References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31, 2015. 9, 23

[2] J Alammar. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 249–257, Online, August 2021. Association for Computational Linguistics. 8

[3] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pages 279–290. PMLR, 2020. 23

[4] John R Anderson, C Franklin Boyle, and Brian J Reiser. Intelligent tutoring systems. *Science*, 228(4698):456–462, 1985. 2

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 21

[6] Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. Causalqa: A benchmark for causal question answering. In *ACL*, 2022. 9, 23

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 6, 21, 23, 24

[8] Angelos Chatzimparmpas, Rafael M Martins, Kostiantyn Kucher, and Andreas Kerren. Stackgenvis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1547–1557, 2020. 9

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 23

[10] Zhuo Chen, Yufen Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. *ArXiv*, abs/2207.12888, 2022. 9, 24

[11] Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. Coarse-to-fine question answering for long documents. In *ACL*, 2017. 9, 23

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2, 9, 24

[13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3, 4, 7, 8, 9, 24

[14] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 2020. 23

[15] Misha Denil, Alban Demiraj, and Nando De Freitas. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*, 2014. 8

[16] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 9, 24

[17] Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020. 2, 18

[18] Alican Dogan and Derya Birant. A weighted majority voting ensemble approach for classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pages 1–6, 2019. 2, 9

[19] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019. 23

[20] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 23

[21] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 2021. 23

[22] Dan Goldwasser and Dan Roth. Learning from natural instructions. *Machine learning*, 94(2):205–232, 2014. 23

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 18

[24] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021. 9, 24

[25] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*, 2022. 4

[26] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 4

[27] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022. 4, 19

[28] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 24

[29] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 9, 23

[30] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *arXiv preprint arXiv:2208.05592*, 2022. 2, 9

[31] Anmol Jain, Aishwary Kumar, and Seba Susan. Evaluating deep neural network ensembles by majority voting cum meta-learning scheme. In *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 2*, pages 29–37. Springer, 2022. 9

[32] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *EMNLP*, 2022. 9

[33] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland, May 2022. Association for Computational Linguistics. 4

[34] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2, 9, 23

[35] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017. 23

[36] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1244–1254, 2021. 4

[37] Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*, 2022. 4

[38] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 4

[39] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR, 2021. 24

[40] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 23

[41] Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang, and Ziwei Liu. Sparse mixture-of-experts are domain generalizable learners. *arXiv preprint arXiv:2206.04046*, 2022. 24

[42] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023. 2

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 4

[44] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2, 3, 9, 18, 24

[45] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 4

[46] Shuang Li, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *ArXiv*, abs/2210.11522, 2022. 9, 24

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 18

[48] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022. 2, 18

[49] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022. 20, 23

[50] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 4

[51] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 4

[52] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023. 2

[53] Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *arXiv preprint arXiv:2201.12382*, 2022. 23

[54] Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *arXiv preprint arXiv:2202.10284*, 2022. 2, 23

[55] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes words and sentences from natural supervision. *ArXiv*, abs/1904.12584, 2019. 2, 9, 24

[56] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 9, 24

[57] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2, 18

[58] John McCarthy et al. *Programs with common sense*. RLE and MIT computation center Cambridge, MA, USA, 1960. 23

[59] Fundamental AI Research Diplomacy Team Meta, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 2022. 2, 9, 24

[60] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021. 23

[61] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021. 23

[62] Hyacinth S Nwana. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277, 1990. 2

[63] OpenAI. Gpt-4 technical report. 2023. 2

[64] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 21, 23

[65] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, 2020. 2, 24

[66] Silviu Pitis, Michael R Zhang, Andrew Wang, and Jimmy Ba. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*, 2023. 2

[67] Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations. *arXiv preprint arXiv:2212.04231*, 2022. 4, 19

[68] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. Open-retrieval conversational question answering. In *ACM SIGIR*, 2020. 23

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4, 9, 24

[70] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 24

[71] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 4, 9, 24

[72] Nazneen Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *ACL*, 2019. 23

[73] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 2019. 23

[74] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 3

[75] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 24

[76] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018. 9, 24

[77] Shailaja Keyur Sampat, Maitreya Patel, Subhasish Das, Yezhou Yang, and Chitta Baral. Reasoning about actions over visual and linguistic modalities: A survey. *arXiv preprint arXiv:2207.07568*, 2022. 9, 23

