International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Vision LLMs: Bridging Language and Visual Understanding - Case study of IndoAI

Vivek Gujar

Founder Director, IndoAI Technologies P Ltd, Pune, India
Email: *vivek [at]indo.ai*

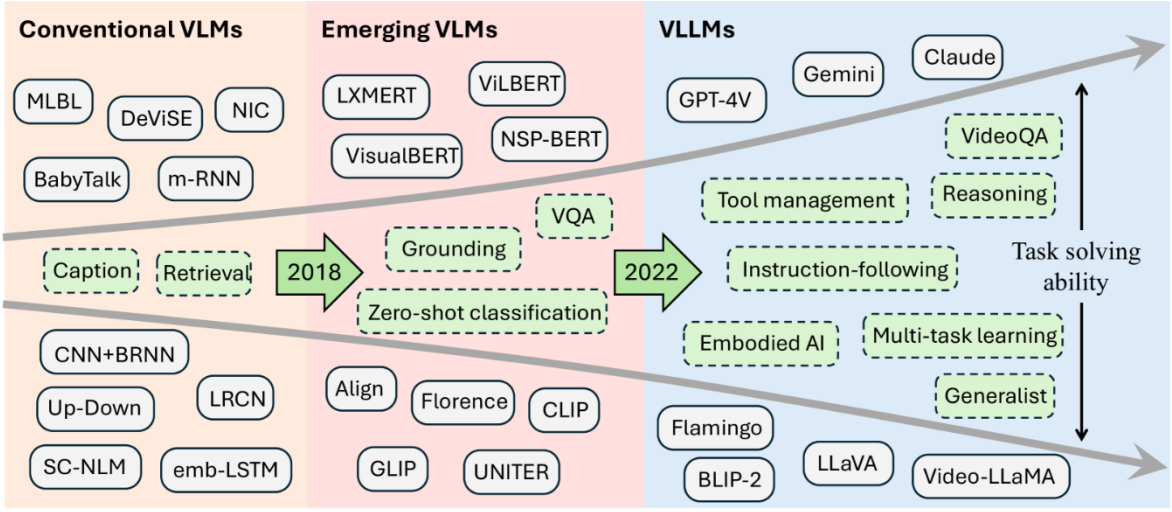
Abstract: ***Vision LLMs are trained on vast datasets containing paired image-text samples, allowing them to perform tasks such as image captioning, visual question answering (VQA) and multimodal reasoning. These Models (Vision LLMs) mark a transformative leap in artificial intelligence by merging visual and linguistic understanding, enabling seamless human-machine communication, power groundbreaking applications-from automated diagnostic reporting in healthcare to real-time scene analysis in autonomous systems. Yet, key challenges remain, including computational inefficiency, embedded biases in training data and limited interpretability which currently restrict broader deployment. Cutting-edge research is tackling these obstacles through optimized model architectures, fairness-aware dataset curation and advanced explainable AI methods. As these advancements progress, Vision LLMs are poised to revolutionize AI-driven solutions across industries such as healthcare, robotics, autonomous vehicles. Their continued evolution is redefining the landscape of interdisciplinary AI, fostering more intuitive, ethical and scalable intelligent systems. This article provides an overview of Vision LLM architectures, their applications and the challenges they face and case study of how building of AI Models through visionLLM may help IndoAI AI camera system.***

Keywords: vision LLM, LAION, NLP, LLM, GPT, IndoAI, AI Camera

1. Introduction

The rapid advancement of artificial intelligence has significantly transformed several key domains, including computer vision, natural language processing and geospatial applications [1] [2]). The emergence of Large Language Models (LLMs) has altered the course of the AI revolution and its integration with Computer Vision is teaching enterprise AI how to both see and speak [3]. Prior to 2023, large language models (LLMs) such as GPT-3 and LLaMA operated exclusively on textual data, require separate computer vision systems for any image processing tasks, it could not analyze images, videos directly or interpret visual context without human-provided descriptions and unable to perform tasks requiring spatial reasoning (e.g., object detection). These constraints significantly restricted the real-world applicability of even advanced LLMs like GPT-4 [4] and LLaMA [5], despite their remarkable performance on text-based tasks and also this division created substantial challenges in developing truly integrated multimodal AI systems. Visual-language models [6] have emerged as a powerful tool for learning a unified embedding space for vision and language. The emergence of Vision-Language Models (VLMs) has begun to address these limitations by integrating convolutional neural networks (CNNs) or vision transformers (ViTs) with traditional LLM architectures [7].

