Editorial | Uncover This Tech Term

eISSN 2005-8330
https://doi.org/10.3348/kjr.2025.1813
Korean J Radiol 2026;27(4):375-378

Korean Journal of Radiology
KJR

Uncover This Tech Term: Large Vision-Language Models in Radiology

Shahriar Faghani[1,2], Yae Won Park[3], Ji Eun Park[4]

[1]Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA
[2]Radiology Informatics Lab, Department of Radiology, Mayo Clinic, Rochester, MN, USA
[3]Department of Radiology and Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Republic of Korea
[4]Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University, Baltimore, MD, USA

Keywords: Large vision-language model; Vision-language model; Large multimodal model; Large language model; Artificial intelligence; Transformer

WHAT ARE LVLMs?

Large multimodal models are typically transformer-based foundational models that can process and generate multiple types of data (modalities), including text, images, audio, and video [1,2]. Large vision-language models (LVLMs) are a subset of large multimodal models that specifically focus on aligning and integrating visual and linguistic representations. Traditional artificial intelligence (AI) systems are trained to perform well-defined narrow tasks and have limited adaptability. By contrast, LVLMs generalize across diverse tasks and support flexible downstream applications without requiring task-specific retraining.

Received: November 27, 2025 **Revised:** December 30, 2025
Accepted: January 12, 2026
Corresponding author: Yae Won Park, MD, PhD, Department of Radiology and Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
• E-mail: yaewonpark@yuhs.ac

HOW ARE LVLMs BUILT?

The construction of LVLMs generally follows the standard foundational model pipeline: 1) largescale pretraining, 2) task- or domain-specific fine-tuning, and 3) post-training for behavior, format, and safety control [3]. LVLMs can be classified into three main categories according to their training strategies:

1) Contrastive image–text pretraining learns a shared embedding space from paired data and enables zero-shot transfer (including CLIP and BioMedCLIP). Variants, such as SigLIP, replace batch softmax with sigmoid or binary cross-entropy to better handle multilabel or noisy pairs.

2) Alignment-based models start with pretrained vision encoders and large language models and learn to connect them via a lightweight adaptor or cross-attention (including BLIP-2 and InstructBLIP). This category also includes instruction-tuned multimodal large language models (including LLaVA and LLaVA-Med) and two-stage or modular pipelines that first ground or align vision to text and then instruction-tune the model for chat, visual question answering (VQA), or reporting (such as MiniGPT-4 and mPLUG-Owl).

3) End-to-end or unified tokenization approaches treat visual inputs as discrete tokens alongside text and train the entire model jointly within a single transformer (such as MedGemma).

Lightweight multimodal connectors, such as simple projection layers, are commonly used to align visual features with language tokens while minimizing computational overhead. Medical variants (including RadFM and Med-Flamingo) typically involve domain-specific pretraining

layered on top of one of these three patterns [4].

APPLICATIONS OF LVLMs IN RADIOLOGY

LVLM applications are rapidly expanding across the radiology workflow (Fig. 1). Core applications include visual recognition tasks, such as detection, segmentation, and classification, in which LVLMs leverage largescale paired image–text representations to identify abnormalities, delineate organs or lesions, and categorize disease patterns. Their language-generation capabilities further support radiology report drafting, image captioning, and structured annotation, while cross-modal reasoning enables VQA, image–text retrieval, and case-based search.

Beyond basic image interpretation, LVLMs have begun to perform various functions in clinical practice, infrastructure management, and education. They can streamline workflow efficiency and quality assurance by generating preliminary reports, flagging critical findings, and retrieving similar cases from archives. In parallel, radiology education can be enhanced through interactive image–text learning modules and explanatory feedback. Their ability to integrate imaging data, narrative reports, and structured clinical information also opens new avenues for knowledge discovery, including cohort construction, cross-modal retrieval, and the identification of latent patterns. Furthermore, multimodal access to images, reports, and metadata enables automated auditing of bias, performance drift, and potential ethical risks, positioning LVLMs as valuable tools for fairness and governance in clinical AI systems.

