**	**84.7**	**81.9**	**83.3**

Table 4. Ablation study on multi-vision sensor reasoning performance according to the number of training images per sensor n

Model	MMMU	MME	MMBench	Q-Bench	SEED^{I}
Phantom-7B	47.8	**2126**	79.8	**69.9**	75.3
Phantom-7B +DNA	**49.3**	2113	**80.2**	68.2	**75.7**

Table 5. Performance comparison of Phantom-7B with and without DNA optimization across various benchmarks.

5.2. Experiment Result

5.2.1. Evaluation on MS-PR Benchmark

In this section, we conduct a comprehensive evaluation using the proposed MS-PR benchmark, a rigorous framework designed to assess the capabilities of large-scale Vision-Language Models (VLMs) in two target tasks: Multi-vision Perception and Multi-vision Reasoning. Multi-vision Perception presents the averaged performance on four dimensions for evaluating visual perception. Meanwhile, Multi-vision Reasoning demonstrates the averaged performance on two dimensions for evaluating the VLMs' ability to understand and reason about multi-vision sensor data. As shown in Table 1, the evaluation revealed that performance varies significantly depending on the type of multi-vision sensor used to capture the input images. VLMs generally have moderate scores in multi-vision perception tasks, but vary significantly in multi-vision contextual and sensory reasoning. Sensory reasoning requires VLMs to not only recognize and describe images but also to understand the physical principles underlying the sensor data. For example, interpreting thermal data involves understanding heat signatures, while depth data requires an understanding of the need for spatial geometry beyond naive 2D interpretation. The experiment demonstrates VLMs' limited proficiency in interpreting sensor data to its physical meaning. We recruited human participants from the crowd sourcing platform Prolific and asked them to evaluate the proposed benchmark. The results shows significant alignment with human assessments. Details of human agreement on our benchmark can be found in the supplementary materials.

5.2.2. Evaluation on the Effects of DNA Optimization

In Table 2, we validate that our proposed Diverse Negative Attribute (DNA) optimization significantly improves the multi-vision reasoning performance in VLMs. As we already mentioned in the introduction, DNA optimization is flexible and adaptable enough so that it is applicable to other VLMs without changing the network architecture. With supervised fine-tuning (SFT), Phantom-7B[33] and Qwen2-VL-7B[24] shows slight improvements across all metrics, particularly in General Description and Contextual Reasoning. With the proposed DNA optimization, Phantom-7B[33] and Qwen2-VL-7B[24] see significant improvements in almost all metrics, especially in Multi-vision Reasoning. This demonstrates DNA optimization significantly enhances multi-vision reasoning, especially in tasks that require an understanding of sensor-specific information.

5.3. Generalizability of DNA Optimization

The proposed DNA optimization has demonstrated exceptional performance in multi-vision sensor reasoning tasks. To assess its generalization capability in general benchmark, we conduct evaluation experiments using the MMMU[29], MME[27], MMBench[31], Q-Bench[30], and SEED^{I}[28] benchmark, which encompasses various disciplines and domains. The results are shown in Table 5. DNA optimization has a comparable performance on other benchmarks. This experiment result underscores its capability to generalize to downstream VLM understanding and reasoning tasks. Furthermore, the fine-tuning process using our synthetic data does not detract from the VLMs reasoning abilities in other benchmarks; rather, it enhances its generalizability.

5.4. Ablation on the Number of Negative Sample

Table 3 presents an ablation study on multi-vision sensor reasoning performance based on the number of negative samples, denoted as k. Baseline model is Phantom-7B[33] with DNA optimization. This table demonstrates the impact of using different numbers of negative samples in DNA optimization on the performance of multi-vision sensor reasoning. As a result, increasing number of negative sample k generally improves multi-vision perception and reasoning scores, especially for Contextual and Sensory reasoning tasks, suggesting that more negative samples help VLMs better differentiate relevant features in sensor-specific contexts. In other words, using diverse negative samples can provide the most balanced and comprehensive understanding for VLMs across various multi-vision sensors.

5.5. Ablation on the Number of Training Images

Table 4 presents an ablation study analyzing multi-vision sensor reasoning performance as a function of the training image count, denoted by n. This table illustrates how adjusting the number of training images per sensor influences multi-vision sensor reasoning performance in both SFT and DNA optimization methods. Even with a limited quantity of training images, DNA optimization surpasses SFT, suggesting that DNA optimization can effectively yield results comparable to those obtained with a larger dataset.

6. Conclusion

In this study, we focus on assessing and improving the ability of large-scale Vision-Language Models (VLMs) to understand and process multi-vision sensor inputs. As VLMs are increasingly deployed in real-world applications, their ability to accurately interpret and reason about data from diverse vision sensors has become crucial. To address this, we propose a new evaluation benchmark called MS-PR, which generates samples aimed at specific physical sensor understanding. We also propose novel DNA optimization to improve the multi-vision sensor reasoning ability. Through extensive experiments, we assess the performance of understanding sensor knowledge in the latest state-of-the-art VLMs handling multi-vision input. Moreover, extensive experimental results validate that the proposed DNA optimization significantly improve the performance of multi-vision sensor reasoning in VLMs. We believe that integrating a sensor knowledge annotated evaluation benchmark and tailored optimization pave the way for promising future applications of VLMs.

Appendix A. Detailed Description on Dataset

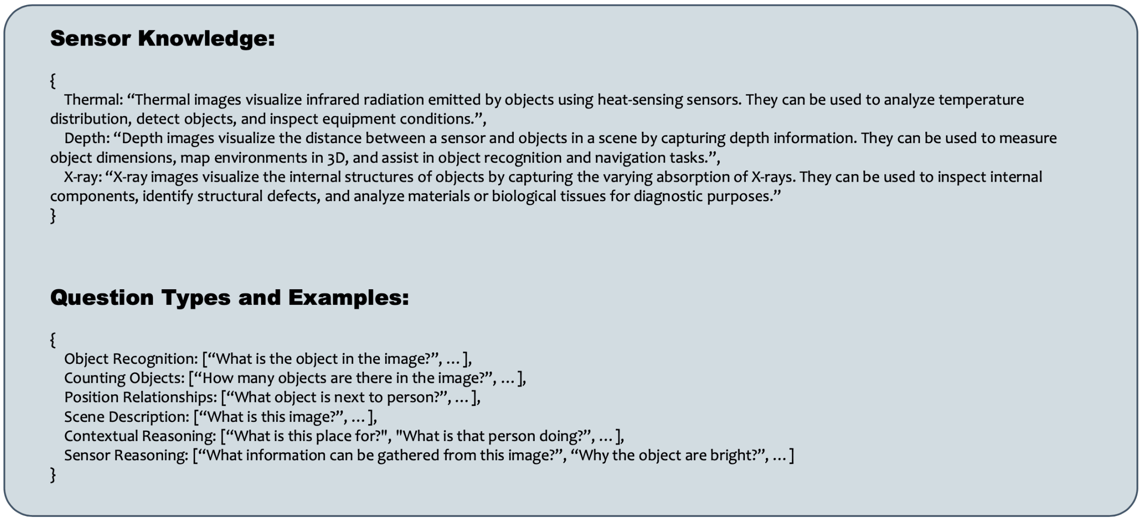
We collect 13 different datasets for each multi-sensor vision task type, together with 7k images and 10k unique questions and answers in total. To ensure the generalizability of MS-PR benchmark, we gather a wide variety of situations and object types from various different datasets. These datasets are mainly classified into three categories according to multi-vision sensors. 1) For thermal sensor datasets, we collect 2.2k images from 8 different thermal datasets, including M3FD[43], Dog&People[44], Pet[45], TCVP[46], HIT-UAV[47], AnimalDet[48], CTFD[49], and IFSOD[50]. 2) Additionally, we gather 3.1k images from 3 different datasets for depth sensor, including DIODE[51], NYUv2[52], and DIML[53]. 3) Finally, we sampled 2.6k images from the two different public X-ray datasets, including UNIFESP[54] and BDXR[55] datasets.

