Words Over Pixels? Rethinking Vision in Multimodal Large Language Models

Anubhooti Jain , **Mayank Vatsa** and **Richa Singh**
IIT Jodhpur
{jain.44, mvatsa, richa}@iitj.ac.in

Abstract

Multimodal Large Language Models (MLLMs) promise seamless integration of vision and language understanding. However, despite their strong performance, recent studies reveal that MLLMs often fail to effectively utilize visual information, frequently relying on textual cues instead. This survey provides a comprehensive analysis of the vision component in MLLMs, covering both application-level and architectural aspects. We investigate critical challenges such as weak spatial reasoning, poor fine-grained visual perception, and suboptimal fusion of visual and textual modalities. Additionally, we explore limitations in current vision encoders, benchmark inconsistencies, and their implications for downstream tasks. By synthesizing recent advancements, we highlight key research opportunities to enhance visual understanding, improve cross-modal alignment, and develop more robust and efficient MLLMs. Our observations emphasize the urgent need to elevate vision to an equal footing with language, paving the path for more reliable and perceptually aware multimodal models.

1 Introduction

Large Vision Models (LVMs) and Large Language Models (LLMs) have made remarkable progress, achieving human-like performance across a wide range of complex tasks. Their success has led to the development of Multimodal Large Language Models (MLLMs), systems that seamlessly integrate vision and language reasoning, leading to powerful models such as GPT-4V [Achiam *et al.*, 2023], LLaVA [Liu *et al.*, 2023], and InternVL [Chen *et al.*, 2024e]. MLLMs are typically built on pretrained unimodal foundation models, leveraging well-performing LLMs like LLaMA [Touvron *et al.*, 2023] or Vicuna [Chiang *et al.*, 2023]. Their architecture generally consists of a vision encoder, an adapter (or connector) module, and an LLM, where the vision encoder and LLM remain frozen while the adapter is trained to bridge the gap between the two modalities. The image is processed by the vision encoder, passed through the projection module, and subsequently fused with language tokens before being fed into the LLM for final predictions. While various training paradigms and adapter strategies have been introduced to refine this pipeline and enhance MLLM capabilities, evaluating their true multimodal understanding remains a fundamental challenge in light of different tasks and applications.

In this paper, we critically examine the vision component of MLLMs from two key perspectives: application-level performance and architectural design. Visual understanding is essential for robust multimodal reasoning, yet current MLLMs exhibit significant deficiencies. As illustrated in Figure 1, their performance varies across different application categories. While these models excel in high-level reasoning tasks, they often struggle with low-level vision tasks, such as basic image classification, fine-grained object recognition, and direct reliance on visual cues for decision-making. Multiple studies highlight that vision remains the weaker modality in MLLMs. We explore these shortcomings and discuss existing benchmarks designed to evaluate visual understanding. While various benchmarks have been introduced to assess MLLM performance [Li and Lu, 2024], many fail to effectively evaluate individual model components, particularly from a vision-centric perspective. Most current evaluations prioritize high-level reasoning tasks while overlooking fundamental visual processing limitations.

From an architectural standpoint, effective representation learning and the fusion of visual and textual information remain open challenges. Different vision encoders process visual information in varying ways, leading to inconsistencies in multimodal alignment. We analyze the strengths and weaknesses of these encoders and their impact on MLLM performance. Finally, we identify key challenges and research opportunities, offering insights into bridging the existing gaps in visual understanding. By addressing these limitations, we aim to contribute to the development of more reliable, perceptually aware, and efficient MLLMs.

2 Rethinking Visual Understanding in MLLMs

For humans, vision and language are fundamental to perceiving and interacting with the world. This dual understanding has been replicated at scale in MLLMs, achieving impressive language-based reasoning. However, visual representations remain a major bottleneck in several aspects. At the application level, as illustrated in Figure 1, we classify current

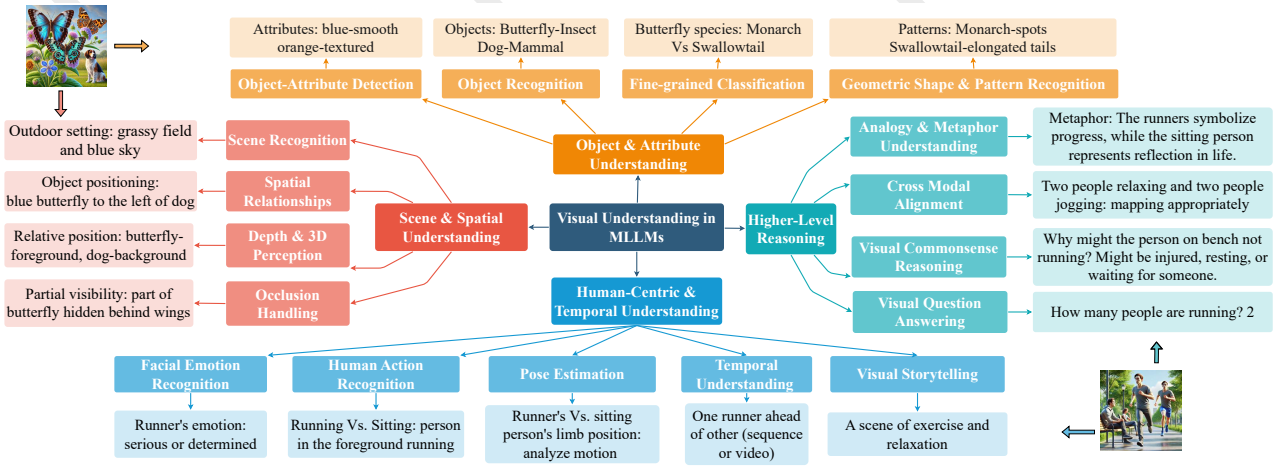

Figure 1: Visual Understanding in Multimodal LLMs.

visual-based tasks into four key categories that encompass the essential components of visual understanding in MLLMs.