Below figure [7] of evolution of VLMs includes three phases: conventional VLMs (before 2018), emerging VLMs (2018-2022), and VLLMs (2022 onward).



To address the limitations in human-like cognitive processing, researchers [8] have developed Vision-Language Models (VLMs) - an advanced class of neural networks that achieve multimodal understanding by jointly processing visual and textual data. These models demonstrate exceptional capabilities in cross-modal comprehension and generation, enabling sophisticated functionality across key tasks including visual question answering, image captioning, and text-to-image generation. By effectively bridging visual and linguistic modalities, VLMs represent a significant

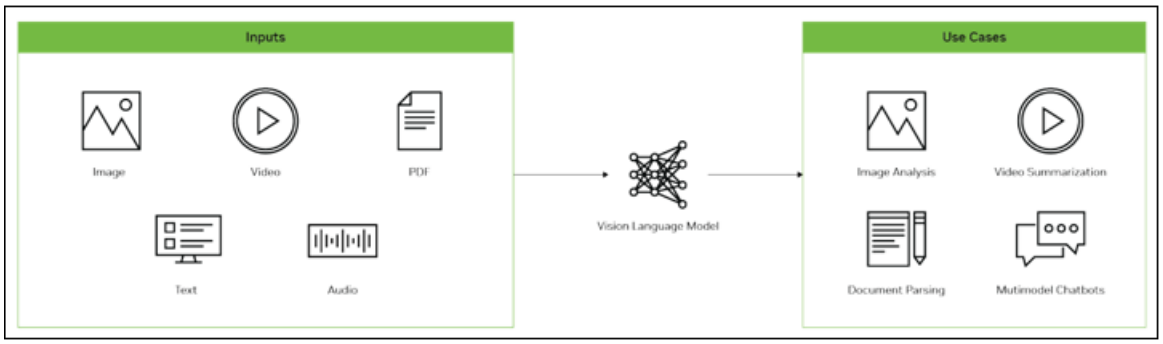
breakthrough in artificial intelligence, offering unprecedented performance in interpreting and generating multimodal content through their deeply integrated architecture.

Vision LLMs overcome previous barriers through:

- Training on massive datasets of paired image-text samples [9]
- Development of sophisticated multimodal architectures [10]
- Capabilities extending to complex tasks like image captioning and visual question answering

2. Architecture of Vision LLMs

Vision language models, according to Nvidia [11] are multimodal AI systems built by combining a large language model (LLM) with a vision encoder, giving the LLM the ability to "see." With this ability, VLMs can process and provide advanced understanding of video, image, and text inputs supplied in the prompt to generate text responses.



Overall architecture of the proposed by authors [12] Vision LLM consists of three parts: a unified language instruction designed to accommodate both vision and vision-language tasks, an image tokenizer that encodes visual information guided by language instructions and an LLM-based open ended task decoder that executes diverse tasks defined by language instructions.

Thus, Vision LLMs typically consist of two main components and LLM:

2.1 Visual Encoder

The visual encoder processes raw images into a structured representation. Common approaches include:

- Convolutional Neural Networks (CNNs): Convolutional neural networks (CNNs) are one of the main types of neural networks used for image recognition and classification [13]. Traditional models like ResNet [63] extract hierarchical features from images.
- Vision Transformers (ViTs): Introduced by [14] ViTs apply self-attention mechanisms to image patches, improving scalability and performance. Image recognition tasks that play an important role **in** digital health applications [15]. Vision Transformers (ViT) have emerged as a promising option for convolutional neural networks (CNN) for image analysis tasks, offering scalability and improved performance [16]

