DAFTAR PUSTAKA

[1] H. A. Alawwad, A. Zafar, A. Alhothali, U. Naseem, A. Alkhathlan, and A. Jamal, “Evaluating multimodal large language models on educational textbook question answering,” *Proc. Int. Generative AI Comput. Lang. Model. Conf. (GACLM)*, 2025.

[2] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, “LVLM-eHub: A comprehensive evaluation benchmark for large vision-language models,” *arXiv preprint* arXiv:2306.09265, Jun. 2023.

[3] J. Yi, J. Yin, J. Xu, P. Bao, Y. Wang, W. Fan, and H. Wang, “ImageRef-VL: Enabling contextual image referencing in vision-language models,” *arXiv preprint* arXiv:2501.12418, Jan. 2025.

[4] G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, and R. Ji, “Cheap and quick: Efficient vision-language instruction tuning for large language models,” *arXiv preprint* arXiv:2305.15023, Oct. 2023.

[5] R. Janssens, P. Wolfert, T. Demeester, and T. Belpaeme, “Integrating visual context into language models for situated social conversation starters,” *IEEE Trans. Affective Comput.*, vol. 16, no. 1, pp. 223–236, 2025.

[6] L. Yan, L. Zhao, V. Echeverria, Y. Jin, R. Alfredo, X. Li, D. Gašević, and R. Martinez-Maldonado, “VizChat: Enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots,” in *Artificial Intelligence in Education 2024*, LNCS, vol. 14830, pp. 180–193, 2024.

[7] E. Markin, V. Zuparova, and A. Martyshkin, “Integration of large language models and computer vision algorithms in LMS: A methodology for automated verification of software tasks and multimodal analysis of educational data,” *Proc. Int. Russian Smart Industry Conf. (SmartIndustryCon)*, vol. 2025, pp. 777–781, 2025.

[8] M. Y. Lee, “Building multimodal AI chatbots,” Bachelor’s thesis, Apr. 2023.

[9] A. Mohammed, “Multimodal AI architectures: Integrating vision and language for enhanced scene understanding,” *Int. J. Multidisciplinary Res. Sci. Eng. Technol.*, vol. 8, no. 5, pp. 8591–8606, May 2025.

[10] C. Chen, D. Han, and C.-C. Chang, “MPCCT: Multimodal vision-language learning paradigm with context-based compact transformer,” *Pattern Recognition*, vol. 147, art. 110084, Mar. 2024.

[11] J. Chan and Y. Li, “Enhancing higher education with generative AI: A multimodal approach for personalised learning,” *New Technology in Education and Training*, LNEDT, pp. 50–57, 2025.

[12] W. Wang et al., “Image as a foreign language: BEiT pretraining for vision and vision-language tasks,” *arXiv preprint* arXiv:2208.10442, Aug. 2022.

[13] A. Bewersdorff et al., “Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education,” *Learning and Individual Differences*, vol. 118, art. 102601, 2025.

[14] S. Ahmed, M. T. Younes, A. Moustafa, A. Allam, and H. Moustafa, “MSA at ImageCLEF 2025 multimodal reasoning: Multilingual multimodal reasoning with ensemble vision language models,” *CLEF 2025 Working Notes*, 2025.

[15] X. Xu et al., “Development and evaluation of a retrieval-augmented large language model framework for enhancing endodontic education,” *Int. J. Med. Inform.*, vol. 203, art. 106006, Nov. 2025.

[16] S.-V. Oprea and A. Bâra, “Transforming product discovery and interpretation using vision–language models,” *J. Theor. Appl. Electron. Commer. Res.*, vol. 20, no. 3, art. 191, 2025.

[17] Y. Qin, J. Chang, L. Li, and M. Wu, “Enhancing gastroenterology with multimodal learning: The role of large language model chatbots in digestive endoscopy,” *Frontiers in Medicine*, vol. 12, art. 1583514, 2025.

[18] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, “MultiModal-GPT: A vision and language model for dialogue with humans,” *arXiv preprint* arXiv:2305.04790, Jun. 2023.

[19] A. Mihalache, R. S. Huang, M. M. Popovic, et al., “Accuracy of an artificial intelligence chatbot’s interpretation of clinical ophthalmic images,” *JAMA Ophthalmology*, vol. 142, no. 4, pp. 321–326, 2024.