M3FD[43] dataset contains images from three primary scenes: road views, university campuses, and resort settings. The dataset comprises 24-bit grayscale infrared and visible images, each with a resolution of 1024×768 pixels. Thermal Dogs and People[44] dataset includes 203 thermal infrared images captured at varying distances from people and dogs in park and home environments. Images are available in both portrait and landscape orientations with a spectral color palette applied. Pet dataset[45] features 640×640 images depicting diverse activities and motions of cats, dogs, and humans. This dataset has 640×640 image size. Thermal Computer Vision Project(TCVP) dataset[46] focuses on heat detection in groups of humans, with an average image size of 640×640 pixels. A high-altitude infrared thermal dataset for object detection applications on Unmanned Aerial Vehicles(HIT-UAV)[47] is a high-altitude infrared thermal dataset for object detection applications involving unmanned aerial vehicles (UAVs). It includes 2,898 infrared images derived from 43,470 video frames captured in diverse scenarios. Animal detection(AnimalDet) dataset[48] consists of thermal images of eight animal species —deer, bear, cow, dog, elephant, fox, goat, and wild boar. The average image size is 369×363 pixels. The Chips Thermal Face Dataset[49] comprises over 1,200 thermal face images of male and female subjects aged 18–23 from three continents. It supports research in advanced thermal facial classification and recognition systems using deep learning techniques. IFSOD dataset[50] contains thermal sensor images of various objects, including bicycles, birds, dogs, and humans, with an average resolution of 640×480 pixels. A Dense Indoor and Outdoor DEpth Dataset(DIODE)[51] provides high-resolution color images paired with precise, dense, long-range depth measurements. It is the first publicly available RGBD dataset featuring both indoor and outdoor scenes captured using a single sensor suite. The NYU-Depth V2 dataset[52] includes video sequences of indoor environments captured with the RGB and depth cameras of Microsoft Kinect. It contains 1,449 densely labeled pairs of aligned RGB and depth images. Digital Image Media Laboratory(DIML)/Computer Vision Laboratory(CVL) RGB-D dataset[53] contains 2 million color images paired with depth maps, covering diverse indoor and outdoor scenes. The RGB images have a resolution of 1920×1080, while depth maps are captured at 512×424 pixels. UNIFESP X-ray Body Part dataset[54] comprises X-ray images of various body parts, such as the knee, leg, hip, ankle, thigh, and pelvis. It stands out for its diversity of human anatomical coverage. Baggage Detection X-Ray(BDXR) dataset contains X-ray images of baggage inspected at airports to ensure diversity and generalization. The average image resolution is 1225×954 pixels. We described the overall distribution of data sources of the MS-PR benchmark in Figure 3.

Appendix B. Detailed Description on Prompt

We designed the input prompts to create the proposed MS-PR benchmark, ensuring the prompts are comprehensive and tailored to extract meaningful multi-vision sensor capabilities from challenging question and answer sets. These prompts require five additional information to effectively guide the VLMs in generating benchmark data:

- To provide sufficient sensor information to the LLM, we developed Sensor Knowledge (Figure 5) and incorporated it into <sensor_knowledge>. This information contains detailed descriptions and context about thermal, depth, and X-ray sensors. This ensures the VLMs understand the unique physical properties and contextual applications of each sensor type.
- The appropriate multi-vision sensor type is included in <sensor_type>. This explicitly informs the VLMs which sensor (thermal, depth, or X-ray) the prompt relates to, ensuring that the generated examples are relevant to the specific sensor.
- The desired question type and corresponding examples are provided in <question_type> and < question_examples>, respectively (Figure 5). This ensures that the model understands the format and context of the questions it needs to generate.
- The number of negative samples to be generated is specified in <negative_samples_num>. These negative samples are designed to include plausible yet incorrect answers, encouraging the model to distinguish correct answers from distractors.



Sensor Knowledge:

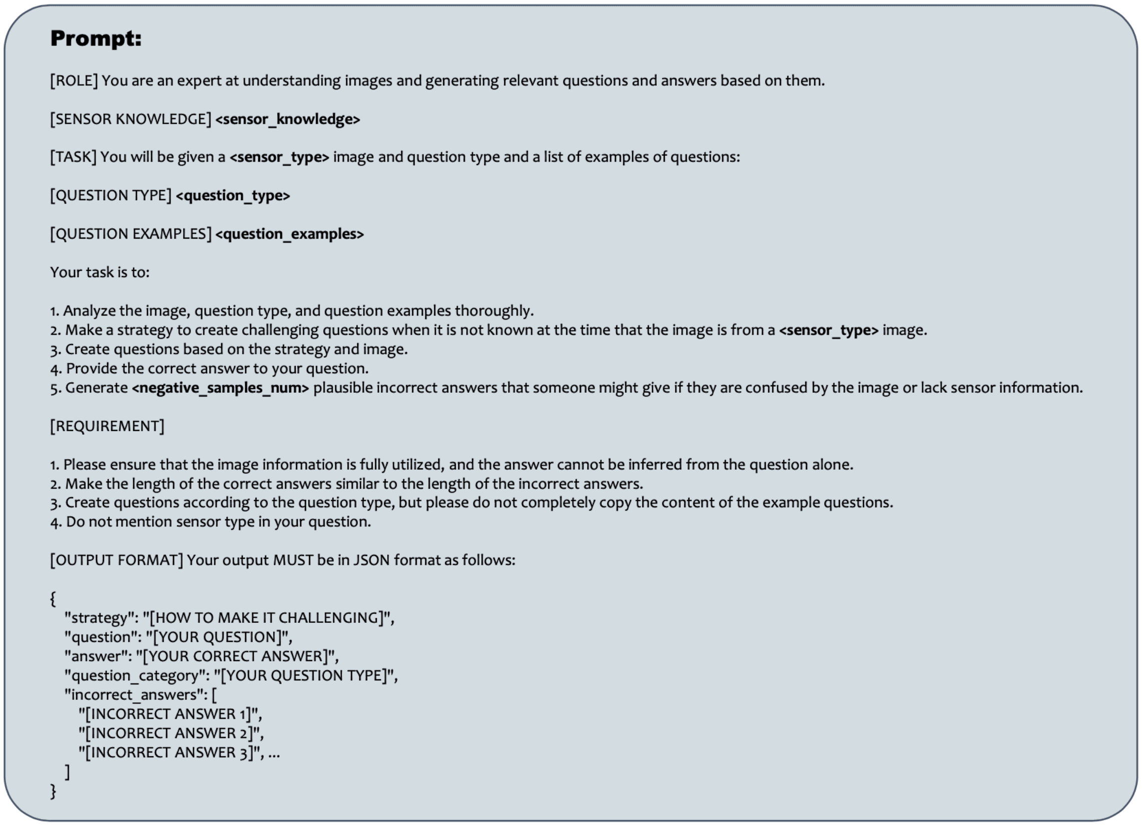
```
{
   Thermal: "Thermal images visualize infrared radiation emitted by objects using heat-sensing sensors. They can be used to analyze temperature distribution, detect objects, and inspect equipment conditions.",
   Depth: "Depth images visualize the distance between a sensor and objects in a scene by capturing depth information. They can be used to measure object dimensions, map environments in 3D, and assist in object recognition and navigation tasks.",
   X-ray: "X-ray images visualize the internal structures of objects by capturing the varying absorption of X-rays. They can be used to inspect internal components, identify structural defects, and analyze materials or biological tissues for diagnostic purposes."
}
```