2.2 Language Model Integration

The encoded visual features are fused with a pre-trained LLM (e.g., GPT, BERT) using cross-modal attention mechanisms. Popular approaches include:

- Flamingo [17]: A model that Bridge powerful pretrained vision-only and language-only models [18] and interleaves visual and textual tokens for seamless multimodal processing. Alayrac et al [17] presented Flamingo, a new class of visual-language AI systems designed to master both images and text which solves three fundamental challenges: (1) connecting state-of-the-art vision and language models, (2) processing mixed media content (like alternating images and text) and (3) working with both photos and video clips. This unique flexibility allows Flamingo to learn from massive online datasets containing random mixes of pictures and words - the key to its remarkable ability to quickly adapt to new tasks with just a few examples.
- BLIP-2 [19]: Uses a lightweight querying transformer to align visual and language features efficiently. BLIP-2 is a scalable multimodal pre-training method that enables any LLMs to understand images while keeping their parameters entirely frozen. It effectively Bootstraps Language-Image Pre-training with frozen image encoders and frozen LLMs [20]. Dzabraev et al [21] present an unsupervised method for enhancing an image captioning model (BLIP2) using reinforcement learning and vision-language models like CLIP and BLIP2-ITM as reward models [21].

3. Training Datasets for Vision LLMs

The effectiveness of Vision-Language Models (VLMs) is fundamentally tied to the datasets they are trained on. The progress in vision-language models (VLMs) has been intrinsically linked to the availability of large-scale datasets [22]. These models require massive, high-quality multimodal datasets that pair visual content (images/videos) with relevant textual descriptions. Pre-trained vision and language models [23, 24] have demonstrated state-of-the-art capabilities over existing tasks involving images and texts, including visual question answering [26]. The authors [25] assume that the visual concepts, if captured by pre-trained VLMs, can be extracted by their vision-language interface with text-based concept prompts.

Below are the key datasets, their characteristics and associated challenges.

3.1 Common Types of Training Data

Since 2020, most LLM builders compose their training data with two types of datasets: targeted sourced data and broad chunks of web crawl data [27].

Vision LLMs typically train on:

a) *Web-Scraped Data*: Web scraping is a powerful technique that extracts data from websites, enabling automated data collection, enhancing data analysis capabilities and minimizing manual data entry efforts [28].
   - Billions of image-text pairs from public web sources (e.g., LAION-5B).
   - dataset consisting of 5.85 billion CLIP-filtered image-text pairs, of which 2.32B contain English language [29]

b) *Curated Datasets*: Human-annotated collections (e.g., COCO, Flickr30k) with precise captions- COCO Annotator is an image annotation tool that allows the labelling of images to create training data for object detection and localization [30] and authors undertook a comprehensive reevaluation of the COCO segmentation annotations. By enhancing the annotation quality and expanding the dataset to encompass 383K images with more than 5.18M panoptic masks, introduced COCONut, the COCO **N**ext **U**niversal segmen**T**ation dataset [31]. On the other, authors [32] took a step further in pushing the limits of vision-and-language pre-training data by relaxing the data collection pipeline used in Conceptual Captions 3M (CC3M) [32] and introduced the Conceptual 12M (CC12M), a dataset with 12 million image-text pairs specifically meant to be used for vision-and-language pre-training [33].

c) *Instruction-Tuning Data*: Task-specific examples (e.g., visual question-answering pairs from VQA-v2). This dataset contains twice as many question-answer pairs as the old version, with 200K images and 1.1M question-answer pairs. Each image has three questions, and each question also has ten answers. VQA v2.0 has both open-end and multiple-choice questions [34].

