Generalize LMMs to Versatile Visual Modalities via Fabricated Modality Synthesis

Shihao Yuan[1], Yuanze Li[1], Ruyi Zhang[1], Ming Liu[1](✉), and Wangmeng Zuo[1]

{csshihao, sqleopop, csmliu}@outlook.com, {ruyi.zhang.maggie, cswmzuo}@gmail.com

[1]Faculty of Computing, Harbin Institute of Technology, Harbin, China

Abstract. Despite the advancements of Large Multimodal Models (LMMs) in RGB vision, their ability to generalize to unseen visual modalities remains a largely unexplored challenge. We argue that different visual modalities are merely distinct samplings of the same physical world. Therefore, effective generalization requires models to possess both modality-agnostic perception of scene semantics and the adaptability to modality-specific characteristics. To achieve this, we propose a training framework, **VVM-Tuning**, to equip LMMs with these capabilities through modality synthesis and modality contexts. Specifically, we synthesize diverse appearance-varied images from RGB scenes, training the model to disentangle invariant semantics from varying visual appearances, and align these appearances with language for visual concepts decoupled from modalities. We then introduce modality contexts in the prompt and use instruction tuning to assist the model in mapping these appearance variations back to modality-related attributes, enabling zero-shot adaptation to unseen modalities during inference. To facilitate research in this direction, we introduce **VVM-Bench**, a comprehensive benchmark featuring 6 real and synthetic modalities to evaluate semantic perception and modality understanding. Experiments demonstrate that, via our training on synthetic modalities, 5 tested models exhibit consistent improvements on both real-world and novel synthetic modalities without in-modality training. Source code and data will be publicly available at https://github.com/Hunter-Will/VVM-Tuning.

Keywords: LMMs · Instruction Tuning · Synthetic Data

1 Introduction

Current Large Multimodal Models (LMMs) demonstrate impressive visual performance on RGB images [10]; however, the generalization boundary on other visual modalities (*e.g.* infrared, depth, *etc.*) remains largely unexplored yet. Existing non-RGB vision in current LMMs primarily depends on the non-RGB data incorporated during the data scaling process [1,4,33]. This leads to a generalization gap when encountering unseen modalities, as the non-RGB vision is built on in-modality data and limited modality coverage.

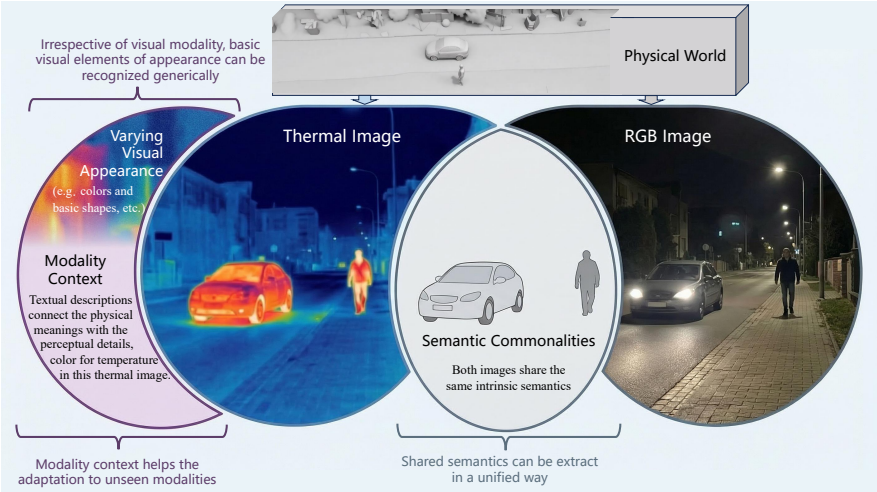

Fig. 1: An overview of our idea. Different visual modalities (Thermal and RGB) are both signal samplings of our physical world, sharing the same underlying semantics (formed by the person and the car). They also share basic visual concepts, as the color itself is invariant across modalities. The modality-specific physical meaning is encoded by the rearrangement and remapping of basic visual concepts. Thus, we introduce textual context to complement such knowledge during inference.

To address this gap and explore the generalization boundary of LMMs, we raise the question: *Is it possible for LMMs to generalize across versatile visual modalities without training on in-modality data?*

The solution to the problem starts from the nature of visual modalities, as shown in Fig. 1, though with different appearances, they are essentially digital signals collected by sensors (thermal and RGB camera in Fig. 1) from the same physical space. Therefore, images from different visual modalities can be viewed as distinct samplings of identical real-world scenes. This commonality lies the foundation for pan-modal generalization, as the understanding of underlying semantics is modality-agnostic. Furthermore, through observation, we find that in human vision, images from other visual modalities can still be **seen** even without understanding the modality. Take Fig. 2 as an example, our brain can still read the semantics of the image from basic visual elements. This gives us inspiration that LMMs can learn this process to recognize semantics even without certain modality knowledge, which we call Modality-unaware Perception. Also, those visual elements, belonging to general visual concepts, are not completely combined with semantics. Some of them (like the colors in the thermal image in Fig. 1) might not be a necessary part for understanding the whole image semantics under this modality-unaware circumstance. Instead, they are supposed to combine with the physical meaning of the visual modality to achieve a higher understanding of the image, such as the color-coded thermal image in Fig. 1. Although they can only be understood superficially without modality knowledge, they are still basic elements of both human vision and machine vision, which are completely

