KJD-25K: A 3D Multi-modal Dataset and Benchmark for Knee Joint Diagnosis

Yihui Li[1], Along He[1](✉), Yuli Wang[2], Kai Guo[2], Yanlin Wu[3], and Ben Niu[4,5](✉)

[1] School of Artificial Intelligence, Shenzhen University, Guangdong, China
healong2020@163.com, drniuben@gmail.com
[2] Shenzhen Second People's Hospital, Shenzhen, Guangdong, China
[3] Nankai University, Tianjin, China
[4] College of Management, Shenzhen University, Guangdong, China
[5] Institute of Digital Intelligence Management and Risk Decision-Making, Shenzhen University, Guangdong, China

Abstract. Multi-modal large language models (MLLMs) have demonstrated increasing promise in medical image understanding, yet their capacity to support clinically aligned knee MRI radiology reports remains limited because of the scarcity of high-quality datasets that reflect real-world orthopaedic practice. Interpreting knee MRI requires not only recognizing imaging patterns but also structured diagnostic reasoning and severity assessments that reflect orthopedic clinical decision-making. In this work, we introduce **KJD-25K**, a large-scale, expert-annotated 3D multi-modal dataset specifically tailored for knee joint analysis. It comprises 25K high-quality MRI volume-report pairs with expert-curated Chinese radiology reports and structured severity annotations. The dataset was developed through over **4,000** hours of expert clinician effort, ensuring high fidelity to clinical workflows, diagnostic consistency, and semantic richness. Furthermore, we benchmark eight state-of-the-art MLLMs on KJD-25K, establishing comprehensive baselines for clinically grounded knee joint diagnosis. By providing a scalable, clinically validated resource, KJD-25K lays the foundation for developing interpretable, reliable, and workflow-integrated MLLMs for knee joint diagnosis. The dataset is publicly available at https://huggingface.co/datasets/YiHui0124/KneeCoT.

Keywords: Orthopaedic diagnosis · Multi-modal large language models · Dataset benchmarking

1 Introduction

Musculoskeletal disorders impose a substantial and enduring burden on global health, contributing significantly to chronic pain, loss of function, and long-term disability across populations worldwide [22]. Among these conditions, the knee joint diseases are particularly prevalent and clinically consequential. In clinical practice, magnetic resonance imaging (MRI) enables comprehensive assessment

of soft-tissue structures, providing diagnosis, staging, and longitudinal monitoring of knee disorders. However, the increasing scale and heterogeneity of knee imaging data present growing challenges for consistent interpretation, and downstream clinical utilization. Accurate diagnosis of knee disorders and the clinical reports constitute a complex and time-intensive task that heavily relies on the expertise and experience of orthopaedic specialists. Given the growing clinical demand and the potential for inter-observer variability, there is a critical need for computer-aided diagnosis (CAD) models that can assist clinicians in interpreting knee imaging and producing consistent, accurate, and clinically actionable diagnostic assessments.

Deep learning has been increasingly explored to assist in knee imaging analysis [25,1,14,10]. Existing studies have demonstrated promising results in detecting anterior cruciate ligament (ACL) ruptures [23], grading osteoarthritis severity [13,19,27] or knee joint region segmentation [29]. Recently, multi-modal large language models (MLLMs) have achieved remarkable progress across a broad spectrum of clinical applications, thanks to their strong capabilities in integrating visual and textual information [12,4,21,28]. Leveraging MLLMs for automated generation of knee joint diagnostic report presents a promising avenue to reduce radiologists' workload and improve clinical efficiency. However, the advancement of MLLMs in knee disease diagnosis remains significantly hindered by the lack of large-scale, high-quality multi-modal datasets. Such datasets are needed to capture real-world clinical complexity—including expert-aligned imaging findings, structured severity assessments, and diagnostic reasoning. This gap underscores the critical need for a dedicated benchmark that reflects authentic orthopaedic workflows and supports the development of clinically grounded knee joint diagnostic model.