Dataset	Size	Description	Use Case
LAION-5B	5.8B image-text pairs	Web-crawled, filtered for quality	Pretraining general VLMs
COCO	330K images	Human-annotated with detailed captions	Fine-tuning, evaluation
Conceptual Captions	3.3M images	Alt-text derived from web images	Pretraining
VQA-v2	1.1M Q&A pairs	Questions about COCO images	Instruction-tuning for reasoning

4. Data Challenges

4.1 Scale vs. Quality Trade-off

The trade-off between scale and quality in datasets is a fundamental challenge in training machine learning models [35], particularly for large-scale applications like computer vision or natural language processing. Web-sourced datasets, such as LAION (Large-scale Artificial Intelligence Open Network), offer immense scale—billions of image-text pairs scraped from the internet. This offers models to learn from diverse, real-world data, capturing a wide range of patterns and representations. However, this comes at the cost of noise: web data often includes mislabeled, low-quality or irrelevant content and that this large-scale dataset is non-curated [36]. For instance, LAION's reliance on web crawls means it inherits inconsistencies like incorrect captions or poor image resolution, which can degrade model performance if not filtered effectively.

In contrast, curated datasets like COCO [37] (Common Objects in Context) prioritize quality over quantity. COCO contains around 330,000 images with detailed annotations (e.g., object segmentation, captions) crafted by human annotators. This precision makes it ideal for tasks requiring high accuracy, such as object detection, but its smaller size limits the diversity of scenarios it covers. The trade-off is evident in practice: models trained on LAION might generalize better across varied inputs, while those trained on COCO excel in controlled, well-defined tasks. Research [38], emphasizes the value of scale for pre-training, while [39] who introduced COCO, highlight the importance of clean annotations for benchmarking. Training multimodal models requires massive computational resources, limiting accessibility [40]. The black-box nature of deep learning models makes it difficult to understand their decision-making processes [41].

4.2 Bias Propagation

Vision LLMs inherit biases from training datasets, leading to skewed outputs [42]. A classic case is seen in datasets like COCO or ImageNet, where captions or labels might disproportionately associate certain professions e.g. "nurse" with women, "engineer" with men [43] based on the images collected. This reflects real-world biases present in the source material, such as media or user-generated content. Studies like those by [44] on gender bias in visual datasets demonstrate how co-occurrences in training data (e.g., "woman" and "cooking") lead models to overgeneralize these patterns, even when they don't hold universally.

4.3 Licensing Issues

Dataset licensing is currently an issue in the development of machine learning systems [45]. A significant hurdle is the absence of widely accepted, standardized contractual frameworks (i.e. data licenses) [46]. Licensing forms a significant hurdle for many datasets, especially those from web scraping. LAION has faced legal scrutiny because it includes images and text pulled from the public internet, often without explicit consent from copyright holders. This was a flashpoint in 2023 when artists and photographers raised concerns about their work being included in AI training sets without permission, sparking debates over intellectual property in the AI era. The court ruled that AAION did not infringe – this first court test of the EU legal framework for AI training's good news for LAION and anyone interested in training data transparency in general [47]. While ruling 'fair use' for research is acceptable but the lack of clear licensing agreements leaves its usage legally ambiguous.

MS COCO images dataset is licensed under a Creative Commons Attribution 4.0 License [48], thus COCO, not immune to licensing questions, benefits from a more controlled creation process—images were sourced with annotations added under Microsoft's oversight, providing clearer usage terms. The uncertainty forces researchers and companies to navigate a patchwork of laws, risking future challenges.

Summarizing the key challenges in Vision LLM datasets:

Challenge	Description	Examples
Scale vs. Quality Trade-off	Large-scale web datasets offer breadth but suffer from noise and inaccuracies, while manually curated datasets are high-quality but limited in diversity and volume.	LAION-5B (noisy web data) vs. COCO (clean but small-scale annotations).
Bias Propagation	Datasets often encode and amplify societal biases, leading to skewed model outputs (e.g., gender, racial, or cultural stereotypes in generated captions).	Gender bias in occupation descriptions (e.g., "nurse" vs. "doctor" associations).
Licensing Issues	Legal ambiguities around dataset provenance and copyright compliance pose risks for commercial deployment.	LAION lawsuits over copyrighted image scraping; CC12M's non-commercial restrictions.

