References

[1] X. Zhou, Y. Zhang, L. Cui, and D. Huang, ‘Evaluating Commonsense in Pre-trained Language Models’. arXiv, Feb. 11, 2021. doi: 10.48550/arXiv.1911.11931.

[2] X. Tang *et al.*, ‘Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners’. arXiv, Jun. 08, 2023. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2305.14825

[3] X. Zhang, D. Ramachandran, I. Tenney, Y. Elazar, and D. Roth, ‘Do Language Embeddings Capture Scales?’ arXiv, Nov. 24, 2020. doi: 10.48550/arXiv.2010.05345.

[4] S. Aroca-Ouellette, C. Paik, A. Roncone, and K. Kann, ‘PROST: Physical Reasoning of Objects through Space and Time’. arXiv, Jun. 07, 2021. doi: 10.48550/arXiv.2106.03634.

[5] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Günnemann, ‘Graphhopper: Multi-Hop Scene Graph Reasoning for Visual Question Answering’. arXiv, Jul. 13, 2021. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2107.06325

[6] G. Aglionby and S. Teufel, ‘Meaningful relations in a common-sense knowledge graph’. Cambridge Open Engage, Nov. 06, 2020. doi: 10.33774/coe-2020-1ljq8.

[7] G. Aglionby and S. Teufel, ‘Faithful Knowledge Graph Explanations in Commonsense Question Answering’, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10811–10817. doi: 10.18653/v1/2022.emnlp-main.743.

[8] A. Chiatti, E. Motta, and E. Daga, ‘Robots with Commonsense: Improving Object Recognition through Size and Spatial Awareness’.

[9] D. Teney, L. Liu, and A. van den Hengel, ‘Graph-Structured Representations for Visual Question Answering’. arXiv, Mar. 30, 2017. doi: 10.48550/arXiv.1609.05600.

[10] K. Zhou, E. Lai, W. B. A. Yeong, K. Mouratidis, and J. Jiang, ‘ROME: Evaluating Pre-trained Vision-Language Models on Reasoning beyond Visual Common Sense’. arXiv, Oct. 30, 2023. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2310.19301

[11] L. Chen *et al.*, ‘LANGUAGE MODELS ARE VISUAL REASONING COORDINATORS’, 2023.

[12] N. Rezaei and M. Z. Reformat, ‘Utilizing Language Models to Expand Vision-Based Commonsense Knowledge Graphs’, *Symmetry*, vol. 14, no. 8, Art. no. 8, Aug. 2022, doi: 10.3390/sym14081715.

[13] P. West *et al.*, ‘Symbolic Knowledge Distillation: from General Language Models to Commonsense Models’. arXiv, Nov. 28, 2022. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2110.07178

[14] J. D. Hwang *et al.*, ‘COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs’. arXiv, Dec. 16, 2021. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2010.05953

[15] A. Radford *et al.*, ‘Learning Transferable Visual Models From Natural Language Supervision’, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Feb. 28, 2024. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[16] A. Singh *et al.*, ‘FLAVA: A Foundational Language And Vision Alignment Model’. arXiv, Mar. 29, 2022. doi: 10.48550/arXiv.2112.04482.

[17] Z.-Y. Dou *et al.*, ‘Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone’. arXiv, Nov. 18, 2022. doi: 10.48550/arXiv.2206.07643.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[19] X. Liu, D. Yin, Y. Feng, and D. Zhao, ‘Things not Written in Text: Exploring Spatial Commonsense from Visual Signals’. arXiv, Apr. 27, 2022. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2203.08075

[20] H. Zhao *et al.*, ‘MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning’. arXiv, Oct. 02, 2023. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/2309.07915

[21] I. Comsa and S. Narayanan, ‘A Benchmark for Reasoning with Spatial Prepositions’, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16328–16335. doi: 10.18653/v1/2023.emnlp-main.1015.

[22] M. Yatskar, V. Ordonez, and A. Farhadi, ‘Stating the Obvious: Extracting Visual Common Sense Knowledge’, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Knight, A. Nenkova, and O. Rambow, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 193–198. doi: 10.18653/v1/N16-1023.

[23] R. Mirzaee, H. R. Faghihi, Q. Ning, and P. Kordjmashidi, ‘SpartQA: : A Textual Question Answering Benchmark for Spatial Reasoning’. arXiv, Apr. 12, 2021. Accessed: Feb. 29, 2024. [Online]. Available: http://arxiv.org/abs/2104.05832

[24] Z. Shi, Q. Zhang, and A. Lipani, ‘StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts’. arXiv, Apr. 18, 2022. Accessed: Feb. 29, 2024. [Online]. Available: http://arxiv.org/abs/2204.08292