To address these issues, we first constructed a large-scale 3D multi-modal knee dataset called KJD-25K. It includes 25,000 high-quality MRI volume-text paired with expert-annotated clinical reports, covering a wide spectrum of knee pathologies and severity levels. To the best of our knowledge, this is the largest publicly available multi-modal 3D knee dataset to date, offering scale and clinical depth to advance research in orthopaedic diagnosis. Extensive experiments across eight widely used MLLMs reveal that current approaches struggle to generate accurate and clinically coherent knee joint diagnostic reports, underscoring the critical need for a dedicated, high-quality benchmark dataset. Training with our constructed dataset effectively enhances the capacity of MLLMs to perform clinically grounded knee MRI interpretation. The inclusion of expert-curated diagnostic reports and structured severity annotations helps MLLMs better align with clinical descriptions for more accurate and reliable diagnosis. These results demonstrate that KJD-25K provides an effective and scalable foundation for bridging the gap between current MLLMs and the nuanced demands of orthopaedic diagnostic practice. Our contributions are two-fold:

(1) We established KJD-25K, a large-scale 3D multi-modal knee MRI dataset containing 25K volume-report pairs, expert-annotated Chinese radiology reports, and structured severity annotations for knee MRI report generation.

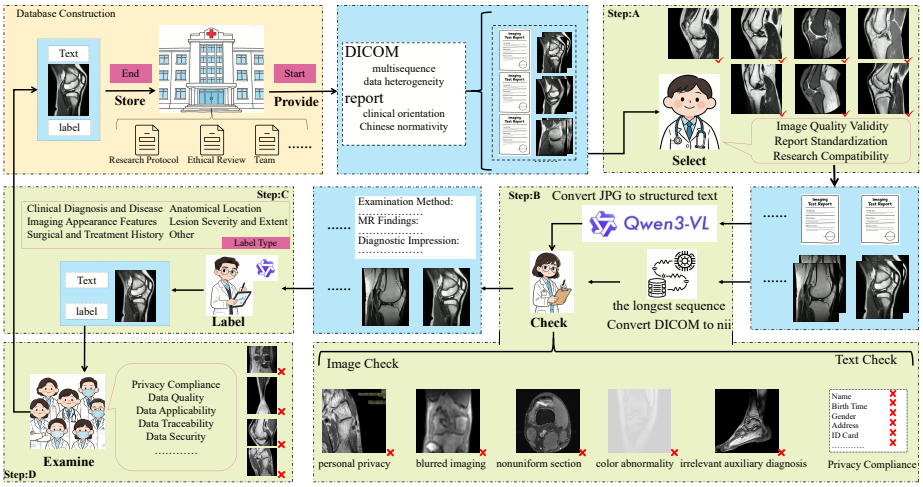

Fig. 1. Curation pipline of KJD-25K dataset. Step A: Experts select qualified images and standardized reports. Step B: Raw data are structured and consistency is verified. Step C: Clinicians label diagnosis, features, locations, and severity. Step D: Dataset is reviewed for privacy, quality, and traceability.

(2) We propose clinical-aware and linguistic metrics for systematic and reproducible evaluation of MLLMs in knee disease diagnosis and prognosis. Using these metrics, we benchmark eight SOTA MLLMs on KJD-25K, establish strong baselines, and provide a unified foundation for future research.