4.4 Application of Vision LLM

Domain	Application	Description
Healthcare	Medical Image Analysis	Vision LLMs assist radiologists by generating diagnostic reports from X-rays and MRIs [49].
	Surgical Assistance	Real-time vision-language models provide guidance during robotic surgeries [50].
	Disease Progression Tracking	Analyzes medical images over time to monitor conditions like tumors or degenerative diseases [51].
	Reading Medical data	developing VLMs to harness multimodal medical data for improved healthcare applications [52].
Autonomous Systems	Self-Driving Cars	Vision LLMs enhance scene understanding by interpreting traffic signs and pedestrian movements [53,54]
	Robotics	Robots use Vision LLMs for object manipulation and human-robot interaction [55].

5. Future Directions

Future research in Vision Large Language Models (LLMs) is poised to address their utility and accessibility. One is efficient training techniques, particularly through model distillation [56]. Distillation involves smaller 'student' models learning from larger and better 'teacher' models [64]. This approach involves transferring knowledge from a large, computationally intensive "teacher" model to a smaller, more efficient "student" model, this will reduce significant computational costs and energy demands associated with training massive Vision LLMs, making them more practical for widespread use without sacrificing performance [57]. DeepSeek used its own AI model, DeepSeek V3, along with other advanced AI models, to create 800,000 step-by-step reasoning examples (Chain of Thought). These examples were then used to train and improve DeepSeek R1 through reinforcement learning, helping it become smarter and more accurate.

Vision LLMs, which integrate visual and textual data, often inherit biases from their training datasets—such as skewed representations of gender, race, or culture. Future efforts may prioritize fairness-aware training protocols, incorporating techniques like adversarial debiasing or balanced dataset curation [58]. These methods aim to fostering trust and ethical deployment in sensitive applications like healthcare or hiring.

Finally, optimizing Vision LLMs for edge devices [59,60]—such as smartphones or wearable—requires lightweight architectures and low-latency inference. This could enable applications like autonomous navigation or augmented reality to operate seamlessly in dynamic environments without relying on cloud computing. Model compression techniques such as quantization, pruning and hardware-aware design will likely play a role in achieving this, balancing speed and accuracy [61]. Together, these advancements promise to make Vision LLMs more efficient, fair and responsive, unlocking their potential across diverse real-world scenarios while addressing current limitations.

6. Conclusion

Vision LLMs now a groundbreaking advancement in AI, merging visual and linguistic understanding to enable more intuitive human-machine interactions. While challenges remain, ongoing research in efficiency, fairness and interpretability will further enhance its capabilities. As these models evolve, they will play an increasingly vital role in AI-driven applications across industries. By combining computer vision with natural language capabilities, these models enable applications such as diagnostic report generation in healthcare and scene interpretation in autonomous systems. However, challenges persist, including computational inefficiency, biases in training data and limited interpretability, which hinder widespread adoption. Ongoing research aims to enhance efficiency through optimized architectures, improve fairness via balanced datasets and increase interpretability with explainable AI techniques. As these issues are addressed, Vision LLMs are expected to play a critical role in AI-driven applications across diverse sectors, including healthcare, robotics, etc shaping the future of interdisciplinary AI innovation.

7. IndoAI's Implementation of Vision LLMs in AI-Powered Imaging Systems [62]

The integration of Vision LLMs into real-world systems represents a significant milestone in multimodal AI research. IndoAI, an emerging leader in applied artificial intelligence, is pioneering the deployment of Vision LLMs to develop next-generation AI models for its intelligent imaging solutions, particularly in its flagship **IndoAI AI Camera** platform. This implementation demonstrates the practical viability of Vision LLMs beyond theoretical frameworks, offering insights into scalability, efficiency and real-time performance.