Question Types and Examples:

```
{
   Object Recognition: ["What is the object in the image?", … ],
   Counting Objects: ["How many objects are there in the image?", … ],
   Position Relationships: ["What object is next to person?", … ],
   Scene Description: ["What is this image?", … ],
   Contextual Reasoning: ["What is this place for?", "What is that person doing?", … ],
   Sensor Reasoning: ["What information can be gathered from this image?", "Why the object are bright?", … ]
}
```

Figure 5. Description of sensor knowledge information and questions types and examples.

Our input prompts for generating MS-PR benchmark is described in Figure 6.



Prompt:

[ROLE] You are an expert at understanding images and generating relevant questions and answers based on them.

[SENSOR KNOWLEDGE] **<sensor_knowledge>**

[TASK] You will be given a **<sensor_type>** image and question type and a list of examples of questions:

[QUESTION TYPE] **<question_type>**

[QUESTION EXAMPLES] **<question_examples>**

Your task is to:

1. Analyze the image, question type, and question examples thoroughly.
2. Make a strategy to create challenging questions when it is not known at the time that the image is from a **<sensor_type>** image.
3. Create questions based on the strategy and image.
4. Provide the correct answer to your question.
5. Generate **<negative_samples_num>** plausible incorrect answers that someone might give if they are confused by the image or lack sensor information.

[REQUIREMENT]

1. Please ensure that the image information is fully utilized, and the answer cannot be inferred from the question alone.
2. Make the length of the correct answers similar to the length of the incorrect answers.
3. Create questions according to the question type, but please do not completely copy the content of the example questions.
4. Do not mention sensor type in your question.

[OUTPUT FORMAT] Your output MUST be in JSON format as follows:

```
{
   "strategy": "[HOW TO MAKE IT CHALLENGING]",
   "question": "[YOUR QUESTION]",
   "answer": "[YOUR CORRECT ANSWER]",
   "question_category": "[YOUR QUESTION TYPE]",
   "incorrect_answers": [
      "[INCORRECT ANSWER 1]",
      "[INCORRECT ANSWER 2]",
      "[INCORRECT ANSWER 3]", ...
   ]
}
```

Figure 6. Description of prompts for generating challenging multiple-choice questions and answers for multi-vision sensor tasks

Appendix C. Human Evaluation

We conducted a human evaluation study to assess how closely our newly proposed MS-PR benchmark aligns with the answers a human would select when viewing sensor images. A total of 20 participants were recruited through the crowd-sourcing Prolific platform. We only accepted reviewers with English as their first language and who had at least bachelor's degree. In the human study, we recruited Prolific participants with approval rates higher than 95% and with at least 200 prior submissions.

Participants were rewarded €9.4/hr for completing all multiple choice questions. We sampled 45 multi-vision reasoning questions from MS-PR benchmark, with 15 questions allocated to each sensor type: thermal, depth, and X-ray. Experiment results on human evaluation are demonstrated in Figure 7. Participants achieved a 95.1% accuracy rate, demonstrating that the proposed MSPR benchmark significantly aligns with human assessment. We also evaluated how other VLMs responded to the 45 sampled questions and verified that their performance on multi-vision reasoning, as shown in Table 1 of the main text, aligns within the margin of error. To be specific, GPT-4o[4] achieved the highest score of 73.3, followed by InternVL2-8B[16] and Phantom-7B[33], both scoring 62.2, Qwen2-VL-7B[24] with 60.0, and LLaVA-1.5-7B[20], which recorded the lowest score of 53.3. The performance difference between the top VLMs (GPT-4o) and human participants is notable at 21.8%, reflecting the challenges that current VLMs face in achieving human-level understanding in multi-vision reasoning tasks.

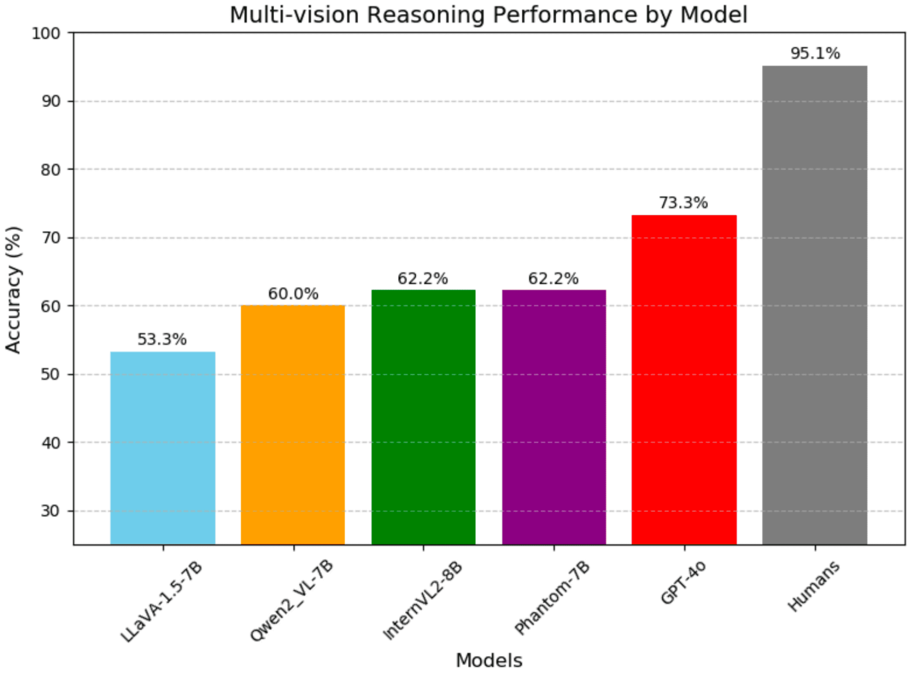


Figure 7. Human Agreement Results on Multi-Vision Sensory Reasoning Performance Across Diverse VLMs

Appendix D. Additional Question and Answer Examples

Figure 8-13 provide examples of benchmark evaluations conducted using various Vision-Language Models (LLaVA-1.5-7B[32], InternVL2-8B[16], Phantom-7B[33], and Phantom-7B with DNA optimization) across three multi-vision sensors: thermal, depth, and X-ray. The answers selected by the models are displayed next to the corresponding options using the models' representative icons and pictograms, and they are color-coded based on correctness: green indicates a correct answer, while red indicates an incorrect answer. By displaying these visual examples with clear indicators and detailed observations, we provide valuable insights into how different VLMs perform on multi-vision sensor reasoning tasks. These examples underscore the importance of tailored optimization techniques, like Diverse Negative Attributes(DNA) optimization, in enhancing the multi-vision sensor reasoning capabilities of VLMs across diverse sensor modalities.