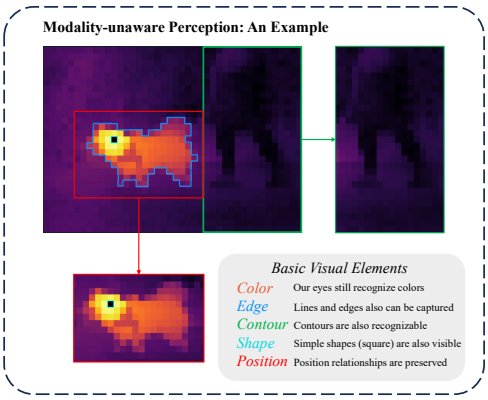

Fig. 2: An example of how human vision adapts to unseen visual modalities. Our eyes can capture basic visual elements regardless of modalities, and our brain recognizes the semantics even through a few visual elements. This demonstrates a compatible modality-unaware perception, which we could implement in LMMs' training.

generalized among diverse visual modalities, as long as they are 3-channel digital images. That means they are still perceivable modality-unawarely, for example, the thermal color in Fig. 1 can still be understood as colors without knowing its connection with temperature. We refer to this part that only perceives basic visual elements without understanding further as *photographic perception*. Correspondingly, the part that comprehends underlying semantics is referred to as *semantic perception*. Both parts compose a modality-unaware perception without modality knowledge. This modality-unaware perception demonstrates a highly possible approach of unseen visual modality generalization; however, it doesn't mean that every single visual modality can be understood in the same way. Different visual modalities have their unique sampling function of the real world, which is defined by the sensor and signal processing procedure. This often leads to varying appearances of the same underlying semantics and usually means a connection between the physical attributes of entities in the real-world scene and the unique visual appearances in digital images, which is, of course, different from regular RGB camera images. Besides, non-RGB modalities usually map an additional physical quantity to the color of the image, known as the *pseudo-color* [13,28], which actually enriches the meaning of basic visual elements. Understanding such a connection or mapping requires modality knowledge, and such modality knowledge is not reachable during training for the unseen visual modalities in our discussion. Thus, we introduce modality contexts to complement the modality understanding by utilizing the instruction following and in-context learning ability of LMMs during inference.

Viewing LMMs' training from this perspective, we argue that the current LMMs are largely overlooking the modality-unaware perception. The visual-language alignment, building primarily on RGB images, doesn't distinguish the invariant semantics and varying visual appearances. This would cause an over-alignment and isolate other visual modalities, as the model needs to learn new

alignments for each different visual modality incorporated in training. Based on the above insights, we proposed a novel training framework called **VVM-tuning** to: 1) Build up a modality-unaware perception via fabricated images; 2) Enable the modality-aware understanding of LMMs through modality synthesis.

For the first part, we leverage RGB images to fabricate non-RGB images to mimic varying appearances from different visual modalities, keeping the semantics invariant from RGB images. Therefore, the model has a chance to learn how to read invariant semantics through varying appearances. Furthermore, we design two sets of descriptions and VQA tasks to disentangle the semantic parts and perceptual parts of the visual-language alignments. By separate alignment, the model is taught to extract the semantics and keep perceptual elements aligned with language at the same time, which would benefit the modality-aware understanding for reconnecting the physical meanings and visual appearances. Because of the fabricated image, we can largely diversify the visual appearances of common scenes; thus, we establish a more general perception of visual appearances, which leads to a base for modality-aware understanding.

For the second part, we utilize modality synthesis to build a general synthetic instruction set to train the model with modality contexts. These modality contexts are fabricated to construct diverse synthesized visual modalities combined with the fabricated images. From these synthetic data, the model learns to connect the modality-specific physical attributes to varying visual appearances and further integrates with the semantics to achieve better modality understanding. With the combination of the above two parts, we enable the zero-shot generalization of LMMs to unseen visual modalities without in-modality data for training.

Finally, we assemble a diverse visual modality benchmark, VVM-bench, containing 6 real and synthesized imaging modalities. From our perspective, this benchmark evaluates the generalization ability through 11 VQA tasks from modality-unaware perception and modality-aware understanding, respectively. Half of the 6 non-RGB visual modalities are collected from VS-TDX [7], and 3 fabricated modalities to mimic unseen visual modalities. Experiments show an average of 8.2% improvements of perception tasks and 3.8% improvements of understanding tasks on this benchmark among 5 different base models and in 6 synthesized and real visual modalities for both perception and understanding.

In summary, our contribution can be listed as follows:

- To our best knowledge, we are the first to explore the potential of LMMs to generalize to unseen visual modalities, and demonstrate a possible approach through modality context and modality synthesis.
- We introduce synthetic data for other visual modalities and propose a fabricated modality synthesis data pipeline, which can produce diverse fabricated images and synthesized modality contexts.
- We propose a training framework establishing modality-unaware perception and modality-aware understanding, and confirm that LMMs can generalize to other visual modalities via modality synthesis and disentangled tasks.
- We assemble a versatile visual modality benchmark to evaluate the visual ability and generalization on visual modalities for LMMs.

2 Related Works

Since the emergence of LLaVA [19], establishing the visual instruction tuning paradigm, both LMMs and benchmarks targeting their RGB visual capabilities have seen significant development [10]. However, the generalization boundary of LMMs to non-RGB visual modalities remains underexplored. Modern LMMs, such as LLaVA-OneVision [1, 16], Qwen-VL [3, 4, 17], and InternVL [6, 33, 41], incorporate diverse and abundant visual data, which may contain other modalities in their training subsets like DocQA [24] and ScienceQA [23]. However, their primary focus remains on scaling data and diversity to improve general visual abilities, rather than investigating the generalization of unseen visual modality.