7.1 Vision LLM as a Foundational Framework

IndoAI's approach leverages Vision LLMs to bridge visual perception and linguistic reasoning, enabling AI systems to perform context-aware scene interpretation. Unlike conventional vision models that operate in a unidirectional (image-to-label) manner, IndoAI's architecture employs bidirectional vision-language alignment, allowing:

- Dynamic scene description generation (e.g., automated security logs, industrial inspection reports)
- Query-based visual reasoning (e.g., "Identify anomalies in this machinery scan")
- Multimodal knowledge retrieval (cross-referencing visual inputs with textual databases)

This framework aligns with recent advancements in large multimodal models (LMMs) but focuses on edge deployment, optimizing for latency and computational efficiency—a critical requirement for embedded AI camera systems.

7.2 Case Study: IndoAI AI Camera

The IndoAI's edge AI Camera serves as a testbed for evaluating Vision LLM performance in practical settings. Key functionalities include:

- Automated Visual Reporting: Generating structured narratives from live footage (e.g., "Construction site safety audit: 2 workers detected without helmets; stranger in the premises").
- Interactive Querying: Enabling users to interrogate visual data via natural language (e.g., "List all vehicles not of this campus;").
- Predictive Diagnostics: Early detection of anomalies in healthcare imaging or manufacturing quality control, supplemented by textual recommendations.

7.3 Future Work and Broader Implications

IndoAI's ongoing research focuses on:

- Lightweight cross-modal architectures for resource-constrained devices.
- Continual learning to adapt Vision LLMs to evolving environments.
- Ethical AI governance frameworks for responsible deployment.

This case study explains the transformative potential of Vision LLMs in industrial and consumer applications, while highlighting the need for further research into robustness, scalability and user trust. As Vision LLMs mature, IndoAI's work provides a blueprint for transitioning from lower scale models to mission-critical systems, especially in edge computing.

References

[1] Balakrishnan Chinnaiyan, Sundaravadivazhagan Balasubaramanian, Mahalakshmi Jeyabalu, Gayathry S. Warrier, (2024), AI Applications – Computer Vision and Natural Language Processing, https://doi.org/10.1002/9781394219230.ch2
[2] Wenfeng Zheng, Mingzhe Liu,Kenan Li, Xuan Liu, (2023), AI for Computational Vision, Natural Language Processing, and Geoinformatics, Appl. Sci. 2023,, 13(24), 13276; https://doi.org/10.3390/app132413276
[3] https://www.chooch.com/blog/how-to-integrate-large-language-models-with-computer-vision/
[4] OpenAI, 2023, GPT-4 Technical Report. arXiv:2303.08774.
[5] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Scialom, T., (2023), Llama 2: Open Foundation and Fine-Tuned Chat Models, arXiv:2307.09288.
[6] Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, Yu Kong, Visual Large Language Models for Generalized and Specialized Applications, https://arxiv.org/html/2501.02765v1
[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, (2021), Learning Transferable Visual Models from Natural Language Supervision, https://arxiv.org/abs/2103.00020
[8] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, Aman Chadha, Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions, https://arxiv.org/html/2404.07214v2
[9] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee, 2023, Visual Instruction Tuning, 37th Conference on Neural Information Processing Systems (NeurIPS 2023), https://llava-vl.github.io
[10] https://encord.com/blog/vision-language-models-guide/]
[11] https://www.nvidia.com/en-us/glossary/vision-language-models/
[12] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, Jifeng Dai, VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks, https://arxiv.org/abs/2305.11175
[13] Mohammad Mustafa TayeORCID, (2023), Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions, Computation 2023, 11(3), 52; https://doi.org/10.3390/computation11030052
[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, (2020), An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://doi.org/10.48550/arXiv.2010.11929,