2 Curation of KJD-25K

2.1 Data Collection and Annotation

As illustrated in Fig. 1, KJD-25K was constructed from patients who visited the orthopaedic department and was collected from a local hospital. KJD-25K includes four standard clinical knee MRI sequences: T1WI, T2WI, T2WIFS, and PDWIFS, acquired on 3.0T scanners with dedicated 8-channel knee coils. The dataset was curated by a 15-person professional team, including five board-certified orthopedic radiologists with over ten years of experience, seven trained physician assistants, and three clinically trained engineers.Data processing was conducted through a human–AI collaborative pipeline, supported by five NVIDIA RTX 4090 GPUs, involving approximately 2,000 clinician-hours, with an average of 130 hours per participant, alongside 2,000 GPU-hours. Raw multi-sequence knee MRI volumes and their corresponding radiology reports were retrospectively extracted, de-identified, and approved by the Institutional Review Board (IRB), and only the MRI volumes and associated clinical reports were retained. The construction followed a four-step pipeline:

Expert-Guided Case Selection. This stage aimed to identify eligible patient studies for dataset construction from all knee MRI examinations acquired be-

tween December 2020 and December 2025. Eligible cases were required to have complete multi-sequence MRI data, a corresponding standardized Chinese radiology report, and sufficient image quality for clinical interpretation. Cases were excluded if they met any of the following criteria: (1) metallic implants causing severe imaging artifacts; (2) incomplete clinical records or missing radiology reports; or (3) acute trauma examinations performed within 72 hours after injury. Only studies satisfying all criteria were retained for subsequent processing. All eligible retained studies received rigorous full-process de-identification before subsequent processing, with all identifiable patient PHI systematically removed, including personal identifiers and privacy-sensitive information in the DICOM headers of MRI scans.

Multi-modal Data Structuring. Clinical reports provided as JPEG images were parsed using a locally deployed Qwen3-VL [3] model to automatically extract three predefined fields: examination method, imaging findings, and diagnostic conclusions. Trained clinicians corrected recognition errors, resolved ambiguities, and unified terminology to ensure consistency and clinical correctness.

In routine clinical practice, a single diagnostic case may contain multiple MRI sequences, including full-coverage scans, clinician-selected key views, and additional sequence of suspicious regions. To ensure effective utilization of MRI sequences for downstream modeling, we developed an automated procedure to identify and extract the full-coverage scan from each case. All extracted MRI volumes were subsequently reviewed by clinical experts to confirm completeness and correctness.

AI-Assisted Label Generation. The label taxonomy was developed based on prior literature [6] [17] and defined through iterative clinician discussions based on routine diagnostic practices. Using the finalized schema, a locally deployed Qwen3-Max model generated preliminary anatomical and clinical labels from structured reports. Clinical experts reviewed the generated labels at the case level and corrected errors where necessary, with random checks conducted to ensure overall labeling reliability.

Rigorous Dataset Validation and Compliance Review. The fully annotated dataset underwent a final examination stage to verify annotation accuracy, privacy protection, and data traceability. Samples failing quality or compliance requirements were removed, resulting in a reliable and ethically compliant benchmark dataset. The final dataset consists of 20K training examples and 5K test examples.

2.2 Dataset Analysis

Table 1 compares existing knee-related datasets, which are primarily designed for conventional tasks, such as classification, segmentation, or registration, with limited support for multi-modal reasoning. Early datasets are limited by their quantity or single modality task. Although 3DReasonKnee introduces question–answer annotations, they often couple visual understanding with explicit localization or detection objectives, which are not always aligned with clinical diagnostic workflows. Therefore, we aim to construct a larger-scale, high-quality

Table 1. Comparison of KJD-25K with other knee datasets. Cls (classification), Seg (segmentation), Det (detection), Reg (regression), Gen (generation), Regist (registration), SL (structured labels), CR (clinical reasoning), RG (Report Generation), VQA (Visual Question Answering), VLM (directly usable for vision–language model training), OA (Open Access).