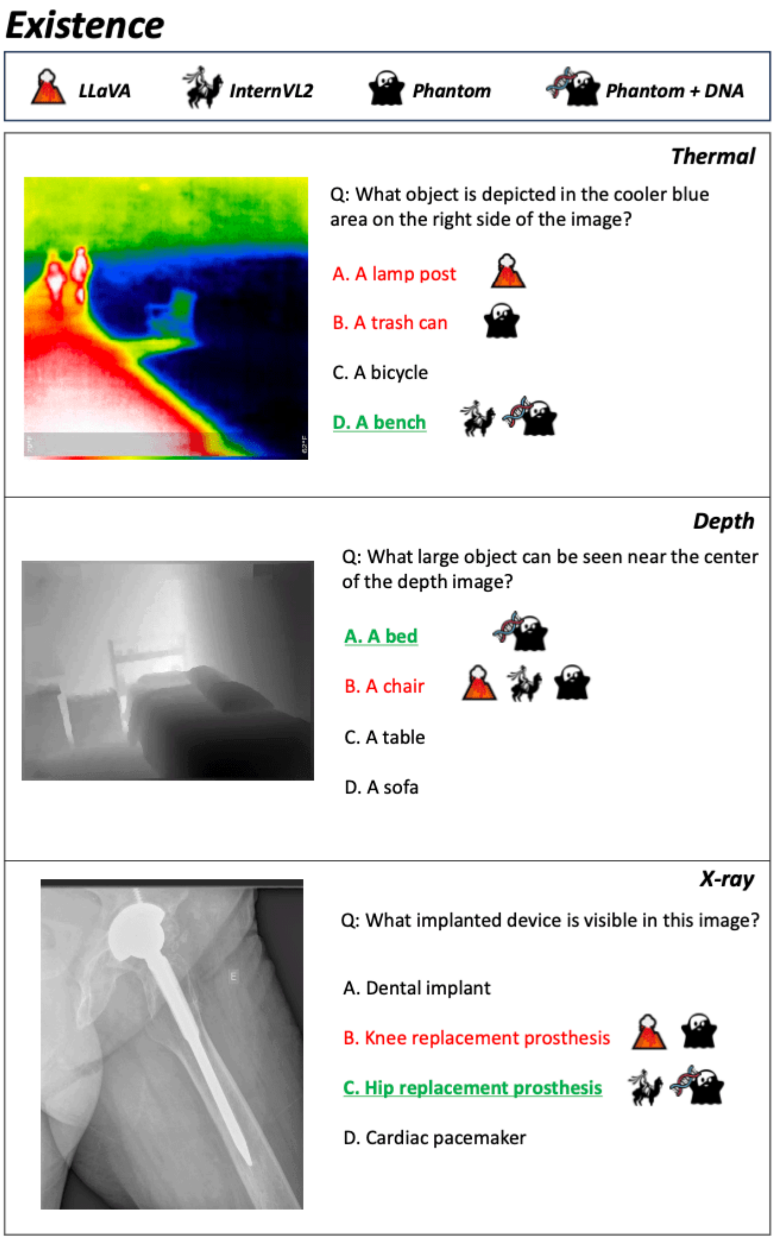
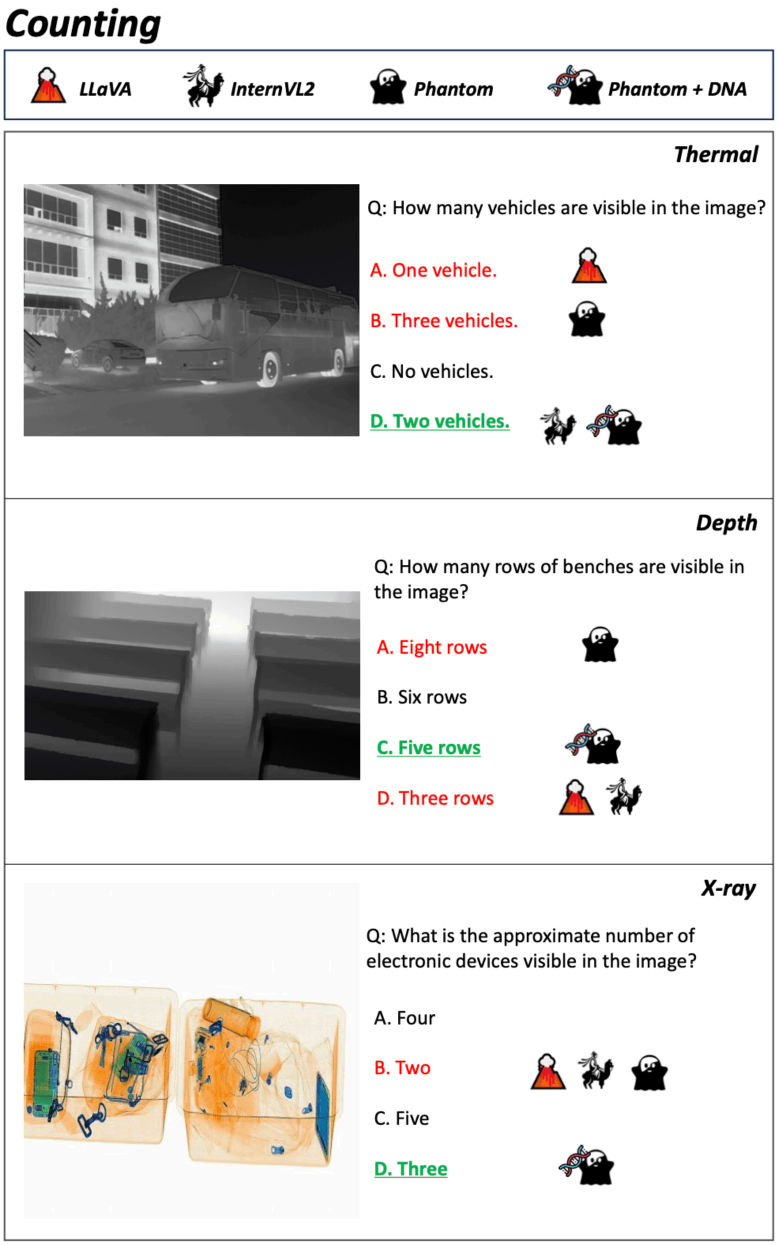

Existence
LLaVA
InternVL2
Phantom
Phantom + DNA
Thermal
Q: What object is depicted in the cooler blue area on the right side of the image?
A. A lamp post
B. A trash can
C. A bicycle
D. A bench
Depth
Q: What large object can be seen near the center of the depth image?
A. A bed
B. A chair
C. A table
D. A sofa
X-ray
Q: What implanted device is visible in this image?
A. Dental implant
B. Knee replacement prosthesis
C. Hip replacement prosthesis
D. Cardiac pacemaker

Figure 8. The comparison of performance across different multi-vision sensors with respect to the representative VLMs in the Multi-vision Perception task (Existence). Green font denotes the correct answer, while red font denotes the incorrect answer.