[15] Al-hammuri, K., Gebali, F., Kanan, A. et al., (2023), Vision transformer architecture and applications in digital health: a tutorial and survey. Vis. Comput. Ind. Biomed. Art 6, 14 (2023). https://doi.org/10.1186/s42492-023-00140-9
[16] Attiapo Acybah Morel Omer, (2024), Journal of Computer and Communications, Vol.12 No.4, April 2024, Image Classification Based on Vision Transformer, DOI: 10.4236/jcc.2024.124005
[17] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan,2022, Flamingo: a Visual Language Model for Few-Shot Learning, https://arxiv.org/abs/2204.14198
[18] https://llmsystem.github.io/llmsystem2024spring/assets/files/Group-Flamingo-98ae9c68fca94cd437716229a2cf42c1.pdf
[19] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, 2023, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, https://doi.org/10.48550/arXiv.2301.12597
[20] https://www.salesforce.com/blog/blip-2/
[21] Dzabraev Maksim, Kunitsyn Alexander, Ivanyuta Andrey; VLRM: Vision-Language Models act as Reward Models for Image Captioning, https://arxiv.org/html/2404.01911v1
[22] Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, Xiaohua Zhai; Scaling Pre-training to One Hundred Billion Data for Vision Language Models, https://arxiv.org/html/2502.07617v1
[23] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, Ming-Wei Chang, 2023, Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? https://doi.org/10.48550/arXiv.2302.11713
[24] Qian-Wei Wang, Yuqiu Xie, Letian Zhang, Zimo Liu, Shu-Tao Xia, (2024), Pre-Trained Vision-Language Models as Partial Annotators, https://doi.org/10.48550/arXiv.2406.18550
[25] Yuan Zang, Tian Yun, Hao Tan, Trung Bui, Chen Sun, 2025, Pre-trained Vision-Language Models Learn Discoverable Visual Concepts, https://doi.org/10.48550/arXiv.2404.12652
[26] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, Ming-Wei Chang, 2023, Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? https://aclanthology.org/2023.emnlp-main.925.pdf
[27] Stefan Baack, A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl, Mozilla Foundation, https://facctconference.org/static/papers24/facct24-148.pdf
[28] https://aclanthology.org/2024.emnlp-main.141.pdf
[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu et al, 2022, LAION-5B: An open large-scale dataset for training next generation image-text models, DOI: 10.48550/arXiv.2210.08402
[30] Daniela Stefanics, Markus Fox, 2022, COCO Annotator: Web-Based Image Segmentation Tool for Object Detection, Localization, and Keypoints, Volume 13, Issue 3, Article No.: 7, Page 1, https://doi.org/10.1145/3578495.3578502
[31] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, Liang-Chieh Chen, 2022, COCONut: Modernizing COCO Segmentation, https://arxiv.org/html/2404.08639v1
[32] S. Changpinyo, P. Sharma, N. Ding and R. Soricut, 2021, Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts, *2021* IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 3557-3567, doi: 10.1109/CVPR46437.2021.00356.
[33] Soravit Changpinyo, Piyush Sharma, Nan Ding, Radu Soricut, 2021, Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts, https://arxiv.org/abs/2102.08981
[34] Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Sunil, Aryal, Imran Razzak, Hakim Hacid,, Visual Question Answering: From Early Developments to Recent Advances - A Survey, https://arxiv.org/html/2501.03939v1
[35] Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, Lingzhong Meng, 2023, A survey on dataset quality in machine learning,Information and Software Technology, Volume 162, 2023, https://doi.org/10.1016/j.infsof.2023.107268
[36] https://laion.ai/blog/laion-400-open-dataset/
[37] Diego Bonilla Salvador, PixLore: A Dataset-driven Approach to Rich Image Captioning, https://arxiv.org/html/2312.05349v1
[38] Schuhmann, C., et al.,2022, LAION-5B: An open large-scale dataset for training next-generation image-text models." arXiv preprint.
[39] Lin, T.-Y., et al. (2014). "Microsoft COCO: Common Objects in Context." ECCV.
[40] Emma Strubell, Ananya Ganesh,Andrew McCallum, 2020, Energy and Policy Considerations for Modern Deep Learning Research, Vol. 34 No. 09: Issue 9: EAAI-20 https://doi.org/10.1609/aaai.v34i09.7123
[41] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders and K. -R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," in Proceedings of the IEEE, vol. 109, no. 3, pp. 247-278, March 2021, doi: 10.1109/JPROC.2021.3060483.
[42] Birhane, A., et al., 2021, Multimodal Datasets: Misogyny, Bias, and Ethical Concerns. FAccT.
[43] Susan Leavy, Gerardine Meaney, Karen Wade, Derek Greene, Mitigating Gender Bias in Machine Learning Data Sets University College Dublin, Ireland
[44] Jieyu Zhao,Tianlu Wang, Mark Yatskar et al, 2017, Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, Conference: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Jan 2017, DOI: 10.18653/v1/D17-1323
[45] Junyu Chen, Norihiro Yoshida, Hiroaki Takada, An investigation of licensing of datasets for machine