Dataset	Size (>20k Vol.)	Modality	OA	Main Task(s)	SL	CR	VLM
MRKR [18]	503K slices	2D X-ray	✓	Cls, Reg	✓	✗	✗
SKM-TEA [5]	✗	3D MRI	✓	Seg, Det	✓	✗	✗
OAI [11]	✓	3D MRI	✓	Cls, Reg	✓	✗	✗
KMAR-50K [26]	✗	3D MRI	✗	Gen, Regist	✓	✗	✗
3DReasonKnee [20]	✗	3D MRI	✓	Cls, Seg, Det, VQA	✓	✓	✓
KJD-25K (Ours)	✓	3D MRI	✓	Cls, RG, VQA	✓	✓	✓

3D dataset to support vision–language model training for knee joint disease diagnosis. It provides structured labels aligned with clinical reporting standards and emphasizes clinically coherent reasoning. The inclusion of expert-curated volume–text pairs and diagnostic narratives enables models to learn the intrinsic correspondence between imaging findings and diagnostic conclusions.

Fig. 2 illustrates the distribution of report text lengths in KJD-25K and the hierarchical structure of the proposed labeling framework. For clarity, Fig. 2 is simplified to retain only the core information. Our dataset is directly usable for end-to-end VLM training and better reflects real-world clinical reasoning processes and has a larger amount of 3D multi-modal data, distinguishing it from prior knee imaging benchmarks.

3 Experiments

3.1 Experiment Setup

We select top-performing MLLMs for comprehensive evaluation, including open-source models InternVL3.5 (8B, 30B) [24], Qwen3-VL-Instruct (8B, 30B), and Qwen3-VL-Thinking (8B, 30B) [3]; closed-source models GPT-5 [15], Claude-Sonnet-4.5 [2] and Gemini-2.5pro; medical MLLMs such as HuatuoGPT-Vision-7B [4], Lingshu (7B) [28], and Hulu-med (4B, 7B, 32B) [9].

For open-source models, experiments are conducted on servers equipped with NVIDIA H100 GPUs, while API-based service is used for closed-source MLLMs. We further assess the effectiveness of the proposed dataset for MLLMs training by fine-tuning Qwen3-VL using Low-Rank Adaptation (LoRA) [8]. During training, both the pre-trained vision encoder and the LLM backbone were frozen, and only the LoRA adapters were optimized. The LoRA rank was set to 8, alpha to 16, and dropout to 0.05. The learning rate was 2e-4, the batch size was 16, and the model was trained for 3 epochs.