Object and Attribute Understanding. This category includes tasks that require understanding object-attribute relationships, such as object detection, fine-grained classification, geometric shape recognition, and pattern recognition. An ideal MLLM should accurately distinguish visually similar objects and recognize multi-object relations, yet these remain significant challenges. Notably, MLLMs often confuse objects and their attributes, such as rigidly associating the color yellow with a banana, even when visual cues suggest otherwise. This over-reliance on learned correlations rather than actual visual perception hinders generalization. Additionally, a robust MLLM should be capable of detecting small objects and their attributes with equal proficiency, leveraging both image and text inputs effectively.

Scene and Spatial Understanding. Beyond object recognition, scene understanding involves grasping the relationships between objects and their environments. This includes scene recognition, spatial relationships, depth perception, occlusion handling, and 3D reasoning. Effective vision-language models must infer object positioning, recognize depth cues, separate foreground from background, and identify partially occluded objects. Furthermore, understanding functional relationships and contextual dependencies is crucial. However, MLLMs often struggle with fine-grained spatial reasoning, such as distinguishing whether an object is to the left or right of another. These limitations significantly impact their ability to reason about real-world spatial relationships and understand basic composition by extension.

Higher-Level Reasoning. Unlike spatial relationships, higher-level reasoning requires abstract, contextual, and cross-modal understanding. This includes visual question answering, visual commonsense reasoning, analogy and metaphor interpretation, and causal inference. The ability to explain events, infer logical relationships, and integrate both literal and symbolic visual cues is an essential goal for MLLM research. While current models perform comparatively well in these tasks, challenges persist in causal reasoning and commonsense inference, particularly in culturally diverse or context-dependent scenarios.

Human-Centric and Temporal Understanding. A critical aspect of multimodal intelligence is the ability to interpret human behavior. This category includes facial emotion recognition, human action recognition, pose estimation, temporal reasoning, and visual storytelling. Understanding moods, interactions, and evolving events over time is inherently complex, as interpretations vary across individuals, cultures, and social contexts. Moreover, MLLMs may inherit cultural and social biases, further complicating their ability to generalize across diverse scenarios.

Despite recent advancements, MLLMs continue to struggle with both low-level and high-level visual reasoning. Challenges persist in fine-grained object recognition, occlusion handling, attribute perception, depth estimation, and spatial alignment, as well as higher-level functions such as causal reasoning, emotional understanding, and social interactions. Visual cues are crucial for these tasks, yet MLLMs frequently overlook them, defaulting instead to language-based reasoning. Next, we examine these limitations in greater depth, with a particular focus on low-level visual functions.

2.1 Visual Understanding at Different Levels

The role of vision in MLLMs is often overlooked due to an implicit bias favoring prior knowledge from language, which is frequently sufficient for generating accurate responses. A study by [Chen *et al.*, 2024a] evaluated MLLMs by isolating the effects of language and vision in a limited data setting. The study examined visual understanding from three perspectives: visual processing, prior knowledge, and reasoning. An intriguing observation was that removing any language component significantly degrades performance, whereas removing the vision component still retains 75% of the model's performance. This finding highlights the current limitations of visual processing in MLLMs, particularly in tasks involving spatial understanding, object recognition, and fine-grained attribute detection.

Benchmark	Focus	Tasks
MME [Fu *et al.*, 2023]	Perception and Reasoning	Coarse-grained and Fine-grained Perception, OCR, and Visual Reasoning
MMStar [Chen *et al.*, 2024d]	Perception and Reasoning	Coarse-grained and Fine-grained Perception, Instance and Logical Reasoning, Mathematics, and Science
MMRel [Nie *et al.*, 2024]	Spatial and Temporal Understanding	Spatial, Action, and Comparative Reasoning
CVBench [Tong *et al.*, 2024a]	Spatial Understanding	Object Counting, Depth Order, and Relative Distance
BLINK [Fu *et al.*, 2024]	Visual Perception	Diverse Visual Prompting, Visual Commonsense, Perception beyond Recognition
CompBench [Kil *et al.*, 2024]	Relative Comparison	Visual Attribute, Existence, State, Emotion, Temporal, Spatial, Quantity, Quality
Q-Bench+ [Zhang *et al.*, 2024d]	Low-level Visual Perception and Understanding	Perception, Description, Image Quality (Single and Pairwise)
MagnifierBench [Li *et al.*, 2023]	Spatial Relations in High-Resolution Images	Object Localization, Counting, and Color
V* Bench [Wu and Xie, 2024]	Visual Grounding on High-Resolution Images	Attribute Recognition and Spatial Reasoning
P^2GB [Chen *et al.*, 2024b]	Comprehensive Image Understanding in High-Resolution Images	Text-Rich Visual Reasoning and Image Understanding
AesBench [Huang *et al.*, 2024]	Image Aesthetics Perception	Visual Perception, Empathy, Assessment, Interpretation
UNIAA [Zhou and others, 2024]	Image Aesthetics Perception	Content and Theme, Composition, Color, Light, Focus, and Sentiment
BlindTest [Rahmanzadehgervi *et al.*, 2024]	Geometric Perspective	Simple Tasks based on Common Geometric Primitives
CRPE [Wang *et al.*, 2024b]	Relation Comprehension	Existence, Subject, Predicate, and Object

Table 1: Benchmarks for visual understanding focusing on different visual capabilities with their associated tasks.