Meanwhile, some previous works, such as Imagebind [12] and PandaGPT [30], also involve thermal and depth images. However, the basic idea of Imagebind [12] is "binding" other general modalities (audio, text, *etc.*) to RGB images in order to eliminate the need for pairwise alignments between these modalities. As a result, RGB-D and RGB-T paired data are still required during the training of Imagebind [12], which are not regarded as **unseen** visual modality and not within our scope of discussion. Similarly, the cross-modal capabilities discussed in PandaGPT [30] refer to the zero-shot instruction following ability among 6 seen modalities inherited from Imagebind [12], which is also beyond the discussion. Concurrently, research dedicated to non-RGB modalities has largely been confined to fine-tuning LMMs on specialized datasets. For instance, Infrared-LLaVA [14] adapts LLaVA for infrared-specific tasks by tailoring the model to the infrared modality. SpatialBot [5] focuses on the spatial understanding with both RGB and depth images, also relying on RGB-D training to inject depth knowledge. Specialized LMMs, for instance, EarthGPT [38] in remote sensing, XrayGPT [32] and RadVLM [8] in radiology, are trained with domain-specific data. Though they also involve multiple visual modalities, they are focusing on application tasks, leaving their inherent ability to generalize across different useen visual modalities behind.

This gap is further reflected in existing evaluation suites. Mainstream benchmarks for general-purpose LMMs, such as MMBench [20], MME [9], MMMU [37], MM-Vet [35], and SEED-Bench [18], are predominantly centered on RGB images. While a few recent efforts have extended evaluation to non-RGB modalities for LMMs, the RGB-Th-Bench [26] focuses on the cross-reference understanding of RGB-T paired images, ignoring the possible generalization to unseen modalities on general visual abilities. The VS-TDX [7] benchmark evaluates general visual abilities on 3 modalities, but still lacks focus on the principles and testing of unseen visual modality generalization. Other specialized evaluations, like MedSG-Bench [36] in medical imaging and RSVQA [21] in remote sensing, are highly domain-specific and not designed to assess general visual adaptability. Consequently, there remains a lack of benchmarks tailored to systematically evaluate the cross-visual-modal generalization capabilities of LMMs. These limitations underscore a critical gap in understanding the true visual versatility of modern LMMs beyond the RGB vision, motivating our study to investigate their generalization potential across different visual modalities.

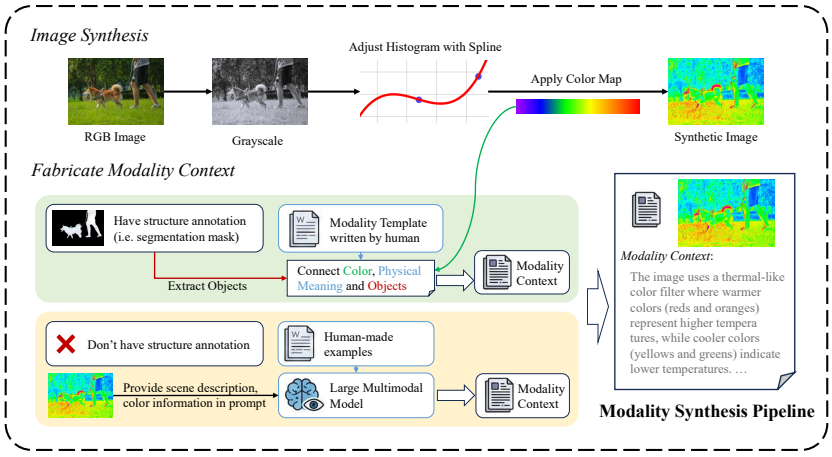

Fig. 3: An overview of our modality synthesis pipeline, composed of the image synthesis and modality context fabrication.

3 Methods

3.1 Preliminary

From the base idea of how different visual modalities follow the basic imaging principles to map the real-world scenarios into photos, our method tries to imitate the same process of how humans understand versatile visual modalities, which leads to two different abilities: modality-unaware perception and modality-aware understanding. To equip LMMs with these two abilities, we design two sets of generated instruction tasks [27] and train the model to learn the adaptability from these tasks with fabricated images and modalities. Our VVM-Tuning mainly follows the paradigm of visual instruction tuning [19]. The training objective follows the common approach of Supervised Fine-Tuning (SFT), which involves the Cross Entropy loss function $\mathcal{L}_{SFT}$.

$$\mathcal{L}_{SFT} = -\sum_{i=1}^{n} \log \pi_\theta(y_i|X, y_1, y_2, ..., y_{i-1}), \tag{1}$$

The whole formalized expression of the training is to optimize $\mathcal{L}_{SFT}(T_a, T_{gt})$ to minimize the differences between the generated T_a and the ground truth T_{gt}.

3.2 Modality Synthesis

Effective generalization requires cross-modal perception and context-guided modality adaptation, which needs the training data covering versatile modal characteristics. Thus, we introduce a visual modality synthesis pipeline, shown in Fig. 3. The pipeline is mainly composed of two parts: 1) image synthesis and 2) modality context fabrication.

The first part is to imitate the unique characteristics that differ from regular RGB photos, such as the pseudo-color pattern [39] and histogram distribution. Here, we introduce a simple pipeline to fabricate non-RGB images from RGB images. As shown in Fig. 3, the RGB source images are first converted to grayscale for histogram adjustments. We apply cubic spline interpolation [25] to randomize the shape of the histogram [29] and increase the diversity by flipping the whole histogram randomly. Then the grayscale images are painted with pseudo-color by randomly applying 22 color maps [40] that are commonly used in scientific computing and visualization. After color mapping, several image augmentations [34], such as noise injection, smoothing, and blurring, are employed to simulate the noisy and low-resolution modalities. With such a simple pipeline, we manage to fabricate diverse images with similar characteristics to the real non-RGB visual modalities, shown by the examples in Fig. 4.

Besides fabricated images, we also make up corresponding physical meanings and explanations to compose modality contexts. This is done by two approaches according to whether there are structure annotations (segmentation mask or object information) of source RGB images, shown in Fig. 3. For the images with annotations, the modality contexts are assembled structure texts by connecting the color maps with objects in the image by labels such as segmentation masks. Then, the value interval mapping colors is ascribed with human-made physical attributes and fit into manual templates to form modality contexts. For the non-annotated images, we leverage the in-context learning ability of LMM to generate color and object-related modality context by prompting it to follow manually written examples. Combining fabricated images and synthesized modality contexts results in modality synthesis that contains diverse characteristics to teach LMMs the varying and invariant features of versatile visual modalities. Further details of the modality synthesis pipeline can be found in the Appendix.