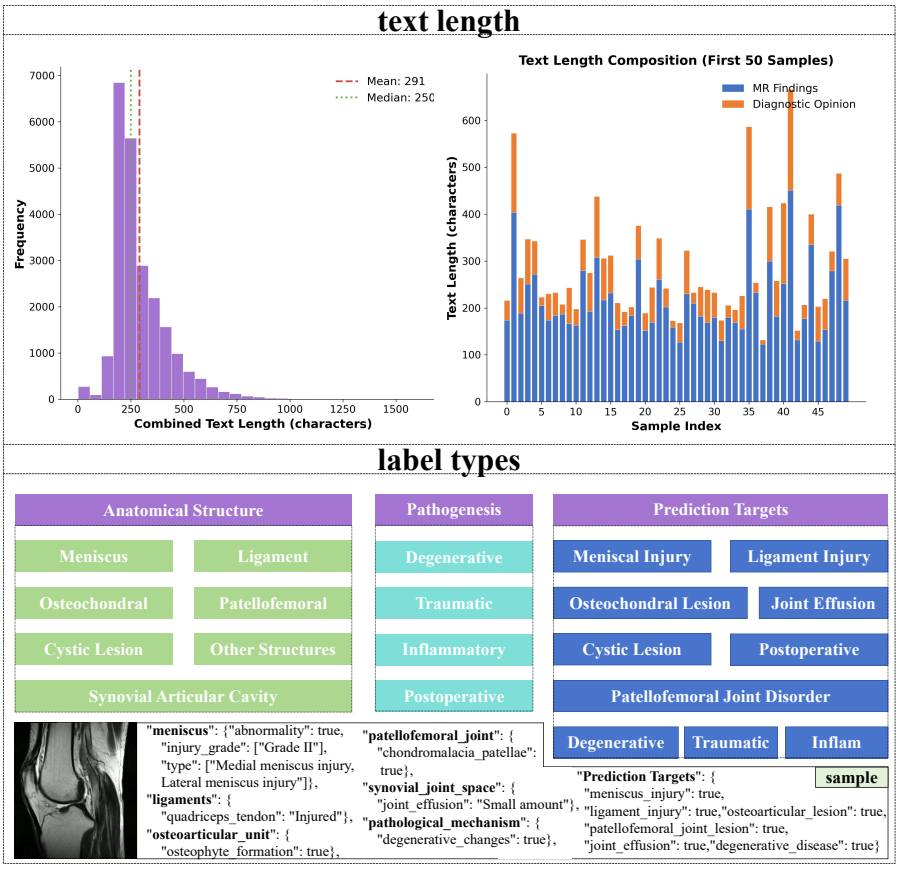

Fig. 2. The data statistics of KJD-25K on label types and text length.

3.2 Evaluation Metrics for KJD-25K

To comprehensively evaluate model performance on KJD-25K, we employ a hybrid protocol that assesses both linguistic similarity and clinical correctness of generated reports. For linguistic evaluation, BLEU [16] and F1 scores are computed to measure consistency. Because textual similarity alone is insufficient to evaluate report-level clinical adequacy, we further introduce three structured LLM-assisted metrics: Lesion Sensitivity (LS), Diagnostic Specificity (DS), and Structured Report Compliance (SRC), which were adapted from routine knee MRI interpretation training and structured reporting practice. LS measures the proportion of ground-truth pathological findings correctly identified in the generated report. DS measures the proportion of normal anatomical structures correctly recognized as normal, thereby penalizing over-diagnosis and unsupported abnormal findings. SRC evaluates whether the generated report follows the predefined reporting fields, including examination method, MRI findings, and diagnostic impression. These criteria have been rigorously reviewed and validated by

Table 2. Performance comparison of different MLLMs on report generation. LS: Lesion Sensitivity; DS: Diagnostic Specificity; SRC: Structured Report Compliance.The best results are marked in red, and the second-best results are marked in blue.

Model	**Text-based**		**Clinical**		
	BLEU ↑	F1 ↑	LS ↑	DS ↑	SRC ↑
General-purpose Open-source MLLMs					
InternVL3.5-8B [24]	33.11	2.80	68.07	37.84	54.07
InternVL3.5-30B [24]	**37.49**	**2.81**	70.97	46.33	55.43
Qwen3-VL-Instruct-8B [3]	28.55	2.20	54.48	23.59	49.54
Qwen3-VL-Instruct-30B [3]	29.67	2.46	56.43	27.89	51.66
Qwen3-VL-Thinking-8B [3]	16.92	0.90	45.81	36.40	51.45
Qwen3-VL-Thinking-30B [3]	19.43	1.86	52.66	39.47	56.47
Closed-source MLLMs					
GPT-5 [15]	28.58	2.47	**77.27**	**55.56**	54.81
Gemini-2.5Pro [7]	29.98	2.19	64.42	49.20	54.41
Claude-Sonnet-4.5 [2]	29.36	2.14	71.39	49.88	54.68
Medical-domain Open-source MLLMs					
HuatuoGPT-Vision-7B [4]	14.03	1.17	48.10	35.99	53.16
Lingshu-7B [28]	30.16	1.83	66.51	39.22	54.44
Hulu-4B [9]	18.24	1.07	49.92	11.87	53.49
Hulu-7B [9]	21.67	1.38	47.54	38.99	54.75
Hulu-32B [9]	23.44	1.43	57.10	44.85	**58.68**
Ours	**42.71**	**3.26**	**79.63**	**58.47**	**66.42**

experienced radiologists to ensure alignment with clinical judgment, with each report assigned a score ranging from 0 to 100. For clinical-metric evaluation, we used GPT-4o as an automated evaluator. The evaluator was provided with the ground-truth report, structured labels, and model-generated report, and was instructed to assign LS, DS, and SRC scores on a 0–100 scale according to predefined scoring criteria. The evaluation prompts were iteratively refined on 100 randomly sampled cases, and senior orthopedic radiologists confirmed strong qualitative agreement with expert assessment.