MLLMs are Blind (Spatial Understanding). MLLMs have demonstrated a notable reliance on textual priors, often producing consistent outputs even in the absence of visual input [Chen *et al.*, 2024a]. This suggests that their decisions are frequently uninfluenced by visual data. A study by [Tong *et al.*, 2024b] found that MLLMs perform poorly on basic visual patterns such as orientation, counting, and positional context, attributing these failures to weak visual representations. Further, models capable of multimodal generation have exhibited blindness to low-level visual features, as demonstrated by [Zheng *et al.*, 2024]. This blindness is largely due to vision encoders losing finer details during encoding.

Additionally, [Rahmanzadehgervi *et al.*, 2024] characterized MLLMs as effectively "blind" after they failed simple low-level visual tasks such as counting intersecting lines or circles—tasks trivial for humans with minimal world knowledge. The study suggested that late fusion of visual input might be responsible for this poor performance, as the models do not sufficiently integrate visual information. A more comprehensive spatial reasoning study by [Wang *et al.*, 2024a] compared MLLMs with unimodal LLM counterparts using vision-only, text-only, and vision-text-based inputs on maze and map navigation tasks. Surprisingly, the models performed better with text descriptions alone than with text combined with visual inputs. Adding noisy or mismatched images further did not significantly alter their performance, reinforcing the argument that visual information plays a minimal role in decision-making.

Looking Deeper (Fine-Grained Recognition). MLLMs have struggled with even basic vision tasks, such as image classification, in both closed and open-world settings [Zhang *et al.*, 2024b]. While encoded visual information may exist within the model, it often cannot be effectively decoded for classification tasks. Fine-tuning with additional fine-grained class samples has been shown to improve performance, but the baseline remains weak. A similar trend was observed in fine-grained object recognition, where [Chandhok *et al.*, 2024] identified vision encoders as the primary source of failure. The study noted that visual information loss during encoding results in poor spatial understanding, though some of these shortcomings can be partially mitigated using language priors. Further, MLLMs consistently struggle with vision-dependent queries, rarely outperforming unimodal LLMs when spatial reasoning is required. Recently, Fine-grained Language-informed Image Representations (FLAIR) [Xiao *et al.*, 2024] was introduced to enhance fine-grained alignment between text and images by learning localized image embeddings alongside global vision-text representations. However, its training was conducted on relatively smaller datasets, limiting its scalability.

Other Tasks and Vulnerabilities. MLLMs exhibit further limitations in various vision-centric tasks. A study by [Zhai *et al.*, 2023b] examined catastrophic forgetting in MLLMs, specifically analyzing whether vision encoders retain their standalone classification abilities after integration into multimodal architectures. The results indicated that MLLMs failed to maintain the same classification performance as their pretrained vision encoders, suggesting that multimodal alignment may degrade visual capabilities. Additionally, in-context learning (ICL) has been analyzed in multimodal set-

tings, where textual information was found to be far more significant than visual input [Chen *et al.*, 2023]. This aligns with previous findings that removing or corrupting images has minimal impact on model performance.

These visual vulnerabilities have also been exploited for adversarial attacks and jailbreaking of MLLMs. A study by [Li *et al.*, 2024c] found that visual inputs can act as alignment backdoors, increasing multimodal attack surfaces. Furthermore, weak vision models have been linked to hallucinations in MLLMs, as highlighted by [Bai *et al.*, 2024]. This phenomenon arises from information loss during visual encoding, where weak perception leads to incorrect visual inferences. Additionally, poor alignment between weak vision encoders and powerful language models can exacerbate hallucinations, further reducing multimodal reliability.

From losing fine-grained information during encoding to ineffective integration and suboptimal decoding of visual information, every step contributes to vision modality lagging and needs to be addressed.

2.2 Benchmarks for Visual Understanding

A variety of benchmarks have been proposed to evaluate MLLMs across multiple capabilities. However, many of these fail to effectively assess the visual competence of these models. For instance, Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) [Yue *et al.*, 2024] and MathVista [Lu *et al.*, 2024] benchmarks exhibited less than a 5% performance gap between multimodal and language-only settings [Tong *et al.*, 2024a], indicating that textual cues alone are often sufficient for solving benchmark tasks. This suggests that these benchmarks do not sufficiently rely on visual input and thus fail to measure true vision-language integration. In response, CV-Bench was introduced as a vision-centric benchmark, specifically targeting spatial relationships, object counting, depth ordering, and relative distance. This benchmark curates samples from existing datasets to provide a more fine-tuned evaluation of visual-spatial reasoning in MLLMs.

Other notable benchmarks include MLLM Evaluation (MME) [Fu *et al.*, 2023], which evaluates coarse-grained and fine-grained perception, Optical Character Recognition (OCR) capabilities, and visual reasoning. It found that top-performing MLLMs struggle with object perception, instruction following, and reasoning, often exhibiting hallucinations. Additionally, [Tong *et al.*, 2024b] proposed Multimodal Visual Patterns (MMVP), leveraging Contrastive Language-Image Pre-training (CLIP) embeddings [Radford *et al.*, 2021] to create blind pairs. Pairs that include images that are visually different but appear similar in the CLIP embedding space. This exposed MLLMs' weaknesses in evaluating nine fundamental visual patterns, including orientation, counting, positional context, and color understanding.

Furthermore, [Chen *et al.*, 2024d] observed that visual content is often unnecessary for answering multimodal queries correctly. They also identified unintentional data leakage in training datasets, allowing models to answer image-dependent questions correctly without actually processing visual content. To mitigate this, they proposed MMStar, a benchmark that explicitly ensures visual dependency while minimizing data leakage. These benchmarks highlight the limitations of current evaluation frameworks and emphasize the need for more vision-centric assessments. A summary of key benchmarks is provided in Table 1.