learning based on the GQM model, https://arxiv.org/abs/2303.13735
[46] https://mlcommons.org/2025/03/unlocking-data-collab/
[47] Paul Keller, oct 2024, https://openfuture.eu/blog/laion-vs-kneschke/.
[48] https://viso.ai/computer-vision/coco-dataset/]
[49] Li M, Jiang Y, Zhang Y, Zhu H.,2023, Medical image analysis using deep learning algorithms. Front Public Health. 2023 Nov 7; 11:1273253. doi: 10.3389/fpubh.2023.1273253.
[50] Samuel Schmidgall, Joseph Cho. Cyril Zakka, William Hiesinger, GP-VLS: A general-purpose vision language model for surgery, https://arxiv.org/html/2407.19305v2
[51] Rafael Zamora-Resendiz, Ifrah Khuram, Silvia Crivelli, 2024, Towards Maps of Disease Progression: Biomedical Large Language Model Latent Spaces For Representing Disease Phenotypes And Pseudotime, doi: https://doi.org/10.1101/2024.06.16.24308979
[52] Iryna Hartsock, Ghulam Rasool, 2024, Vision-language models for medical report generation and visual question answering: a review, Front. Artif. Intell., 19 Nov 2024, Sec. Medicine and Public Health Volume 7 – 2024, https://doi.org/10.3389/frai.2024.1430984
[53] Dianwei Chen, Zifan Zhang, Yuchen Liu, Xianfeng Terry Yang, 2025, INSIGHT: Enhancing Autonomous Driving Safety through Vision-Language Models on Context-Aware Hazard Detection and Edge Case Evaluation, https://arxiv.org/html/2502.00262v1
[54] Haoxiang Gao, Yu Zhao, 2025, Application of Vision-Language Model to Pedestrians Behavior and Scene Understanding in Autonomous Driving, https://doi.org/10.48550/arXiv.2501.06680
[55] Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo,Tadayoshi Aoyama,,Yasuhisa Hasegawa, 2024, Enhancing the LLM-Based Robot Manipulation Through Human-Robot Collaboration, IEEE Robotics and Automation Letters, vol. 9, no. 8, pp. 6904-6911, Aug. 2024., DOI:https://doi.org/10.1109/LRA.2024.3415931
[56] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, Tianyi Zhou, A Survey on Knowledge Distillation of Large Language Models, https://doi.org/10.48550/arXiv.2402.13116, 2024
[57] Chuanpeng Yang, Wang Lu, Yao Zhu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, Yiqiang Chen, 2024, Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application, https://doi.org/10.48550/arXiv.2407.01885
[58] Xiahua Wei, Naveen Kumar, Han Zhang, 2025, Addressing bias in generative AI: Challenges and research opportunities in information management, Information & Management, Vol 62, Issue 2,2025, https://doi.org/10.1016/j.im.2025.104103.
[59] Ahmed Sharshar, Latif U. Khan, Waseem Ullah, Mohsen Guizani,2025, Vision-Language Models for Edge Networks: A Comprehensive Survey, https://arxiv.org/html/2502.07855v1
[60] https://semiengineering.com/vision-is-why-llms-matter-on-the-edge/
[61] Dong Liu, 2024, Contemporary Model Compression on Large Language Models Inference, https://arxiv.org/html/2409.01990v1
[62] www.indo.ai
[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, https://www.cv-foundation.org/
[64] https://www.bruegel.org/policy-brief/how-deepseek-has-changed-artificial-intelligence-and-what-it-means-europe