[20] A. Vlachos, G. Filandrianos, M. Lymperaiou, N. Spanos, I. Mitsouras, V. Karampinis, and A. Voulodimos, “Analyze-Prompt-Reason: A collaborative agent-based framework for multi-image vision-language reasoning,” *arXiv preprint* arXiv:2508.00356, Aug. 2025.

[21] J. Swacha and M. Gracel, “Retrieval-augmented generation chatbots for education: A survey of applications,” *Applied Sciences*, vol. 15, no. 8, art. 4234, Apr. 2025.

[22] J. Qi, D. Li, S. Sun, X. Li, W. Li, and Y. Jiang, “RoRA-VLM: Robust retrieval-augmented vision–language models,” *arXiv preprint* arXiv:2410, Oct. 2024.

[23] N. N. Bhat, J. Mondal, and S. Sarkar, “ExpertNeurons at SciVQA 2025: RAVQA-VLM—retrieval-augmented VQA for scholarly document processing,” *Proc. 5th Workshop on Scholarly Document Processing*, pp. 221–229, 2025.

[24] V. N. Rao et al., “RAVEN: Multitask retrieval-augmented vision–language learning,” *arXiv preprint* arXiv:2406, Jun. 2024.

[25] H. Han et al., “Retrieval-augmented generation with graphs (GraphRAG),” *arXiv preprint* arXiv:2406, Jan. 2025.

[26] M. M. Abootorabi et al., “Ask in any modality: A survey of multimodal retrieval-augmented generation,” *Findings of ACL*, pp. 16776–16809, 2025.

[27] N. Drushchak et al., “Multimodal retrieval-augmented generation: Unified information processing across text, image, table, and video modalities,” *Proc. MAGMaR Workshop (ACL 2025)*, 2025.

[28] P. Xia et al., “MMed-RAG: Multimodal retrieval-augmented generation for medical vision–language models,” *arXiv preprint* arXiv:2403, Mar. 2025.

[29] P. Xia et al., “RULE: Reliable multimodal retrieval-augmented generation for factuality in medical vision–language models,” *arXiv preprint* arXiv:2310, Oct. 2024.

[30] A. Martin et al., “Seeing through the MiRAGE: Evaluating multimodal retrieval-augmented generation,” *arXiv preprint* arXiv:2510, Oct. 2025.

[31] M. Mortaheb et al., “RAG-Check: Evaluating multimodal retrieval-augmented generation performance,” *arXiv preprint* arXiv:2501, Jan. 2025.

[32] W. Hu et al., “MRAG-Bench: Vision-centric evaluation for retrieval-augmented multimodal models,” *Proc. ICLR*, 2025.

[33] J. Yu et al., “MIND-RAG: Multimodal context-aware and intent-aware retrieval-augmented generation for educational publications,” *Proc. ICCV Workshops*, 2025.

[34] Z. Reza et al., “Small models, big support: A local LLM framework for teacher-centric content creation and assessment using RAG and CAG,” *arXiv preprint* arXiv:2506, Jun. 2025.

[35] K. Katharakis, S. Rossi, and R. R. Mukkamala, “Small language models for curriculum-based guidance: AI teaching assistants for personalised learning,” *arXiv preprint* arXiv:2510, 2025.

[36] S. Xing et al., “Re-Align: Aligning vision–language models via retrieval-augmented direct preference optimization,” *arXiv preprint* arXiv:2502.13146, Sep. 2025.

[37] S. Yu et al., “VisRAG: Vision-based retrieval-augmented generation on multi-modality documents,” *arXiv preprint* arXiv:2410.10594, Mar. 2025.

[38] X. Fan et al., “Test-time retrieval-augmented adaptation for vision–language models,” *Proc. ICCV*, 2025.

[39] D. G. Lee, A. Park, H. Lee, H. Nam, and Y. Maeng, “Typed-RAG: Type-aware decomposition of non-factoid questions for retrieval-augmented generation,” *arXiv preprint* arXiv:2503.15879, 2025.

[40] Y. Sun et al., “VisRAG 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation,” *arXiv preprint* arXiv:2510.09733, Oct. 2025.