More recently, agentic AI has emerged as a prominent research direction, positioning vision-language models as the cognitive core of autonomous radiology assistants capable of planning tasks, calling external tools, navigating PACS and RIS systems, retrieving prior studies, and iteratively refining outputs through feedback [5]. By linking visual perception with multimodal reasoning and actionable instructions, these models can integrate detection, reporting, retrieval, and decision support into a coherent end-to-end intelligent workflow. This shift aligns with the development of generalist AI models—vision–language foundation models—trained on extensive and diverse datasets to support a broad spectrum of downstream tasks. Current state-of-the-art applications can be broadly classified into two major categories.

VQA: "Read Like a Radiologist"

LVLMs can mimic radiologists by performing both image-based tasks (including detection and classification) and text-based tasks (including generating coherent reports). However, current models often struggle to produce clinically grounded and diagnostically coherent summaries. Zhong et al. [6] introduced a region-based generation framework embedded within a structured diagnostic workflow

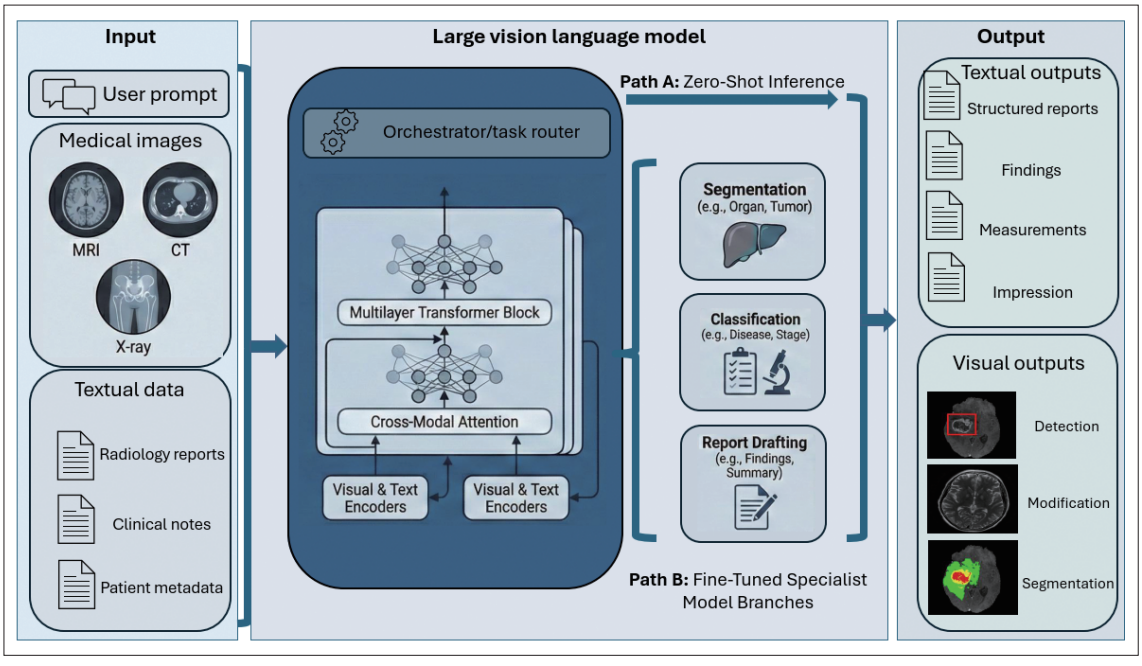

Fig. 1. Schematic of large vision-language models.

https://doi.org/10.3348/kjr.2025.1813

Korean Journal of Radiology
KJR

comprising: 1) abnormality identification, 2) region-level finding generation using VQA, 3) organ-centered VQA, and 4) abnormality-specific finding synthesis. This sequential, clinically oriented approach outperformed holistic captioning methods that relied solely on global image prompts, producing more precise and relevant descriptions of abnormalities. Representative medical VQA benchmark datasets include VQA-RAD, SLAKE, and PMC-VQA.

Zero-Shot Cross-Modal Inference: Image–Text Pair Inference

Zero-shot learning enables LVLMs to perform downstream tasks without additional training by learning an aligned latent embedding space for visual and linguistic representations. This capability allows retrieval of relevant text from image queries and vice versa, with direct clinical utility [7,8]. For example, Lee et al. [9] applied LVLMs to automatically remove burned-in protected health information from imaging studies, including patient names, identification numbers, dates of birth, and demographic markers, thereby facilitating secure data exchange and reducing the manual workload associated with de-identification. Another emerging application is the generation of visual abstracts. Lee et al. [10] evaluated LVLM-driven visual abstract creation for radiology journals using structured prompts derived from key study elements. This approach provides a feasible first-draft tool that, when refined through human–AI co-development, may improve the quality and efficiency of scientific communication.