Counting

LLaVA InternVL2 Phantom Phantom + DNA

Thermal

Q: How many vehicles are visible in the image?

A. One vehicle.

B. Three vehicles.

C. No vehicles.

D. Two vehicles.

Depth

Q: How many rows of benches are visible in the image?

A. Eight rows

B. Six rows

C. Five rows

D. Three rows

X-ray

Q: What is the approximate number of electronic devices visible in the image?

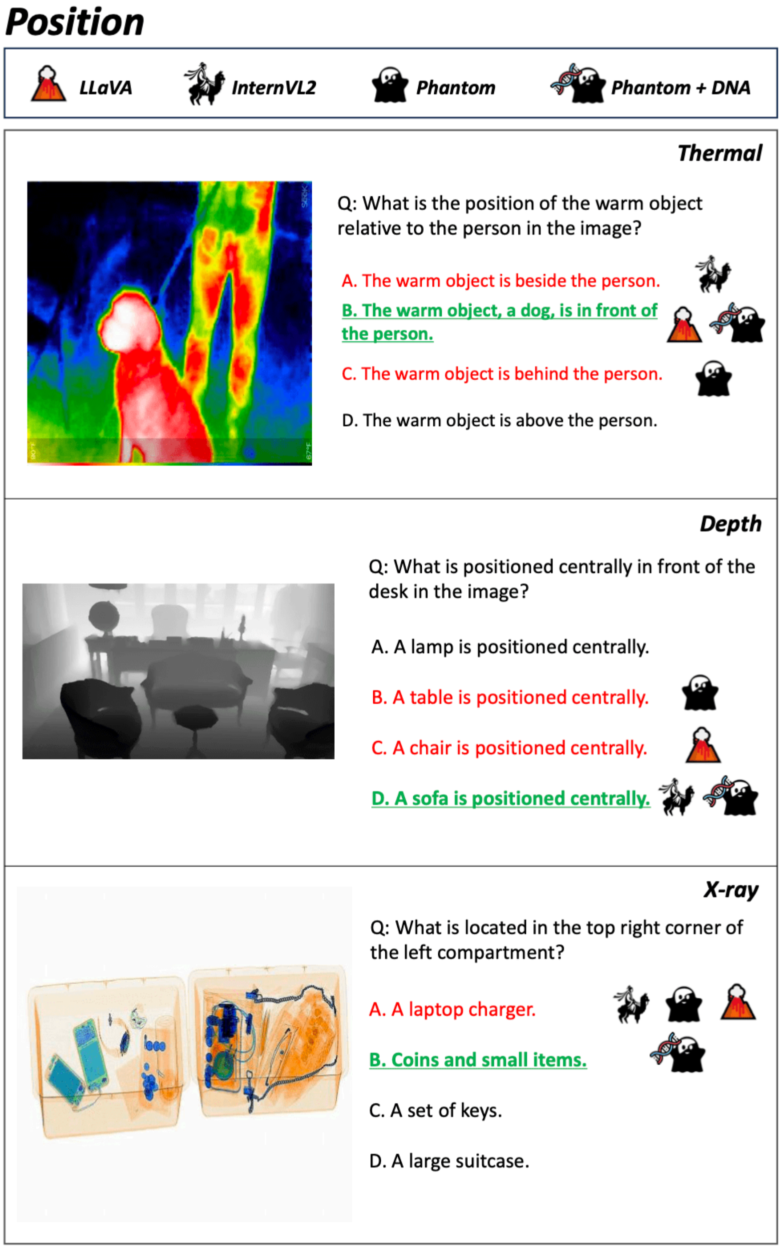
A. Four

B. Two

C. Five

D. Three

Figure 9. The comparison of performance across different multi-vision sensors with respect to the representative VLMs in the Multi-vision Perception task (Counting). Green font denotes the correct answer, while red font denotes the incorrect answer.



Q: What is the position of the warm object relative to the person in the image?

A. The warm object is beside the person.

B. The warm object, a dog, is in front of the person.

C. The warm object is behind the person.

D. The warm object is above the person.

Q: What is positioned centrally in front of the desk in the image?

A. A lamp is positioned centrally.

B. A table is positioned centrally.

C. A chair is positioned centrally.

D. A sofa is positioned centrally.

Q: What is located in the top right corner of the left compartment?

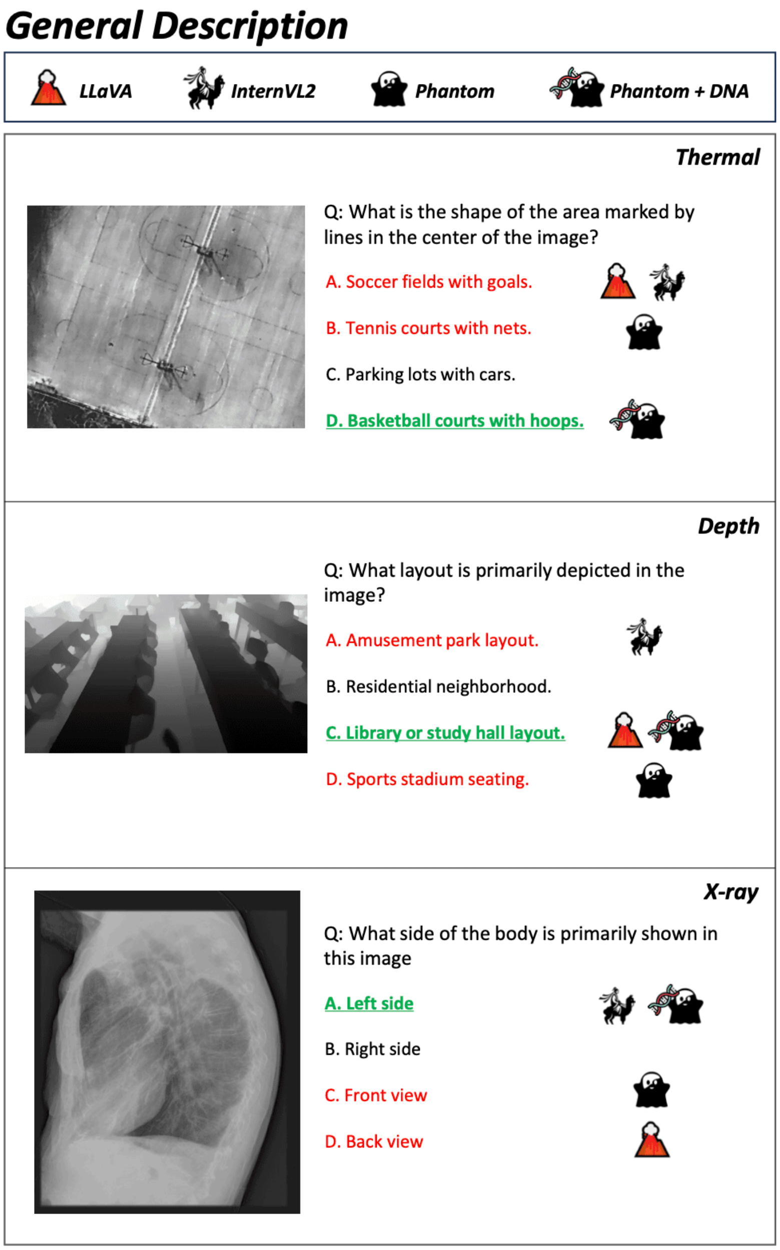
A. A laptop charger.