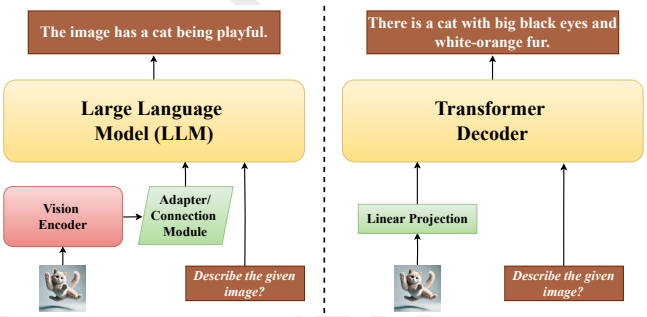

Figure 2: Generic architecture overview of MLLMs. Left: MLLMs with a separate vision encoder and adapter like LLaVA; Right: MLLMs without a separate vision encoder like Fuyu.

3 Dissecting MLLMs: Vision Encoders, Representation, and Fusion

We have discussed the challenges MLLMs face with visual understanding, categorizing them into four key areas. At the architectural level (see Figure 2), two critical components contribute to these issues: the vision encoder and the fusion mechanism. The vision encoder determines how images are processed and encoded into visual representations, while the fusion mechanism dictates how this encoded visual information is integrated with textual inputs. In this section, we examine both components in depth.

3.1 Exploring the Vision Encoders in MLLMs

Vision encoders are expected to play an equal role alongside language models in MLLMs. However, they often underperform, exhibiting issues such as information loss and weak visual semantic reasoning. Most MLLMs use pretrained vision transformers (ViTs) [Dosovitskiy *et al.*, 2021] as vision encoders, particularly ViTs from Contrastive Language-Image Pre-training (CLIP) models, which have become the de facto choice. However, as we discuss in this section, CLIP-based encoders introduce their own set of challenges, prompting exploration into alternative encoders.

CLIP—the Preferred Vision Encoder. CLIP [Radford *et al.*, 2021] employs a contrastive learning objective to align vision and language encoders in a shared representation space. Its ability to effectively align these modalities has made CLIP's vision encoder a popular choice for MLLMs. Several MLLMs use variations of CLIP's ViT, such as ViT-G from EVA-CLIP [Sun *et al.*, 2023], which strengthens the vision backbone, and SigLIP [Zhai *et al.*, 2023a], which incorporates a sigmoid-based loss to improve zero-shot capabilities. These encoders, trained to align with text, enable MLLMs to share a representation space with LLMs, facilitating multimodal learning. However, compared to commonly used LLMs such as LLaMA (which ranges from 7 to 65 billion parameters), ViTs are significantly smaller, with ViT-G

having only around 1 billion parameters.

Recent studies have identified a modality gap within CLIP, which impacts its effectiveness in certain tasks. While this gap [Liang *et al.*, 2022] helps preserve performance in zero-shot recognition, it also contributes to poor fine-grained attribute recognition. Research by [Schrodi and others, 2024] attributes this gap to information imbalance, as text captions contain less information than images. The study further highlights that reducing the modality gap is not always beneficial, emphasizing the need for a balanced trade-off between textual and visual information.

Beyond CLIP—Alternative Vision Encoders. Several studies have explored alternative vision encoders beyond CLIP [Jiang *et al.*, 2023]. These encoders can be supervised (like CLIP) or self-supervised (such as DINO [Oquab *et al.*, 2024]). More recently, [Fan *et al.*, 2025] explored the scalability of self-supervised learning (SSL) models such as DINO. They found that SSL not only performs well on visual tasks when compared with language-supervised models like CLIP but can match the latter's performance when scaled. The gains are evident across a diverse set of Visual Question Answering (VQA) tasks involving general knowledge, world knowledge, OCR, chart understanding, and vision-centric datasets. These SSL models were scaled in both model size and training data volume, affirming the potential of using such models as vision encoders in MLLMs to enhance overall performance. Different encoder architectures have also been explored, including convolution-based, generative, and self-supervised encoders, as well as hybrid models. For instance, ConvNeXT [Liu *et al.*, 2022], a convolution-based encoder, has been used with approximately 198 million parameters.

Research by [Wang *et al.*, 2023] examined the science behind selecting an effective visual tokenizer or encoder, finding that self-supervised encoders offer better fine-grained perception, whereas supervised encoders excel in semantic understanding. This suggests a necessary trade-off between the two approaches to optimize vision encoders for MLLMs. Another study [Shi *et al.*, 2024] demonstrated that unfreezing the vision encoder can significantly improve MLLM performance, particularly for high-resolution image comprehension. However, this approach comes with a high computational cost. Unfreezing the encoder allows for better alignment of the vision modality with the LLM, but the trade-off in efficiency must be carefully considered.

Several studies highlight that traditional vision encoders lack effective pixel-level and object-level understanding [Zhang *et al.*, 2024a]. To address this, OMG-LLaVA [Zhang *et al.*, 2024a] introduced a novel multimodal model, leveraging a universal perception model, OMG-Seg [Li *et al.*, 2024b]. Similarly, Libra [Xu *et al.*, 2024] proposed an ideal vision system, arguing that an effective vision encoder should be somewhat independent from the LLM to enhance visual understanding and cross-modal interaction. The choice of an optimal vision encoder depends on the specific end task. For example, MLLMs designed for text-rich environments would benefit more from vision encoders specialized in OCR recognition and panoptic segmentation. However, specialized MLLMs can be computationally expensive, making efficient adaptation crucial.

Mixture-of-Encoders—Combining Different Vision Models. An alternative strategy to enhance vision encoding is to combine multiple encoders, as different encoders capture distinct relationships more effectively. Using a mixture-of-encoders within a mixture-of-experts framework [Shi *et al.*, 2024; He and others, 2024] allows for richer, globally and locally aware image embeddings. This approach can improve vision-language alignment by integrating both fine-grained object relations and coarse-level semantic understanding, contributing to a more unified multimodal model.