3.3 Modality-unaware Perception

The foundation of generalizing to unseen visual modalities is an invariant perception ability across them. This ability is called modality-unaware perception, which contains two parts: 1) the underlying semantic perception and 2) the photographic perception of basic visual elements. In regular RGB visual-language alignment, these two parts of perception are usually entangled together, aligned

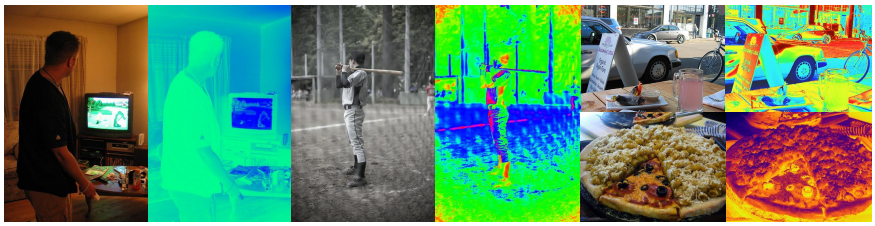

Fig. 4: Different examples of fabricated images with different color maps produced by our image synthesis pipeline.

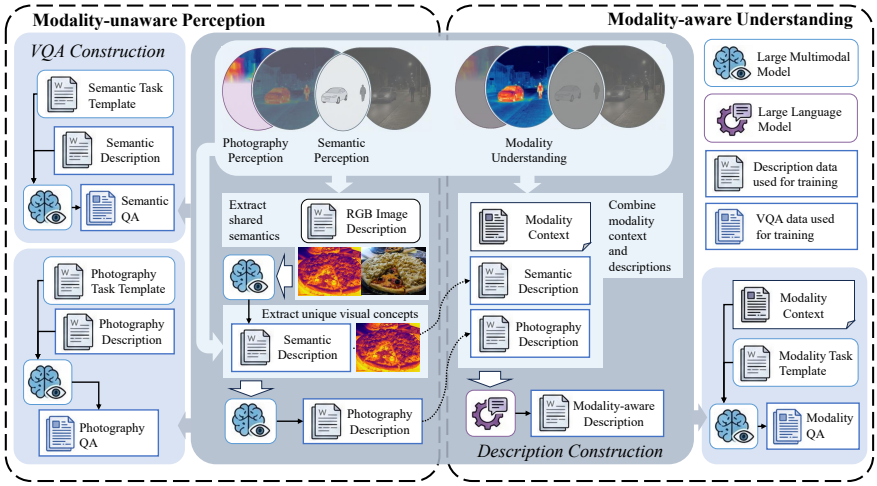

Fig. 5: An overview of the task construction of **VVM-Tuning**, including modality-unaware perception tasks and modality-aware understanding tasks. Starting from our core insight at the top, the region illustrates the construction of the semantic description and photographic description tasks to form modality-aware descriptions. Then, the surrounding VQA tasks are generated from task templates with descriptions.

with the language as a whole. However, when other visual modalities are involved, the entangled perception isolates the shared semantics between visual modalities due to different appearances. This causes the LMMs need to learn new alignments for each different visual modality. As the underlying semantics are encoded in different forms of expression by different visual modalities, we first design scene semantic tasks to equip LMM with the ability to extract the underlying semantics. Then, photographic perception tasks are involved to capture unique visual clues introduced by different visual modalities. With the image synthesis pipeline, the fabricated images are convenient to create the same semantics with different visual appearances. Based on them, we design several instructions to train the model to recognize the physical entities in various appearances and activate the common knowledge with such entities. Furthermore, the photographic perception tasks ensure the LMM captures the visual details and remains loyal to the image instead of being confused by semantics.

For better adaptability, both semantic perception and photographic perception are composed of two forms of tasks, image description [15] and visual question-answering (VQA) [2, 19]. As shown in the left part of Fig. 5, they form four tasks: semantic description, semantic QA, photographic description, and photographic QA. First, for the semantic description, we leverage the common semantics of fabricated images and prompt an LMM to extract the overlapped parts from the description of the RGB image according to the RGB image and the corresponding fabricated image. That leads to a semantic description that extracts the shared semantics of fabricated and RGB images. Then, the photographic description is generated by prompting the LMM to extract the unique

parts from the fabricated image that are not in the semantic description. However, due to the lack of perception of these fabricated images, the colors in the photographic description need to be recorrected by introducing the color names during the image synthesis pipeline. As for the VQA instructions, we design 11 different sub-tasks for semantic and photographic QA as human-made question templates and utilize a large-sized LMM to generate diverse questions and answers according to RGB images, fabricated images, and corresponding descriptions, respectively. Finally, the questions are randomly organized as a mix of non-choice questions and choice questions to combine with the descriptions to create a completely modality-unaware training set. Further details of the task distribution and templates can be found in the Appendix.

3.4 Modality-aware Understanding

The purpose of this part of VVM-Tuning is to enhance the understanding of modality contexts. Modality context is a textual prompt, including specific characteristics and physical meanings of the modality of the image. The two ways of synthesizing modality context are described in Sec. 3.2. The modality-aware understanding tasks leverage synthesized modality contexts to equip the model with the ability to understand modality contexts and connect the physical meaning to specific visual concepts or appearances. Same as the modality-unaware perception, the modality-aware training also consists of image description and VQA tasks, called modality description and modality QA, shown in the right part of Fig. 5. The construction of modality description is achieved by leveraging a large language model to combine the semantic description, photographic description, and synthesized modality context together to form an image description with modality knowledge. The modality QA is generated by LMM according to the modality context and human-made templates. We don't use the modality description here as the VQA tasks are concentrated to particular part of the image and the modality description only provides a general description of the entire image. However, scene and photographic descriptions are involved to make sure the LMM for generation understands the non-RGB images enough. The question templates of modality QA are specially designed to mainly cover different circumstances that require a knowledge of the visual modality. Further details can be found in the Appendix.