B. Coins and small items.

C. A set of keys.

D. A large suitcase.

Figure 10. The comparison of performance across different multi-vision sensors with respect to the representative VLMs in the Multi-vision Perception task (Position). Green font denotes the correct answer, while red font denotes the incorrect answer.



Q: What is the shape of the area marked by lines in the center of the image?

A. Soccer fields with goals. — LLaVA, InternVL2

B. Tennis courts with nets. — Phantom

C. Parking lots with cars.

D. Basketball courts with hoops. — Phantom + DNA

Q: What layout is primarily depicted in the image?

A. Amusement park layout. — InternVL2

B. Residential neighborhood.

C. Library or study hall layout. — LLaVA, Phantom + DNA

D. Sports stadium seating. — Phantom

Q: What side of the body is primarily shown in this image

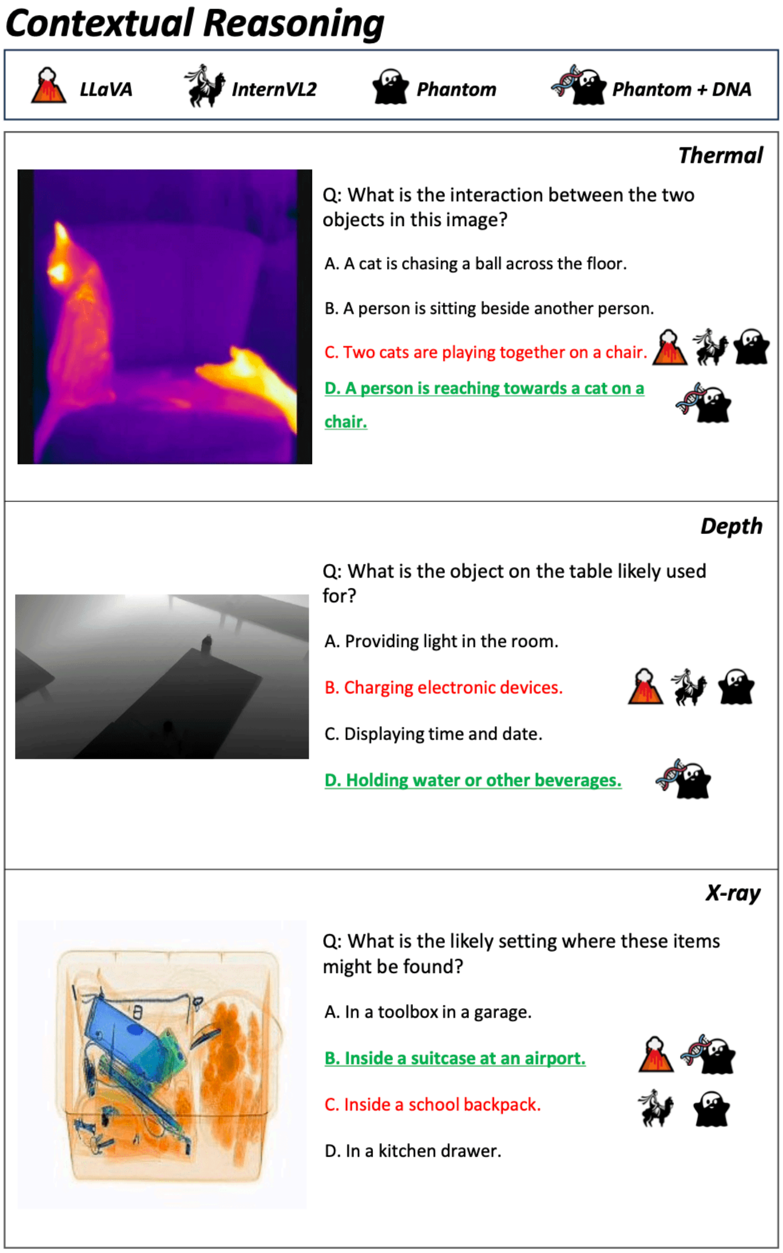
A. Left side — InternVL2, Phantom + DNA

B. Right side

C. Front view — Phantom

D. Back view — LLaVA

Figure 11. The comparison of performance across different multi-vision sensors with respect to the representative VLMs in the Multi-vision Perception task (General Description). Green font denotes the correct answer, while red font denotes the incorrect answer.



Thermal

Q: What is the interaction between the two objects in this image?

A. A cat is chasing a ball across the floor.

B. A person is sitting beside another person.

C. Two cats are playing together on a chair.

D. A person is reaching towards a cat on a chair.

Depth

Q: What is the object on the table likely used for?

A. Providing light in the room.

B. Charging electronic devices.

C. Displaying time and date.

D. Holding water or other beverages.

X-ray

Q: What is the likely setting where these items might be found?