The Incorporating Visual Experts (IVE) framework [He and others, 2024] demonstrated the effectiveness of using multiple task-specific vision encoders, each specializing in semantics, low-level visual features, and document-related information. More recently, EAGLE [Shi *et al.*, 2024] introduced a hybrid encoder approach, leveraging: CLIP for image-text matching, ConvNeXT for image classification, EVA for object detection, Pix2Struct for text recognition, SAM for image segmentation, and DINO for self-supervised feature learning.

However, integrating multiple encoders poses challenges, as they may generate conflicting representations, leading to inconsistent predictions and redundant feature fusion. Despite their advantages, mixture-of-encoders approaches come at a high computational cost and introduce complex feature fusion challenges.

3.2 Visual Representation and Input Tokens

Encoded visual tokens play a crucial role in MLLM performance. The resolution of images and the number of visual tokens significantly impact the efficiency and accuracy of the model [McKinzie *et al.*, 2024]. Handling multiple resolutions increases computational overhead due to the larger number of tokens, making efficient token management essential.

To address this, DeepStack [Meng and others, 2024] restructures visual tokens into a stacked architecture, where each layer is directly connected to corresponding LLM layers. This approach reduces computational overhead, enabling more effective handling of high-resolution images. Similarly, OtterHD [Li *et al.*, 2023] moves away from fixed-resolution MLLMs, enabling models to process images at various resolutions dynamically. More recently, LLaVA-OneVision [Li *et al.*, 2024a] introduced the AnyRes strategy, allowing processing of both high-resolution images and video frames, striking a balance between performance and computational cost.

Another approach to improving visual token formulation is augmenting tokens with extra perceptual information. Perception tokens were introduced in [Bigverdi *et al.*, 2024], supplementing standard tokens with depth maps, bounding box coordinates, and spatial information. These additional tokens, derived from an encoder-decoder auxiliary model, significantly enhance spatial reasoning tasks. However, such models require fine-tuning to interpret perception tokens effectively, increasing training complexity.

A persistent challenge with visual tokens is their higher computational footprint compared to text tokens. Pruning strategies have been explored to reduce unnecessary visual to-

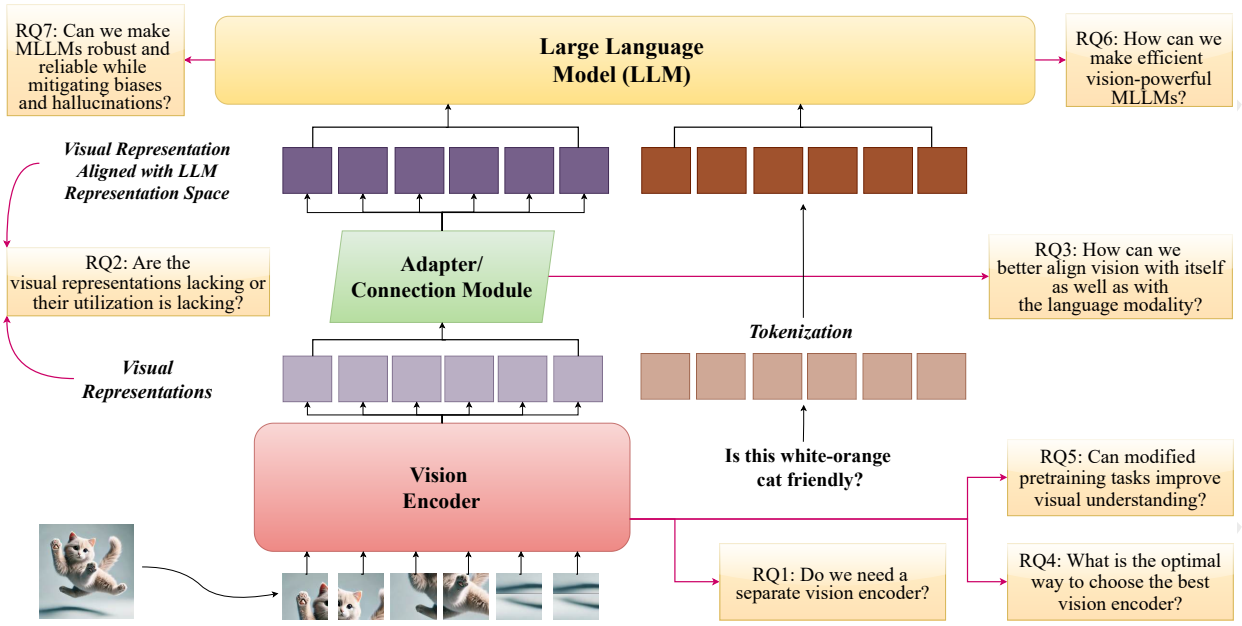

Figure 3: Possible research opportunities framed as research questions for MLLMs from a vision perspective.

kens, but determining the most relevant tokens remains data-dependent. Instead of pruning tokens directly, [Zhang *et al.*, 2024c] proposed a structural pruning approach, where inactive attention heads and redundant transformer layers are selectively removed. This method focuses on optimizing model architecture rather than filtering data tokens, ensuring more effective use of computational resources.

4 Challenges and Research Opportunities

Despite recent advancements, MLLMs continue to face significant challenges in visual understanding, both from an application-based and architectural perspective. The former includes tasks spanning from low-level recognition to high-level reasoning, while the latter concerns the limitations of vision encoders and their alignment with LLMs. Even though vision encoders are powerful, their full capabilities are often not retained fully when integrated with LLMs. Weak vision encoders, suboptimal cross-modal alignment, and poor fusion strategies collectively hinder the model's ability to effectively process and utilize visual information.