3.3 Evaluation Results

Table 2 summarizes the performance of different MLLMs with linguistic and clinical metrics. First, models trained on our dataset achieve systematically better performance across all evaluation metrics. In particular, compared with general-purpose MLLMs without domain-specific adaptation, model trained on KJD-25K shows clear improvements in diagnostic correctness and clinical impact, indicating that the curated multi-modal data provides more informative supervision for knee joint diagnostic tasks. Second, while closed-source models

(e.g., GPT-5 and Claude-Sonnet-4.5) exhibit strong baseline performance, their gains are not uniformly dominant across all metrics. Additionally, under identical prompts, Linshu-7B generates outputs in multiple language versions, which leads to anomalous performance on the F1 and BLEU metrics compared with other models. In contrast, model trained on KJD-25K demonstrates more stable and balanced improvements, particularly on expert-oriented criteria such as diagnostic consistency and clinical usefulness. This suggests that domain-specific, high-quality datasets play a critical role.

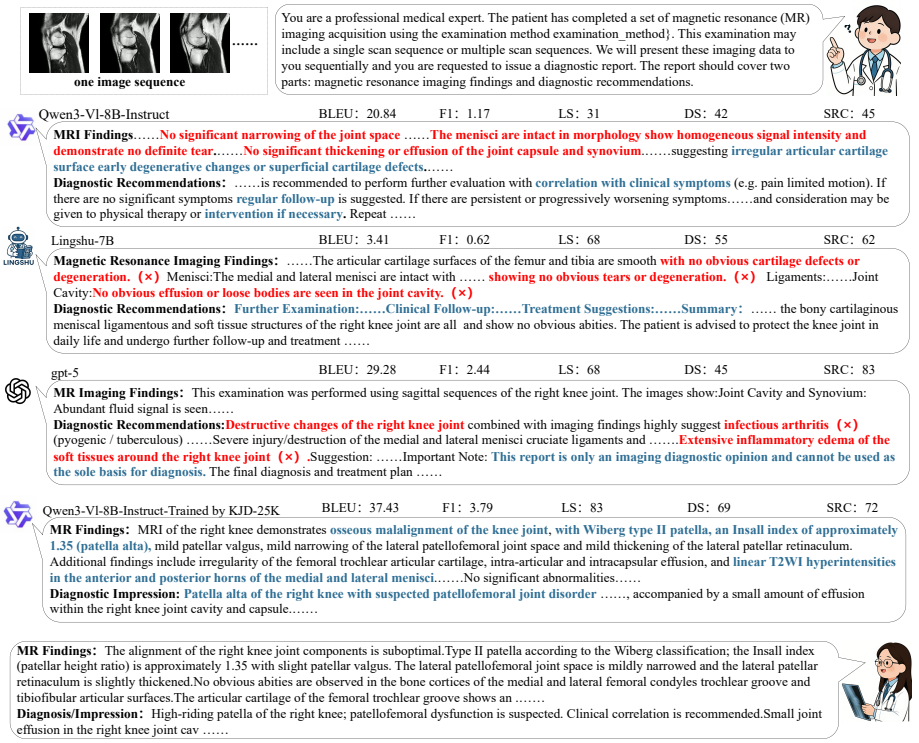

Fig. 3. Case study of knee joint diagnosis report generation with different MLLMs. Correct diagnostic statements are highlighted in blue, while obvious errors are highlighted in red.