3.5 Benchmark

To mitigate the lack of evaluation data and confirm the feasibility of our training framework, we assembled a benchmark that contains multiple real and synthesized visual modalities to test the perception and understanding abilities of LMMs. The VVM-Bench, as shown in Fig. 6, is composed of 6 visual modalities and 11 VQA tasks belonging to perception and understanding, respectively. Among them, three modalities (thermal, depth, and X-ray) are collected from VS-TDX [7] with minimal changes of output formation to insert modality contexts in understanding tasks. The other three modalities are fabricated from

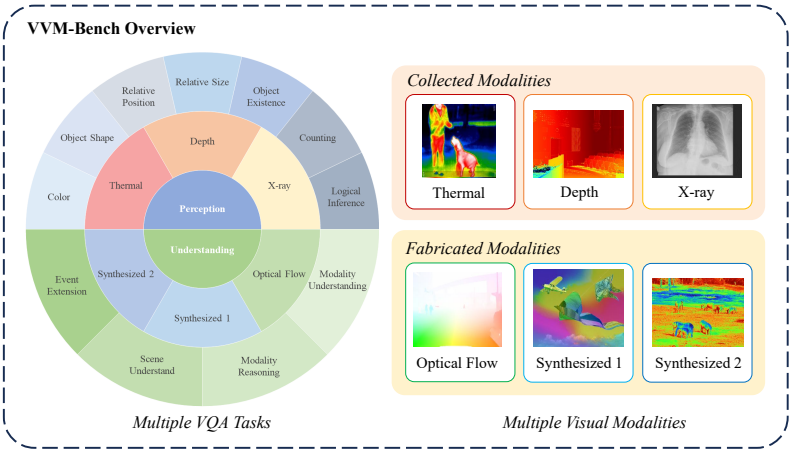

Fig. 6: An overview of VVM-Bench. Our VVM-Bench contains 6 visual modalities and 11 VQA tasks for perception and understanding evaluation, including 3 real modalities collected from previous work and 3 fabricated from RGB images.

RGB images, including Optical Flow, Synthesized 1, and Synthesized 2. Specifically, the optical flow images are visualizations of predicted optical flows from RAFT [31] to serve as signals from an almost entirely new and unknown imaging sensor with real physical meanings. The other two visual modalities are completely synthesized as unseen visual modalities. The images in Synthesized 1 are fabricated using a style transfer algorithm, StyleShot [11], to apply a visualization style on normal RGB images. As for images for Synthesized 2, we use the same image synthesis method as described in Sec. 3.2 with changed hyperparameters to explore the generative diversity of our image synthesis pipeline.

Our benchmark evaluates two different aspects of LMMs corresponding to our insights, the modality-unaware perception and modality-aware understanding, abbreviated as perception and understanding in Fig. 6. The perception aspect contains 7 sub-tasks, and all tasks focus on the semantics or visual attributes perceivable without modality knowledge. No modality contexts are provided in these tasks. Instead, the understanding tasks are all about the modality understanding or knowledge association with the visual modality, which provides modality contexts in the prompts. All modality contexts are written by human and we use choice questions in VVM-bench for the convenience of evaluation. Detailed task templates and modality contexts can be found in the Appendix.

4 Experiments

4.1 Implementation Details

We have employed the training framework on multiple base LMMs, including LLaVA-1.5 [19], Qwen2.5-VL [17], and Qwen3-VL [4] with different sizes in our capabilities. All the experiments are conducted on a 4 NVIDIA A800 GPU server

with a total batch size of 32. Following the common setting of Supervised Fine-Tuning (SFT), we apply the AdamW [22] as optimizer and set the learning rate to $1e-5$. We adopt a single-staged fine-tuning with roughly 50k training samples in total for 1 epoch, involving all tasks in modality-unaware perception and modality-aware understanding. To preserve the general abilities of LMMs, we retain 12k RGB visual instruction data in the 50k training dataset.

Our benchmark follows the common VQA benchmark setting, with the form of choice questions. To quantify the results, the LMMs are instructed to output the option's letter only to calculate accuracy. Specifically, for tasks that need modality contexts, all modality contexts are inserted between the image and the question as contexts. Further evaluation details can be found in the Appendix.

4.2 Results of Modality-unaware Perception

The results in Tab. 1 show the VQA accuracy on the perception subset of VVM-Bench. The results in Tab. 1 report the total accuracy of all samples in each modality for both baselines and our tuned models. From the results in Tab. 1, experiments show our VVM-tuning brought an average 8.2% improvement for all base models among 6 modalities on our VVM-Bench perception subset. Across all tested models, LLaVA-1.5 gained the biggest average improvement of 13.6% because of its relatively low base. The 4B model of Qwen-3-VL got the lowest improvement, but still exceeded 5%. This exhibits the adaptability of our modality-unaware perception training among different base models and confirms that the perception of LMMs on unseen visual modalities still has room for improvement.

Table 1: The results of the Modality-unaware Perception subset of VVM-bench. "Δ" represents accuracy difference between the model we tuned and the baselines. "Average" shows the average accuracy and difference across all tested modalities. "Average Δ" shows the average improvements across all tested models for each visual modality. Note that there is no modality context for this perception subset.