It is evident that even state-of-the-art MLLMs exhibit substantial gaps in visual perception, leaving room for improvement. To advance the field, we highlight seven key research questions, illustrated in Figure 3, that present future opportunities for enhancing MLLMs.

RQ1: Do We Need a Separate Vision Encoder? Most MLLMs employ a separate vision encoder with a connector module that projects visual features into an LLM's representation space (see Figure 2). However, is this architectural separation necessary? Models like Fuyu [Bavishi *et al.*, 2023] have explored a single-decoder approach, demonstrating that multimodal fusion can be achieved without relying on a multi-stage, multi-encoder setup. While such approaches still need further scaling and benchmarking, they offer a potential path toward unified multimodal processing. A unified encoder-decoder architecture could ensure that vision and language modalities are treated equally, potentially improving fine-grained visual recognition. An alternative approach to unifying vision and text was proposed in LaViT [Jin *et al.*, 2024a], where a visual tokenizer translates image tokens into an LLM-understandable format, treating them like a foreign language. This framework dynamically adjusts token length based on sparsity and interdependence between image patches, ensuring more efficient processing. However, the impact of such tokenization strategies on complex vision-language reasoning tasks remains an open question.

RQ2: Are the Visual Representations Weak, or is Their Utilization Lacking? While MLLMs are often criticized for lacking visual understanding, is the problem truly with their visual representations? Studies have shown that well-trained vision and language encoders exhibit high semantic similarity [Maniparambil *et al.*, 2024]. This suggests that rather than being inherently weak, visual representations may not be optimally utilized or aligned. A key issue is that current architectures attempt to translate visual context into the language space, rather than fully integrating both modalities. Exploring new visual representation strategies tailored for LLMs could help enhance visual comprehension. For example, [Zhong *et al.*, 2024] proposed the visual table, a hierarchical text-based representation of semantic scene elements designed to improve visual reasoning. While effective, generating such representations introduces computational overhead and requires information-rich structured data. Scalability for such approaches is an open challenge.

RQ3: How Can We Better Align Vision With Itself and With the Language Modality? Alignment issues persist at

multiple levels within MLLMs. One major challenge is cognitive misalignment, where ambiguous visual representations prevent LLMs from accurately interpreting images. As explored by [Zhao *et al.*, 2024], introducing rich and structured representations can help mitigate this misalignment, although such methods often require labeled supervision. At the same time, perceptual misalignment within vision encoders can negatively affect multimodal fusion. A study by [Sundaram *et al.*, 2024] demonstrated that leveraging human perception knowledge can enhance visual representations, making them more general-purpose. However, effectively integrating such representations with LLMs remains an open challenge.

RQ4: What is the Optimal Strategy for Selecting the Best Vision Encoder? Rather than dismissing a vision encoder as inherently weak, selecting the right encoder for the task is key to building robust multimodal models. Among various pretrained encoders, the AC policy [Yang *et al.*, 2024] was proposed as a method for selecting an optimal vision encoder by leveraging CLIP as a reference model. It demonstrated a weak correlation with OCR-based tasks but a strong link between cross-modal alignment and overall model performance. This suggests that a data-driven approach for selecting or adapting encoders, rather than relying on fixed architectures, could significantly enhance MLLM performance. Additionally, research into combining multiple encoders or dynamically adjusting them based on the task remains an area of opportunity.

RQ5: Can Modified Pretraining Tasks Improve Visual Understanding? Modifying pretraining objectives could help improve visual representations and alignment within MLLMs. Lyrics [Lu *et al.*, 2023] introduced multi-task pretraining, incorporating objectives like image-text contrastive learning, image-grounded captioning, and masked spatial prediction. This framework aimed to align both fine-grained and coarse-grained visual features with text, enhancing multimodal reasoning. A novel task, Pixel Value Prediction (PVP) [Gou *et al.*, 2024], was proposed to predict RGB values at specific image coordinates. Adapting the vision encoder with PVP during training has been empirically shown to improve perception capabilities. The integration of vision-centric or alignment-specific pretraining remains a promising direction for improving MLLMs' fine-grained recognition and spatial reasoning abilities.

RQ6: How Can We Develop Efficient Yet Vision-Powerful MLLMs? Scalability and efficiency are crucial for practical deployment, like scaling SSL models [Fan *et al.*, 2025]. On the other hand, scaling vision encoders might seem like a natural solution, but it is not always the most efficient approach. Instead, better token selection strategies can help balance computational cost and performance. Efficient MLLM architectures were explored in [Jin *et al.*, 2024b], emphasizing the need for lightweight, generalizable models that maintain high visual fidelity. One such model, EVLM [Chen *et al.*, 2024c], employed a mixture-of-experts mechanism and hierarchical visual feature modeling to achieve improved visual comprehension while maintaining efficiency. Developing lightweight yet visually powerful MLLMs remains an active area of research.

RQ7: How Can We Make MLLMs More Robust and Reliable While Mitigating Hallucinations and Biases? The dominance of language in MLLMs can lead to over-reliance on text priors, introducing biases and hallucinations that compromise model robustness. We have already discussed how hallucinations can emerge from visual shortcomings in these models. These shortcomings can also leave the model vulnerable to jailbreaks [Wang and others, 2024]. A key research direction is designing methods to systematically analyze and mitigate such biases. One approach is to conduct a trustworthiness analysis, systematically evaluating how each component of an MLLM contributes to bias and hallucination risks [Liu and others, 2024; Mittal *et al.*, 2024a]. A deeper investigation into error attribution within multimodal architectures could enable targeted improvements, leading to more reliable and fair multimodal AI systems. The models as well as datasets [Mittal *et al.*, 2024b] also need to be systematically reviewed for ethical considerations.