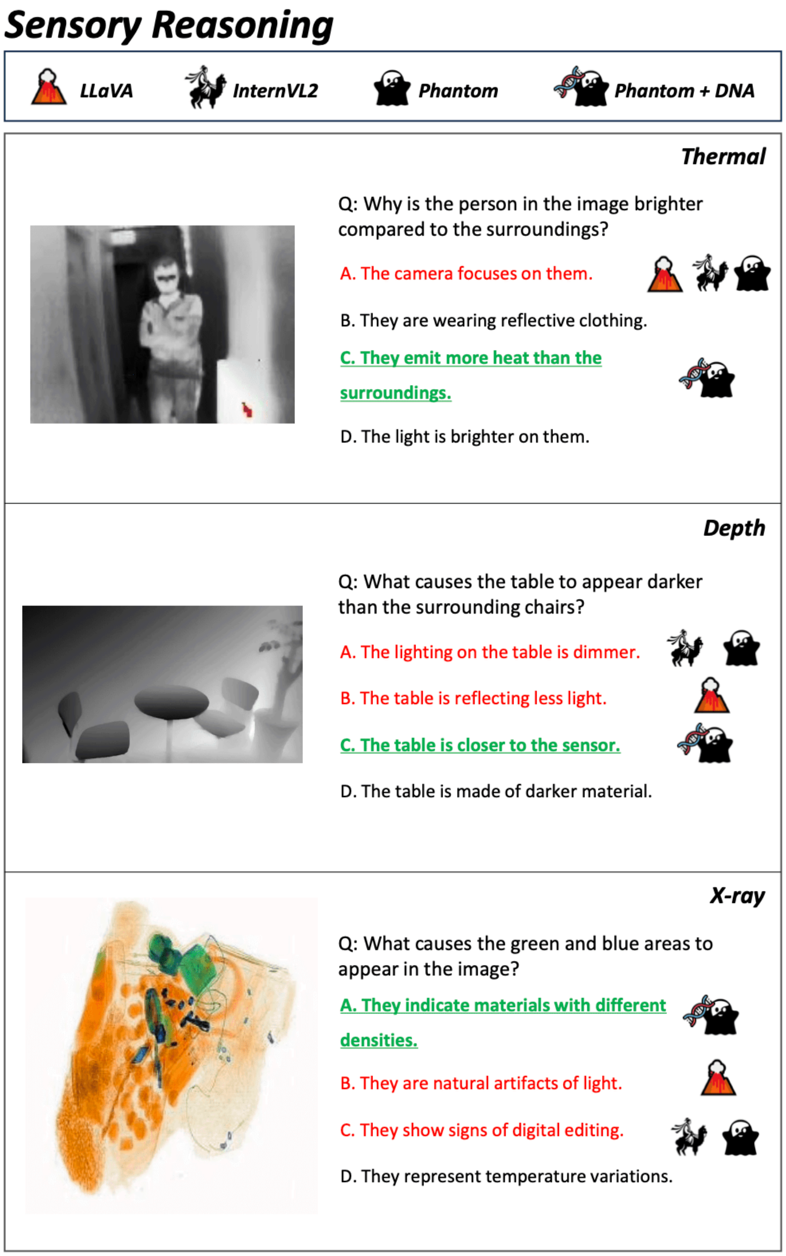
A. In a toolbox in a garage.

B. Inside a suitcase at an airport.

C. Inside a school backpack.

D. In a kitchen drawer.

Figure 12. The comparison of performance across different multi-vision sensors with respect to the representative VLMs in the Multi-vision Reasoning task (Contextual Reasoning). Green font denotes the correct answer, while red font denotes the incorrect answer.



Q: Why is the person in the image brighter compared to the surroundings?

A. The camera focuses on them.

B. They are wearing reflective clothing.

C. They emit more heat than the surroundings.

D. The light is brighter on them.

Q: What causes the table to appear darker than the surrounding chairs?

A. The lighting on the table is dimmer.

B. The table is reflecting less light.

C. The table is closer to the sensor.

D. The table is made of darker material.

Q: What causes the green and blue areas to appear in the image?

A. They indicate materials with different densities.

B. They are natural artifacts of light.

C. They show signs of digital editing.

D. They represent temperature variations.

Figure 13. The comparison of performance across different multi-vision sensors with respect to the representative VLMs in the Multi-vision Reasoning task (Sensory Reasoning). Green font denotes the correct answer, while red font denotes the incorrect answer.

References

1. ^Koh JY, Salakhutdinov R, Fried D. "Grounding language models to images for multimodal inputs and outputs." In: International Conference on Machine Learning. PMLR; 2023. p. 17283–17300.

2. ^Ren S, Yao L, Li S, Sun X, Hou L (2024). "Timechat: A time-sensitive multimodal large language model for long video understanding". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323.
3. ^Ye J, Hu A, Xu H, Ye Q, Yan M, Dan Y, Zhao C, Xu G, Li C, Tian J, et al. mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499. 2023.
4. a, b, c, dOpenAI (2024). "Hello GPT-4o". https://openai.com/index/hello-gpt-4o/.
5. a, bMao J, Qian Y, Zhao H, Wang Y (2023). "Gpt-driver: Learning to drive with gpt". arXiv preprint arXiv:2310.01415.
6. a, bXu Z, Zhang Y, Xie E, Zhao Z, Guo Y, Wong KYK, Li Z, Zhao H (2024). "Drivegpt4: Interpretable end-to-end autonomous driving via large language model". IEEE Robotics and Automation Letters. **2024**. Published by IEEE.
7. ^Guo Z, Yagudin Z, Lykov A, Konenkov M, Tsetserukou D (2024). "VLM-Auto: VLM-based Autonomous Driving Assistant with Human-like Behavior and Understanding for Complex Road Scenes". arXiv. arXiv:2405.05885 [cs.RO].
8. ^Chu X, Qiao L, Lin X, Xu S, Yang Y, Hu Y, Wei F, Zhang X, Zhang B, Wei X, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886. 2023.
9. ^Dinh QM, Ho MK, Dang AQ, Tran HP (2024). "TrafficVLM: A Controllable Visual Language Model for Traffic Video Captioning". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 7134-7143.
10. a, bCho Y, Kim T, Shin H, Cho S, Shin D (2024). "Pretraining Vision-Language Model for Difference Visual Question Answering in Longitudinal Chest X-rays". arXiv. Available from: https://arxiv.org/abs/2402.08966.
11. ^Gao J, Sarkar B, Xia F, Xiao T, Wu J, Ichter B, Majumdar A, Sadigh D. "Physically grounded vision-language models for robotic manipulation." In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2024. p. 12462–12469.
12. ^Huang W, Wang C, Li Y, Zhang R, Fei-Fei L (2024). "ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation". arXiv. arXiv:2409.01652 [cs.RO].
13. ^Duan J, Pumacay W, Kumar N, Wang YR, Tian S, Yuan W, Krishna R, Fox D, Mandlekar A, Guo Y (2024). "AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation". arXiv. arXiv:2410.00371 [cs.RO].
14. ^Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, Ding T, Driess D, Dubey A, Finn C, Florence P, Fu C, Gonzalez Arenas M, Gopalakrishnan K, Han K, Hausman K, Herzog A, Hsu J, Ichter B, Irpan A, Joshi N, Julian R, Kalashnikov D, Kuang Y, Leal I, Lee L, Lee TW, Levine S, Lu Y, Michalewski H, Mordatch I, Pertsch K, Rao K, Reymann K, Ryoo M, Salazar G, Sanketi P, Sermanet P, Singh J, Singh A, Soricut R, Tran H, Vanhoucke V, Vuong Q, Wahid A, Welker S, Wohlhart P, Wu J, Xia F, Xiao T, Xu P, Xu S, Yu T, Zitkovich B. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. 2023. Available from: https://arxiv.org/abs/2307.15818.

15. ^Huang S, Chang H, Liu Y, Zhu Y, Dong H, Gao P, Boularias A, Li H (2024). "A3VLM: Actionable Articulation-Aware Vision Language Model". arXiv. Available from: https://arxiv.org/abs/2406.07549.

16. a, b, c, d, e, f, g, hOpenGVLab. InternVL2: Better than the Best—Expanding Performance Boundaries of Open-Source Multimodal Models with the Progressive Scaling Strategy. 2024. Available from: https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

17. a, b, c, d, eYao Y, Yu T, Zhang A, Wang C, Cui J, Zhu H, Cai T, Li H, Zhao W, He Z, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800. 2024.