Model	Size	Collected Visual Modalities			Fabricated Visual Modalities			Average
		Thermal	Depth	X-Ray	Optical Flow	Synthesized 1	Synthesized 2	
LLaVA-1.5	7B	55.6%	62.4%	58.1%	53.2%	70.7%	62.6%	60.4%
LLaVA-1.5 (Ours)	7B	69.1%	69.8%	67.3%	75.4%	85.8%	77.0%	74.1%
Δ	-	+13.5%	+7.4%	+9.2%	+22.2%	+15.1%	+14.4%	+13.6%
Qwen-2.5-VL	3B	72.0%	67.9%	68.1%	67.2%	83.0%	74.9%	72.2%
Qwen-2.5-VL (Ours)	3B	81.7%	77.1%	77.5%	83.1%	88.9%	83.0%	81.9%
Δ	-	+9.7%	+9.2%	+9.4%	+15.9%	+5.9%	+8.1%	+9.7%
Qwen-2.5-VL	7B	74.5%	73.3%	75.7%	77.0%	87.1%	81.1%	78.1%
Qwen-2.5-VL (Ours)	7B	79.8%	79.2%	79.6%	88.4%	91.2%	86.4%	84.1%
Δ	-	+5.3%	+5.9%	+3.9%	+11.4%	+4.1%	+5.3%	+6.0%
Qwen-3-VL	4B	80.6%	77.8%	76.7%	73.0%	81.6%	75.6%	77.6%
Qwen-3-VL (Ours)	4B	82.6%	84.4%	79.5%	81.5%	86.9%	82.1%	82.8%
Δ	-	+2.0%	+6.6%	+2.8%	+8.5%	+5.3%	+6.5%	+5.3%
Qwen-3-VL	8B	79.8%	78.5%	73.8%	73.3%	82.4%	78.0%	77.6%
Qwen-3-VL (Ours)	8B	82.1%	84.2%	80.3%	86.5%	89.5%	82.1%	84.1%
Δ	-	+2.3%	+5.7%	+6.5%	+13.2%	+7.1%	+4.1%	+6.5%
Average Δ	-	+6.6%	+7.0%	+6.4%	+14.2%	+7.5%	+7.7%	+8.2%

Table 2: The results of the Modality-aware Understanding subset of VVM-bench with modality contexts. "Δ" represents accuracy difference between the model we tuned and the baselines. "Average" shows the average accuracy and difference across all tested modalities. "Average Δ" shows the average improvements across all tested models for each visual modality.

Model	Size	Collected Visual Modalities			Fabricated Visual Modalities			Average
		Thermal	Depth	X-Ray	Optical Flow	Synthesized 1	Synthesized 2	
LLaVA-1.5	7b	77.8%	80.3%	72.4%	76.7%	84.1%	84.8%	79.4%
LLaVA-1.5 (Ours)	7b	82.5%	81.2%	79.3%	89.2%	87.6%	89.8%	84.9%
Δ	-	+4.7%	+0.9%	+6.9%	+12.5%	+3.5%	+5.0%	+5.6%
Qwen-2.5-VL	3b	83.8%	85.5%	83.5%	89.6%	86.9%	91.0%	86.7%
Qwen-2.5-VL (Ours)	3b	89.3%	89.5%	85.9%	93.8%	91.0%	93.8%	90.6%
Δ	-	+5.5%	+4.0%	+2.4%	+4.2%	+4.1%	+2.8%	+3.8%
Qwen-2.5-VL	7b	86.4%	83.4%	84.1%	84.6%	88.8%	89.8%	86.2%
Qwen-2.5-VL (Ours)	7b	88.5%	90.2%	85.9%	93.8%	91.9%	93.5%	90.6%
Δ	-	+2.1%	+6.8%	+1.8%	+9.2%	+3.1%	+3.7%	+4.5%
Qwen-3-VL	4b	87.3%	83.8%	84.3%	90.0%	90.0%	91.8%	87.9%
Qwen-3-VL (Ours)	4b	89.6%	87.0%	88.0%	92.1%	90.5%	92.0%	89.9%
Δ	-	+2.3%	+3.2%	+3.7%	+2.1%	+0.5%	+0.2%	+2.0%
Qwen-3-VL	8b	89.3%	87.2%	83.1%	90.8%	87.9%	91.5%	88.3%
Qwen-3-VL (Ours)	8b	91.6%	91.5%	89.8%	91.7%	90.3%	93.8%	91.5%
Δ	-	+2.3%	+4.3%	+6.7%	+0.9%	+2.4%	+2.3%	+3.2%
Average Δ	-	+3.4%	+3.8%	+4.3%	+5.8%	+2.7%	+2.8%	+3.8%

Meanwhile, although training on synthetic data, all base models show at least an average 6% improvement on visual modalities with real physical meanings, as shown in Thermal, Depth, X-ray, and Optical Flow. This confirms the existence of the generalization foundation, and different real visual modalities do share commonalities in both semantics and perceptual elements. Also, the performance on synthesized modalities indicates that this improvement could generalize to completely unseen visual modalities and further confirm the effectiveness of our simple image synthesis pipeline.

4.3 Results of Modality-aware Understanding

From the results in Tab. 2, we can observe further improvements on modality understanding tasks. These tasks focus on understanding the modality context and linking the modality knowledge and common sense, which is more associated with the language model in LMM. As shown in Tab. 2, our VVM-tuning enhanced the ability to utilize modality context and reinterprete the perceptual features by a 3.8% average improvement across all models and modalities. Similar to before, among all baselines, LLaVA-1.5 has the greatest improvement of 5.6%, and the 4B Qwen-3-VL has the smallest improvement of 2.0%. This result further confirms the possibility of generalization across diverse unseen visual modalities. Tested LMMs demonstrate strong potential for learning from synthesized modalities and adapting to the other unseen visual modalities.

Among the evaluated modalities, the Optical Flow fabricated by us is the most improved, and the other three real modalities are next to it; our synthesized modalities get the lowest improvements. This shows that a synthetic gap still

Table 3: The results of 3 real visual modalities in the understanding subset with/out modality contexts. "M.C." is an abbreviation for Modality Context. "Δ" represents the accuracy difference with or without modality context for each model and modality. "Average Δ" demonstrates the average improvements of providing modality contexts for each modality across all models.