Despite noted shortcomings discussed above, empirical investigation of MLLMs remains limited, offering a promising direction for future research.

5 Conclusion

Multimodal Large Language Models (MLLMs) have significantly advanced in recent years, demonstrating strong performance in high-level reasoning tasks. However, they continue to struggle with low-level and fine-grained visual tasks. Benchmarks consistently highlight their shortcomings in visual-spatial reasoning, text-rich image comprehension, fine-grained object recognition, and basic geometric and classification tasks. While CLIP-based vision encoders remain the preferred choice for many MLLMs, research has explored a variety of alternative encoders to enhance visual understanding. This includes Mixture of Experts (MoE) mechanisms, which allow multiple task-specific vision encoders to be utilized within a single model.

Despite these advancements, the interaction between vision-encoded tokens and textual representations remains a critical bottleneck. The connector module, which facilitates this interaction, has been frequently identified as a source of weak alignment, contributing to poor visual capabilities and information loss. Recent works have proposed solutions such as hierarchical feature representations, transformer-based projection layers, and optimized fusion strategies to address these challenges. However, visual understanding in MLLMs remains an open problem, with lingering issues such as hallucinations, misalignment, and susceptibility to jailbreak attacks.

Despite ongoing research efforts, there is still a pressing need to bridge the gap between vision and language modalities. Addressing these challenges will be essential in building more reliable, perceptually aware, and robust MLLMs that can truly integrate and understand multimodal information.

Acknowledgments

This work is supported by Srijan: Center of Excellence on GenAI at IIT Jodhpur, India, IndiaAI Mission, and Meta.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

[Bai *et al.*, 2024] Zechen Bai, Pichao Wang, et al. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930, 2024.

[Bavishi *et al.*, 2023] Rohan Bavishi, Erich Elsen, et al. Introducing our multimodal models. 2023.

[Bigverdi *et al.*, 2024] Mahtab Bigverdi, Zelun Luo, et al. Perception tokens enhance visual reasoning in multimodal language models. *CoRR*, abs/2412.03548, 2024.

[Chandhok *et al.*, 2024] Shivam Chandhok, Wan-Cyuan Fan, and Leonid Sigal. Response wide shut: Surprising observations in basic vision language model capabilities. *CoRR*, abs/2408.06721, 2024.

[Chen *et al.*, 2023] Shuo Chen, Zhen Han, et al. Understanding and improving in-context learning on vision-language models. *CoRR*, abs/2311.18021, 2023.

[Chen *et al.*, 2024a] Allison Chen, Ilia Sucholutsky, et al. Analyzing the roles of language and vision in learning from limited data. *CoRR*, abs/2403.19669, 2024.

[Chen *et al.*, 2024b] Jiaxing Chen, Yuxuan Liu, et al. Plug-and-play grounding of reasoning in multimodal large language models. *CoRR*, abs/2403.19322, 2024.

[Chen *et al.*, 2024c] Kaibing Chen, Dong Shen, et al. EVLM: an efficient vision-language model for visual understanding. *CoRR*, abs/2407.14177, 2024.

[Chen *et al.*, 2024d] Lin Chen, Jinsong Li, et al. Are we on the right way for evaluating large vision-language models? *CoRR*, abs/2403.20330, 2024.

[Chen *et al.*, 2024e] Zhe Chen, Jiannan Wu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org, 2023. Accessed: 2025-02-01.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[Fan *et al.*, 2025] David Fan, Shengbang Tong, et al. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

[Fu *et al.*, 2024] Xingyu Fu, Yushi Hu, et al. BLINK: multimodal large language models can see but not perceive. *CoRR*, abs/2404.12390, 2024.

[Gou *et al.*, 2024] Chenhui Gou, Abdulwahab Felemban, et al. How well can vision language models see image details? *CoRR*, abs/2408.03940, 2024.

[He and others, 2024] Xin He et al. Incorporating visual experts to resolve the information loss in multimodal large language models. *CoRR*, abs/2401.03105, 2024.

[Huang *et al.*, 2024] Yipo Huang, Quan Yuan, et al. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *CoRR*, abs/2401.08276, 2024.

[Jiang *et al.*, 2023] Dongsheng Jiang, Yuchen Liu, et al. From clip to dino: Visual encoders shout in multi-modal large language models. 2023.

[Jin *et al.*, 2024a] Yang Jin, Kun Xu, et al. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *ICLR*, 2024.

[Jin *et al.*, 2024b] Yizhang Jin, Jian Li, et al. Efficient multimodal large language models: A survey. *CoRR*, abs/2405.10739, 2024.

[Kil *et al.*, 2024] Jihyung Kil, Zheda Mai, et al. Compbench: A comparative reasoning benchmark for multimodal llms. *CoRR*, abs/2407.16837, 2024.

[Li and Lu, 2024] Jian Li and Weiheng Lu. A survey on benchmarks of multimodal large language models. *CoRR*, abs/2408.08632, 2024.

[Li *et al.*, 2023] Bo Li, Peiyuan Zhang, et al. Otterhd: A high-resolution multi-modality model. *CoRR*, abs/2311.04219, 2023.

[Li *et al.*, 2024a] Bo Li, Yuanhan Zhang, et al. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326, 2024.

[Li *et al.*, 2024b] Xiangtai Li, Haobo Yuan, et al. Omg-seg: Is one model good enough for all segmentation? *CoRR*, abs/2401.10229, 2024.

[Li *et al.*, 2024c] Yifan Li, Hangyu Guo, et al. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *CoRR*, abs/2403.09792, 2024.