Fig. 3 qualitatively illustrates the strong generative capability of our model, whose generated reports exhibit closer alignment with the reference answers. Specifically, Qwen3-VL-8B-Instruct and Lingshu-7B produce internally inconsistent reports, such as describing intact meniscal morphology while simultaneously suggesting cartilage degeneration or joint effusion, indicating weak alignment between imaging evidence and textual conclusions. In contrast,GPT-5 tend to over-interpret the imaging inputs, generating unsupported diagnoses such as infectious arthritis, extensive soft-tissue edema, or severe ligament destruction

that are not substantiated by the MRI. These errors consistently stem from the absence of a unified, expert-annotated knee MRI dataset rather than differences in model capacity. The final output of the model trained on KJD-25K is intended to demonstrate improved anatomical consistency, accurate recognition of subtle but clinically meaningful findings such as patellofemoral malalignment and mild joint effusion, and coherent alignment between imaging observations and diagnostic reasoning. These underscore the indispensable role of high-quality dataset for reliable knee joint report generation.

4 Conclusion

In this work, we present KJD-25K, a large-scale multi-modal knee MRI dataset and evaluation benchmark designed to support clinically aligned radiology report generation. By integrating expert-curated image–text pairs, structured diagnostic labels, and clinically grounded evaluation criteria, KJD-25K enables systematic training and assessment of MLLMs. Furthermore, experimental results show that KJD-25K improves the report-generation performance of MLLMs, achieving more coherent report content and better alignment with expert reference reports compared with general-purpose baselines. We believe KJD-25K provides a practical and scalable foundation toward clinically meaningful knee disease assessment.

Acknowledgements Ben Niu was supported by the National Natural Science Foundation of China [Grant 72334004], Natural Science Foundation of Guangdong Province [Grant 2024A1515011712]. Along He was supported by the Guangdong Basic and Applied Basic Research Foundation [Grant 2026A1515011406] and the Youth S&T Talent Support Programme of Guangdong Provincial Association for Science and Technology [Grant SKXRC2026813]. Yihui Li was supported by the Shenzhen University 2026 Graduate Student Independent Innovation Achievement Cultivation Project [Grant SZUGS2026ZZ75]. The Intelligent Computing Center of Shenzhen University provided computational support.

Disclosure of Interests The authors have no competing interests in this paper.

References

1. Angelone, F., Ciliberti, F.K., Tobia, G.P., Jónsson Jr, H., Ponsiglione, A.M., Gislason, M.K., Tortorella, F., Amato, F., Gargiulo, P.: Innovative diagnostic approaches for predicting knee cartilage degeneration in osteoarthritis patients: a radiomics-based study. Information Systems Frontiers **27**(1), 51–73 (2025)
2. Anthropic: Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5/ (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S.,

Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
5. Desai, A.D., Schmidt, A.M., Rubin, E.B., Sandino, C.M., Black, M.S., Mazzoli, V., Stevens, K.J., Boutin, R., Re, C., Gold, G.E., et al.: Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
6. Gill, T.K., Mittinty, M.M., March, L.M., Steinmetz, J.D., Culbreth, G.T., Cross, M., Kopec, J.A., Woolf, A.D., Haile, L.M., Hagins, H., et al.: Global, regional, and national burden of other musculoskeletal disorders, 1990–2020, and projections to 2050: a systematic analysis of the global burden of disease study 2021. The Lancet Rheumatology **5**(11), e670–e682 (2023)
7. Google: Gemini 2.5: Our most intelligent ai model. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/ (2025)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
9. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
10. Khan, S., Khawer, M.A., Zhong, J., Qureshi, R., Asim, M., Chen, W.: Advancing deep learning based knee cartilage segmentation in mri: innovations, challenges and applications. Osteoarthritis and Cartilage Open p. 100702 (2025)
11. Lester, G.: The osteoarthritis initiative: a nih public–private partnership. HSS journal **8**(1), 62–63 (2012)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
13. Mahum, R., Irtaza, A., El-Meligy, M.A., Sharaf, M., Tlili, I., Butt, S., Mahmood, A., Awais, M.: A robust framework for severity detection of knee osteoarthritis using an efficient deep learning model. International Journal of Pattern Recognition and Artificial Intelligence **37**(07), 2352010 (2023)
14. Mead, K., Cross, T., Roger, G., Sabharwal, R., Singh, S., Giannotti, N.: Mri deep learning models for assisted diagnosis of knee pathologies: a systematic review. European Radiology **35**(5), 2457–2469 (2025)
15. OpenAI: Introducing gpt-5. https://openai.com/zh-Hans-CN/index/introducing-gpt-5/ (2025)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Parkar, A.P., Adriaensen, M.E.A.P.M.: Esr essentials: Mri of the knee—practice recommendations by essr. European Radiology **34**, 6590–6599