OUTLOOK

LVLMs reshape radiology by unifying visual understanding with language-based reasoning, enabling applications ranging from image interpretation to reporting, retrieval, and workflow automation. However, their clinical integration remains limited owing to several practical challenges. Processing full 3D imaging volumes carries a substantial computational cost, which currently limits scalability, although ongoing advances in hardware and algorithmic efficiency are expected to mitigate this barrier. Moreover, most existing approaches generate image-level captions rather than finer-level descriptions, highlighting the need for more region-grounded reasoning [11,12]. Safety is a major concern in the clinical implementation of LVLMs. In particular, hallucinations (including confidently describing a pneumothorax that is not actually present in the image) represent a serious risk if such incorrect outputs are accepted without verification. Without transparency (including clear explainability of how a model derives its conclusions), clinicians may find it difficult to detect or challenge such errors, increasing the risk of harm. For the commercial use of LVLMs, validated data sources with transparency and safety mechanisms will ultimately be essential to reduce hallucinations and mitigate future legal issues. Finally, clinical integration requires domain-specific fine-tuning, uncertainty awareness, and interpretable outputs, with proactive oversight to ensure safety and equitable performance [5,13]. Notably, strong task-level generalization does not necessarily translate into clinical generalization, as robust performance on external datasets remains a critical requirement for real-world deployment.

CONCLUSION

With responsible development achieved through close collaboration among clinicians, academia, and industry, LVLMs have the potential to become reliable and transformative tools for imaging-based care.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: all authors. Supervision: Ji Eun Park. Writing—original draft: all authors. Writing—review & editing: all authors.

ORCID IDs

Shahriar Faghani
https://orcid.org/0000-0003-3275-2971
Yae Won Park
https://orcid.org/0000-0001-8907-5401
Ji Eun Park
https://orcid.org/0000-0002-4419-4682

Funding Statement

This study was financially supported by the Faculty Research Grant of Yonsei University College of Medicine (6-2023-0072).

REFERENCES

1. Jung KH. Uncover this tech term: foundation model. *Korean J*

Radiol 2023;24:1038-1041
2. Gupta A, Rangarajan K. Uncover this tech term: transformers. *Korean J Radiol* 2024;25:113-115
3. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, et al. Multimodal large language models in medical imaging: current state and future directions. *Korean J Radiol* 2025;26:900-923
4. Kalpélbé BC, Adaambiik AG, Peng W. Vision language models in medicine. arXiv [Preprint]. 2025 [accessed on December 21, 2025]. Available at: https://doi.org/10.48550/arXiv.2503.01863
5. Faghani S, Moassefi M, Rouzrokh P, Khosravi B, Erickson BJ. Uncover this tech term: agentic artificial intelligence in radiology. *Korean J Radiol* 2025;26:888-892
6. Zhong Z, Wang Y, Wu J, Hsu WC, Somasundaram V, Bi L, et al. Vision-language model for report generation and outcome prediction in CT pulmonary angiogram. *NPJ Digit Med* 2025;8:432
7. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med* 2023;29:2307-2316
8. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. *Nat Med* 2024;30:863-874
9. Lee T, Kim H, Park SH, Chae S, Yoon SH. Evaluation of vision-language models for detection and deidentification of medical images with burned-in protected health information. *Radiology* 2025;315:e243664
10. Lee T, Chae S, Park SH, Kahn CE, You SC, Yoon SH. Using a vision-language model to generate visual abstracts for radiology journals. *Radiology* 2025;316:e251458
11. Xiang J, Wang X, Zhang X, Xi Y, Eweje F, Chen Y, et al. A vision-language foundation model for precision oncology. *Nature* 2025;638:769-778
12. Shen Y, Xu Y, Ma J, Rui W, Zhao C, Heacock L, et al. Multi-modal large language models in radiology: principles, applications, and potential. *Abdom Radiol (NY)* 2025;50:2745-2757
13. Faghani S, Gamble C, Erickson BJ. Uncover this tech term: uncertainty quantification for deep learning. *Korean J Radiol* 2024;25:395-398

https://doi.org/10.3348/kjr.2025.1813