[Liang *et al.*, 2022] Weixin Liang, Yuhui Zhang, et al. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

[Liu and others, 2024] Xin Liu et al. Safety of multimodal large language models on images and text. In *IJCAI*, 2024.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, et al. A convnet for the 2020s. In *CVPR*, 2022.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, et al. Visual instruction tuning. In *NeurIPS*, 2023.

[Lu *et al.*, 2023] Junyu Lu, Ruyi Gan, et al. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *CoRR*, abs/2312.05278, 2023.

[Lu *et al.*, 2024] Pan Lu, Hritik Bansal, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

[Maniparambil *et al.*, 2024] Mayug Maniparambil, Raiymbek Akshulakov, et al. Do vision and language encoders represent the world similarly? In *CVPR*, 2024.

[McKinzie *et al.*, 2024] Brandon McKinzie, Zhe Gan, et al. MM1: methods, analysis & insights from multimodal LLM pre-training. *CoRR*, abs/2403.09611, 2024.

[Meng and others, 2024] Lingchen Meng et al. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *CoRR*, abs/2406.04334, 2024.

[Mittal *et al.*, 2024a] Surbhi Mittal, Arnav Sudan, Mayank Vatsa, et al. Navigating text-to-image generative bias across indic languages. In *ECCV*, 2024.

[Mittal *et al.*, 2024b] Surbhi Mittal, Kartik Thakral, et al. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. *Nat. Mac. Intell.*, 2024.

[Nie *et al.*, 2024] Jiahao Nie, Gongjie Zhang, et al. Mmrel: A relation understanding dataset and benchmark in the MLLM era. *CoRR*, abs/2406.09121, 2024.

[Oquab *et al.*, 2024] Maxime Oquab, Timothée Darcet, et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Rahmanzadehgervi *et al.*, 2024] Pooyan Rahmanzadehgervi, Logan Bolton, et al. Vision language models are blind. *CoRR*, abs/2407.06581, 2024.

[Schrodi and others, 2024] Simon Schrodi et al. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. *CoRR*, abs/2404.07983, 2024.

[Shi *et al.*, 2024] Min Shi, Fuxiao Liu, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *CoRR*, abs/2408.15998, 2024.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, et al. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023.

[Sundaram *et al.*, 2024] Shobhita Sundaram, Stephanie Fu, et al. When does perceptual alignment benefit vision representations? In *NeurIPS*, 2024.

[Tong *et al.*, 2024a] Shengbang Tong, Ellis Brown, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *CoRR*, abs/2406.16860, 2024.

[Tong *et al.*, 2024b] Shengbang Tong, Zhuang Liu, et al. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, et al. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[Wang and others, 2024] Taowen Wang et al. Large vision-language model security: A survey. In *FCS*, 2024.

[Wang *et al.*, 2023] Guangzhi Wang, Yixiao Ge, et al. What makes for good visual tokenizers for large language models? *CoRR*, abs/2305.12223, 2023.

[Wang *et al.*, 2024a] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, and Neel Joshi. Is A picture worth A thousand words? delving into spatial reasoning for vision language models. *CoRR*, abs/2406.14852, 2024.

[Wang *et al.*, 2024b] Weiyun Wang, Yiming Ren, et al. The all-seeing project V2: towards general relation comprehension of the open world. *CoRR*, abs/2402.19474, 2024.

[Wu and Xie, 2024] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

[Xiao *et al.*, 2024] Rui Xiao, Sanghwan Kim, et al. FLAIR: VLM with fine-grained language-informed image representations. *CoRR*, abs/2412.03561, 2024.

[Xu *et al.*, 2024] Yifan Xu, Xiaoshan Yang, Yaguang Song, and Changsheng Xu. Libra: Building decoupled vision system on large language models. In *ICML*, 2024.

[Yang *et al.*, 2024] Shijia Yang, Bohan Zhai, Quanzeng You, et al. Law of vision representation in mllms. *CoRR*, abs/2408.16357, 2024.

[Yue *et al.*, 2024] Xiang Yue, Yuansheng Ni, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

[Zhai *et al.*, 2023a] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[Zhai *et al.*, 2023b] Yuexiang Zhai, Shengbang Tong, et al. Investigating the catastrophic forgetting in multimodal large language models. *CoRR*, abs/2309.10313, 2023.

[Zhang *et al.*, 2024a] Tao Zhang, Xiangtai Li, et al. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *CoRR*, abs/2406.19389, 2024.

[Zhang *et al.*, 2024b] Yuhui Zhang, Alyssa Unell, et al. Why are visually-grounded language models bad at image classification? *CoRR*, abs/2405.18415, 2024.

[Zhang *et al.*, 2024c] Zeliang Zhang, Phu Pham, et al. Treat visual tokens as text? but your mllm only needs fewer efforts to see. *arXiv preprint arXiv:2410.06169*, 2024.

[Zhang *et al.*, 2024d] Zicheng Zhang, Haoning Wu, et al. A benchmark for multi-modal foundation models on low-level vision: from single images to pairs. *CoRR*, abs/2402.07116, 2024.

[Zhao *et al.*, 2024] Yaqi Zhao, Yuanyang Yin, et al. Beyond sight: Towards cognitive alignment in LVLM via enriched visual knowledge. *CoRR*, abs/2411.16824, 2024.

[Zheng *et al.*, 2024] Boyang Zheng, Jinjin Gu, Shijun Li, and Chao Dong. LM4LV: A frozen large language model for low-level vision tasks. *CoRR*, abs/2405.15734, 2024.

[Zhong *et al.*, 2024] Yiwu Zhong, Zi-Yuan Hu, et al. Beyond embeddings: The promise of visual table in visual reasoning. In *EMNLP*, 2024.

[Zhou and others, 2024] Zhaokun Zhou et al. UNIAA: A unified multi-modal image aesthetic assessment baseline and benchmark. *CoRR*, abs/2404.09619, 2024.