Model	Size	Thermal			Depth			X-Ray		
		w/o M.C.	w/ M.C.	Δ	w/o M.C.	w/ M.C.	Δ	w/o M.C.	w/ M.C.	Δ
LLaVA-1.5	7b	47.7%	77.8%	+30.2%	40.7%	80.3%	+39.6%	61.0%	72.4%	+11.5%
LLaVA-1.5 (Ours)	7b	63.1%	82.5%	+19.4%	43.5%	81.2%	+37.6%	73.0%	79.3%	+6.3%
Qwen-2.5-VL	3b	76.0%	83.8%	+7.8%	54.4%	85.5%	+31.0%	79.7%	83.5%	+3.8%
Qwen-2.5-VL (Ours)	3b	79.9%	89.3%	+9.5%	64.7%	89.5%	+24.8%	84.4%	85.9%	+1.5%
Qwen-2.5-VL	7b	69.6%	86.4%	+16.8%	46.3%	83.4%	+37.1%	82.9%	84.1%	+1.2%
Qwen-2.5-VL (Ours)	7b	74.5%	88.5%	+13.9%	54.3%	90.2%	+35.9%	84.1%	85.9%	+1.8%
Qwen-3-VL	4b	77.5%	87.3%	+9.8%	47.9%	83.8%	+35.9%	83.4%	84.3%	+0.9%
Qwen-3-VL (Ours)	4b	73.8%	89.6%	+15.8%	55.9%	87.0%	+31.1%	86.9%	88.0%	+1.1%
Qwen-3-VL	8b	71.4%	89.3%	+17.9%	53.8%	87.2%	+33.4%	82.3%	83.1%	+0.8%
Qwen-3-VL (Ours)	8b	79.1%	91.6%	+12.6%	70.4%	91.5%	+21.1%	88.2%	89.8%	+1.6%
Average Δ	-	-	-	+15.37%	-	-	+32.75%	-	-	+3.05%

exists between made-up physical meaning and realistic sensing from the physical world. However, this gap does not reflect in the perception subset, indicating that modality meanings are more specific and harder to imitate than their visual appearances. Nevertheless, the 2.7%-5.8% improvements still support that the model can learn from synthesized modality contexts to generalize to real visual modalities and keep the potential for completely unseen visual modalities.

4.4 Ablation Study

We also conduct ablation experiments to further observe the effectiveness of our training strategy. The first ablation demonstrates the effect of introducing modality context on real visual modalities. The results in Tab. 3 indicate that the modality context is important for modality understanding when encountering unseen modalities, as removing modality context leads to a 32.75% decrease at most. However, as the data scales of modern LMMs are continuously expanding, common non-RGB modalities like Thermal and X-ray are inevitably involved in their knowledge base. Thus, we observe a rising trend as the model size and date increase in Thermal and X-ray compared to Depth, even without modality context. Also, the results confirm that our training does not inject modality knowledge during fine-tuning; instead, improved modality-unaware perception brought an increase in modality understanding even without modality context.

Further ablation provided a result about the effectiveness of two aspects of VVM-tuning. In Tab. 4, we conduct an ablation study on our tuned 7B Qwen-2.5-VL model to explore the influence of modality-unaware perception tasks and modality-aware understanding tasks. As shown in Tab. 4, both parts are necessary for complete modality understanding. The modality-unaware perception tasks solely brought roughly half of the total improvement on both perception and understanding subsets. And the modality-aware understanding training further enhances the improvements on the basis of modality-unaware perception.

Table 4: Ablation results for the modality-unaware perception and modality-aware understanding in VVM-Tuning on Qwen 2.5-VL 7B model. "Perception" and "Understanding" represent modality-unaware perception tasks and modality-aware understanding tasks in VVM-Tuning, respectively. Instead, "P" and "U" represent the perception subset and understanding subset in VVM-Bench for evaluation. "Average" shows an average accuracy across all visual modalities. "✗" and "✔" indicate whether the corresponding task set is used or not used during training.

Perception	Understanding	Subset	Collected Visual Modalities			Fabricated Visual Modalities			Average
			Thermal	Depth	X-Ray	Optical Flow	Synthesized 1	Synthesized 2	
✗	✗	P	74.5%	73.3%	75.7%	77.0%	87.1%	81.1%	78.1%
		U	86.4%	83.4%	84.1%	84.6%	88.8%	89.8%	86.2%
✔	✗	P	75.3%	75.6%	79.6%	81.7%	90.7%	87.6%	81.7%
		U	86.7%	82.5%	86.3%	92.9%	90.3%	91.3%	88.3%
✗	✔	P	76.8%	74.1%	75.4%	74.1%	85.8%	81.7%	78.0%
		U	87.6%	85.4%	85.1%	92.1%	88.8%	90.3%	88.2%
✔	✔	P	79.8%	79.2%	79.6%	88.4%	91.2%	86.4%	84.1%
		U	88.5%	90.2%	85.9%	93.8%	91.9%	93.5%	90.6%

However, without modality-unaware tasks, the perception base is too weak to learn the adaptability of modality understanding for unseen visual modalities. Therefore, the performance nearly drops back to baseline.

5 Conclusion

Through our exploration of unseen visual modality generalization of LMMs, we have demonstrated that the disentangled alignments of semantics and photographic perception are effective as a generalization foundation. Via fabricating images, the modality-unaware tasks equip the model with unified perception across multiple real and synthesized visual modalities. Furthermore, we introduce modality context in the prompt and enable the adaptation to unseen visual modalities through modality synthesis. Our experiments have shown that the training framework is effective on multiple base models and helps them improve 8.2% and 3.8% performance on perception and understanding respectively without in-modality data. Therefore, the results confirmed that generalizing among versatile unseen visual modalities is possible for LMMs with enough perception and modality contexts. We also propose a benchmark to evaluate LMMs' general abilities on diverse visual modalities. We hope our efforts could facilitate the research on the visual modality adaptation of LMMs.