18. ^Shi Y, Gao Y, Lai Y, Wang H, Feng J, He L, Wan J, Chen C, Yu Z, Cao X (2024). "Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models". arXiv preprint arXiv:2402.04178.

19. ^Bazi Y, Al Rahhal MM, Bashmal L, Zuair M (2023). "Vision--language model for visual question answering in medical imagery". Bioengineering. **10** (3): 380.

20. a, b, cLiu H, Li C, Wu Q, Lee YJ (2023). "Visual Instruction Tuning". NeurIPS.

21. ^Liu H, Li C, Li Y, Li B, Zhang Y, Shen S, Lee YJ. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge [Internet]. 2024 Jan. Available from: https://llava-vl.github.io/blog/2024-01-30-llava-next/.

22. a, b, cLi J, Li D, Savarese S, Hoi S. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In: International conference on machine learning. PMLR; 2023. p. 19730-19742.

23. a, b, cCheng Z, Leng S, Zhang H, Xin Y, Li X, Chen G, Zhu Y, Zhang W, Luo Z, Zhao D, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. arXiv preprint arXiv:2406.07476. 2024.

24. a, b, c, d, e, f, gWang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, Wang J, Ge W, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191. 2024.

25. ^Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I (2023). "Imagebind: One embedding space to bind them all". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 15180–15190.

26. ^Su Y, Lan T, Li H, Xu J, Wang Y, Cai D (2023). "Pandagpt: One model to instruction-follow them all". arXiv preprint arXiv:2305.16355.

27. a, bFu C, Chen P, Shen Y, Qin Y, Zhang M, Lin X, Yang J, Zheng X, Li K, Sun X, Wu Y, Ji R (2024). "MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models". arXiv. arXiv:2306.13394 [cs.CV].

28. a, bLi B, Wang R, Wang G, Ge Y, Ge Y, Shan Y (2023). "Seed-bench: Benchmarking multimodal llms with generative comprehension". arXiv preprint arXiv:2307.16125.

29. a, bYue X, Ni Y, Zhang K, Zheng T, Liu R, Zhang G, Stevens S, Jiang D, Ren W, Sun Y, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 9556-9567.

30. a, bWu H, Zhang Z, Zhang E, Chen C, Liao L, Wang A, Li C, Sun W, Yan Q, Zhai G, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181. 2023.

31. a, b *Liu Y, Duan H, Zhang Y, Li B, Zhang S, Zhao W, Yuan Y, Wang J, He C, Liu Z, et al. Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision. Springer; 2025. p. 216–233.*

32. a, b *Liu H, Li C, Li Y, Lee YJ (2023). "Improved Baselines with Visual Instruction Tuning". arXiv. arXiv:2310.03744.*

33. a, b, c, d, e, f, g, h *Lee BK, Chung S, Kim CW, Park B, Ro YM (2024). "Phantom of latent for large language and vision models". arXiv preprint arXiv:2409.14713.*

34. a, b *Gemini Team, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, Tanzer G, Vincent D, Pan Z, Wang S, et al. (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". arXiv preprint arXiv:2403.05530.*

35. a, b *Anthropic. "Claude 3.5 sonnet". https://www.anthropic.com/news/claude-3-5-sonnet, 2024.*

36. ^ *Kahneman D, Treisman A, Gibbs BJ (1992). "The reviewing of object files: Object-specific integration of information". Cognitive psychology.* **24** *(2): 175–219.*

37. ^ *Broadbent DE. Perception and communication. Elsevier; 2013.*

38. ^ *Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems.* **35***: 27730–27744, 2022.*

39. ^ *Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C (2024). "Direct preference optimization: Your language model is secretly a reward model". Advances in Neural Information Processing Systems.* **36***.*

40. ^ *Meng Y, Xia M, Chen D (2024). "Simpo: Simple preference optimization with a reference-free reward". arXiv preprint arXiv:2405.14734.*

41. ^ *Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2024). "Qlora: Efficient finetuning of quantized llms". Advances in Neural Information Processing Systems.* **36***.*

42. ^ *Loshchilov I, Hutter F (2019). "Decoupled Weight Decay Regularization". arXiv. arXiv:1711.05101 [cs.LG].*

43. a, b *Liu J, Fan X, Huang Z, Wu G, Liu R, Zhong W, Luo Z (2022). "Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5802–5811.*

44. a, b ***Roboflow****. "thermal dogs and people x6ejw Dataset". Roboflow Universe. Nov 2022. Available from: https://universe.roboflow.com/object-detection/thermal-dogs-and-people-x6ejw. Visited on 2023-03-29.*

45. a, b *harang. pet Dataset. Roboflow Universe. Roboflow; 2024 Jul. Available from: https://universe.roboflow.com/harang/pet-kjl3x. Visited on 2024-10-28.*

46. a, b *Visual. Thermal Dataset. Roboflow Universe. Roboflow; 2023. Available from: https://universe.roboflow.com/visual-iqhyh/thermal-duv93. visited on 2024-10-22.*

47. a, b *Suo J, Wang T, Zhang X, Chen H, Zhou W, Shi W (2023). "HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection". Scientific Data.* **10***: 227.*

48. [a, b] *one. animal-detection-flir-extra Dataset. Roboflow Universe. Roboflow; 2023. Available from: https://universe.roboflow.com/one-rphct/animal_detection_flir_extra. Visited on 2024-10-28.*

49. [a, b] *Cook J. chips-thermal-face-dataset [Internet]. 2020 Apr [cited 2024 Oct 28]. Available from: https://www.kaggle.com/datasets/kagglechip/chips-thermal-face-dataset.*

50. [a, b] *NJUST. IFSOD Dataset. https://universe.roboflow.com/njust-oxpbo/ifsod, 2023. Roboflow Universe. Roboflow. visited on 2024-10-28.*

51. [a, b] *Vasiljevic I, Kolkin N, Zhang S, Luo R, Wang H, Dai FZ, Daniele AF, Mostajabi M, Basart S, Walter MR, Shakhnarovich G. DIODE: A Dense Indoor and Outdoor DEpth Dataset. 2019. Available from: https://arxiv.org/abs/1908.00463.*

52. [a, b] *Silberman N, Hoiem D, Kohli P, Fergus R (2012). "Indoor Segmentation and Support Inference from RGBD Images". In: ECCV.*

53. [a, b] *Cho J, Min D, Kim Y, Sohn K (2021). "DIML/CVL RGB-D Dataset: 2M RGB-D Images of Natural Indoor and Outdoor Scenes". arXiv. arXiv:2110.11590 [cs.CV].*

54. [a, b] *Farina E, Kitamura F (2022). UNIFESP X-ray Body Part Classifier Competition. Kaggle. Available from: https://kaggle.com/competitions/unifesp-x-ray-body-part-classifier.*

55. ^ *Malek. "X-ray baggage detection dataset". Roboflow Universe. Roboflow; 2022 Apr. Available from: https://universe.roboflow.com/malek-mhnrl/x-ray-baggage-detection. visited on 2024-11-11.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.