(2024). https://doi.org/10.1007/s00330-024-10706-7, https://doi.org/10.1007/s00330-024-10706-7
18. Price, B., Adleberg, J., Thomas, K., Zaiman, Z., Mansuri, A., Brown-Mulry, B., Okecheukwu, C., Gichoya, J., Trivedi, H.: Emory knee radiograph (mrkr) dataset. arXiv preprint arXiv:2411.00866 (2024)
19. Qiu, Z., Xie, Z., Lin, H., Li, Y., Ye, Q., Wang, M., Li, S., Zhao, Y., Chen, H.: Learning co-plane attention across mri sequences for diagnosing twelve types of knee abnormalities. Nature Communications **15**(1), 7637 (2024)
20. Sambara, S., Kim, S.E., Zhang, X., Luo, L., Johri, S., Baharoon, M., Ro, D.H., Rajpurkar, P.: 3dreasonknee: Advancing grounded reasoning in medical vision language models. In: Biocomputing 2026: Proceedings of the Pacific Symposium. pp. 99–113. World Scientific (2025)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Steinmetz, J.D., Culbreth, G.T., Haile, L.M., Rafferty, Q., Lo, J., Fukutaki, K.G., Cruz, J.A., Smith, A.E., Vollset, S.E., Brooks, P.M., et al.: Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the global burden of disease study 2021. The Lancet Rheumatology **5**(9), e508–e522 (2023)
23. Wang, D.y., Liu, S.g., Ding, J., Sun, A.l., Jiang, D., Jiang, J., Zhao, J.z., Chen, D.s., Ji, G., Li, N., et al.: A deep learning model enhances clinicians' diagnostic accuracy to more than 96% for anterior cruciate ligament ruptures on magnetic resonance imaging. Arthroscopy: The Journal of Arthroscopic & Related Surgery **40**(4), 1197–1205 (2024)
24. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
25. Wei, M., Lv, Z., Meng, D., He, S., Yang, G., Wang, Z.: Knock-knee diagnosis in chinese adolescents: Expert evaluation and defensive strategies in image analysis—a population study. Computers in Biology and Medicine **185**, 109513 (2025)
26. Xi, Y., Wang, F., Shi, F., Zhou, Q., Li, R., Li, Y., Wang, Y.: A multi-view open-access dataset of paired knee mri for motion artifact removal. Scientific Data **12**(1), 1173 (2025)
27. Xie, Z., Qiu, Z., Li, Y., Chen, X., Ye, Q., Wang, M., Chen, X., Shao, Y., Liu, C., Song, L., et al.: Development of a multi-task deep learning system for classification of nine common knee abnormalities on mri: a large-scale, multicentre, stepwise validation study. EClinicalMedicine **89** (2025)
28. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
29. Yao, L., Liang, X., Liu, J., Qian, B., Xiao, Z., Zhu, D., Feng, J., Zhou, S., Cui, H., Li, S., et al.: Deep learning-enabled segmentation of knee cartilage in conventional magnetic resonance images: Internal and external validation of different models. The Knee (2025)