This attempt also gives us a new view of understanding the underlying mechanism of LMMs' vision. By distinguishing the differences and commonalities between visual modalities, we have the idea of separating shared semantics and photographic elements, which leads to a rethinking of machine vision: Do LMMs understand the difference between the image and the real world behind it? Or is its "real world" just the image? We wish this could inspire others to a deeper understanding of LMMs' visual abilities and continue to explore the boundaries of LMMs' vision.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2022YFA1004100 and the National Natural Science Foundation of China (NSFC) under Grant No. 62501191.

Besides, we express our gratitude to the anonymous reviewers for their invaluable suggestions and comments. We would like to thank all members of the ILL Lab for their helpful discussions and the computing resources supported by the ILL Lab, Faculty of Computing, HIT. Also, thanks to Sangyun Chung and other authors of VS-TDX for the accessibility of Non-RGB data.

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
5. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: SpatialBot: Precise Spatial Understanding with Vision Language Models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
7. Chung, S., Yu, Y., Kim, S.Y., Chee, Y., Ro, Y.M.: Enhanced Vision-Language Models for Diverse Sensor Understanding: Cost-Efficient Optimization and Benchmarking. arXiv preprint arXiv:2412.20750 (2024)
8. Deperrois, N., Matsuo, H., Ruipérez-Campillo, S., Vandenhirtz, M., Laguna, S., Ryser, A., Fujimoto, K., Nishio, M., Sutter, T.M., Vogt, J.E., et al.: RadVLM: A Multitask Conversational Vision-Language Model for Radiology. arXiv preprint arXiv:2502.03333 (2025)
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv preprint arXiv:2306.13394 (2023)
10. Fu, C., Zhang, Y.F., Yin, S., Li, B., Fang, X., Zhao, S., Duan, H., Sun, X., Liu, Z., Wang, L., et al.: MME-Survey: A Comprehensive Survey on Evaluation of Multimodal LLMs. arXiv preprint arXiv:2411.15296 (2024)
11. Gao, J., Sun, Y., Liu, Y., Tang, Y., Zeng, Y., Qi, D., Chen, K., Zhao, C.: StyleShot: A Snapshot on Any Style. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

12. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: ImageBind: One Embedding Space To Bind Them All. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
13. Hu, J., Peng, X., Xu, Z.: Study of Gray Image Pseudo-Color Processing algorithms. In: 6th International Symposium on Advanced Optical Manufacturing and Testing Technologies: Large Mirrors and Telescopes. vol. 8415, pp. 323–328. SPIE (2012)
14. Jiang, S., Chen, Z., Liang, J., Zhao, Y., Liu, M., Qin, B.: Infrared-LLaVA: Enhancing Understanding of Infrared Images in Multi-Modal Large Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 8573–8591 (2024)
15. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: BabyTalk: Understanding and Generating Simple Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2891–2903 (2013)
16. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-OneVision: Easy Visual Task Transfer (2024)
17. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
18. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: SEED-Bench: Benchmarking Multimodal Large Language Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. Advances in neural information processing systems **36** (2024)
20. 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. pp. 216–233. Springer (2024)
21. Lobry, S., Marcos, D., Murray, J., Tuia, D.: RSVQA: Visual Question Answering for Remote Sensing Data. IEEE Transactions on Geoscience and Remote Sensing **58**(12), 8555–8566 (2020)
22. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (2017)
23. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)
24. Mathew, M., Karatzas, D., Jawahar, C.: DocVQA: A Dataset for VQA on Document Images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2200–2209 (2021)
25. McKinley, S., Levine, M.: Cubic Spline Interpolation. College of the Redwoods **45**(1), 1049–1060 (1998)
26. Moshtaghi, M., Khajavi, S.H., Pajarinen, J.: RGB-Th-Bench: A Dense benchmark for Visual-Thermal Understanding of Vision Language Models. arXiv preprint arXiv:2503.19654 (2025)
27. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction Tuning with GPT-4. arXiv preprint arXiv:2304.03277 (2023)
28. Radewan, C.H.: Digital Image Processing With Pseudo-Color. In: Acquisition and Analysis of Pictorial Data. vol. 48, pp. 50–56. SPIE (1975)

29. Schoenberg, I.: Splines and Histograms. In: Spline Functions and Approximation Theory: Proceedings of the Symposium held at the University of Alberta, Edmonton May 29 to June 1, 1972. pp. 277–327. Springer (1973)
30. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: PandaGPT: One Model To Instruction-Follow Them All. In: Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23 (2023)
31. Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)
32. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.: XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. pp. 440–448 (2024)
33. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv preprint arXiv:2508.18265 (2025)
34. Xu, M., Yoon, S., Fuentes, A., Park, D.S.: A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. Pattern Recognition **137**, 109347 (2023)
35. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. arXiv preprint arXiv:2308.02490 (2023)
36. Yue, J., Zhang, S., Jia, Z., Xu, H., Han, Z., Liu, X., Wang, G.: MedSG-Bench: A Benchmark for Medical Image Sequences Grounding. arXiv preprint arXiv:2505.11852 (2025)
37. Yue, 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. pp. 9556–9567 (2024)
38. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–20 (2024)
39. Zhang, X., Bai, T., Li, H.: Pseudo-color coding method of infrared images based on human vision system. In: Infrared Materials, Devices, and Applications. vol. 6835, pp. 403–410. SPIE (2008)
40. Zhou, L., Hansen, C.D.: A Survey of Colormaps in Visualization. IEEE Transactions on Visualization and Computer Graphics **22**(8), 2051–2069 (2015)
41. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv preprint arXiv:2504.10479 (2025)