[78] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 23

[79] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. 2, 9

[80] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020. 23

[81] Benedikt Schmidt, Reuben Borrison, Andrew Cohen, Marcel Dix, Marco Gärtler, Martin Hollender, Benjamin Klöpper, Sylvia Maczey, and Shunmuga Siddharthan. Industrial virtual assistants: Challenges and opportunities. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 794–801, 2018. 2

[82] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*, 2022. 2, 18

[83] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. *arXiv preprint arXiv:2303.01903*, 2023. 4

[84] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 24

[85] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 18

[86] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 2, 9

[87] Derek Sleeman and John Seely Brown. *Intelligent tutoring systems*. London: Academic Press, 1982. 2

[88] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 22

[89] Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. On the importance of diversity in question generation for qa. In *ACL*, 2020. 23

[90] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 2

[91] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 4

[92] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 4

[93] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*, 2016. 9, 23

[94] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. 2

[95] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2, 3, 9, 18, 24

[96] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358, 2021. 9, 24

[97] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*, 2022. 2

[98] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2

[99] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 7, 8, 23

[100] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 2

[101] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. 3

[102] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. 2, 9

[103] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706, 2019. 9, 18, 23

[104] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021. 5

[105] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, 2022. 2, 24

[106] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *ArXiv*, abs/1910.01442, 2019. 2, 24

[107] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *ArXiv*, abs/1810.02338, 2018. 2, 9, 24

[108] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 4

[109] Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. Conversational question answering: A survey. *Knowledge and Information Systems*, 2022. 9, 23

[110] Rufai Yusuf Zakari, Jim Wilson Owusu, Hailin Wang, Ke Qin, Zaharaddeen Karami Lawal, and Yuezhou Dong. Vqa and visual reasoning: An overview of recent datasets, methods and challenges. *arXiv preprint arXiv:2212.13296*, 2022. 2, 9, 23

[111] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2018. 2, 9, 24

[112] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Neural Information Processing Systems*, 2021. 9, 24

[113] Andy Zeng, Adrian S. Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Peter R. Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *ArXiv*, abs/2204.00598, 2022. 2, 9, 24

[114] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 4, 6

[115] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022. 24

[116] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021. 23

A Experimental Details

In this section, we elaborate on our training and evaluation details, prompt templates, and more qualitative examples for analysis.

A.1 Datasets

Our experiments are conducted on a challenging suite of three diverse visual reasoning tasks, including outside knowledge VQA, visual entailment, and visual spatial reasoning. For each task, we select the following dataset respectively.

Visual Question Answering v2 [23] (VQA v2) is a large-scale benchmark containing over 1 million images from the COCO dataset and more than 250,000 human-generated question-answer pairs. The dataset is designed to test the ability of machine learning models to understand both the visual content of an image and the meaning behind natural language questions. The questions in VQA v2 cover a wide range of topics and are often open-ended, requiring models to reason and generalize about the world. VQA v2 has been widely used to evaluate the performance of state-of-the-art models in the field of computer vision and natural language processing.

Augmented Outside Knowledge VQA [82] (A-OKVQA) contains about 25k questions paired with both multiple choice (MC) answer options. Unlike most existing VQA datasets, the questions in A-OKVQA cannot often be answered by querying the knowledge base, but rather involve some type of commonsense reasoning and outside knowledge about the situation portrayed in the image.

Outside Knowledge VQA [57] (OK-VQA) includes more than 14,000 questions that require external knowledge to answer. The answers are provided in free-text direct answer form. Both A-OKVQA and OK-VQA sample images from the COCO dataset, with no overlapping.

e-SNLI-VE [17] dataset is an extended version of SNLI-VE dataset [103], which contains about 190k question pairs and human-annotated natural language explanations for the ground-truth labels. The text premise provides a statement about the contents of the image. The task is to determine whether the statement is true or false based on the image content.

Visual Spatial Reasoning [48] (VSR) consists of 65 spatial relations (*e.g.,* under, in front of, facing, *etc.*) of instances in images. VSR has more than 10k question pairs, associated with 6940 images from MS COCO [47].

A.2 Finetuning Details

We adopt pretrained BLIP [44][2] and OFA [95][3] as VLMs, and freeze their parameters without updating. The finetuning only happens on the language model part. The training set of each dataset is used for finetuning. We use the whole training set unless otherwise specified in low-data finetuning discussion.

We use an AdaFactor optimizer [85] at the learning rate of 1e-4 for all Cola-FT experiments. The batch size is by default set to 16, though we find Cola-FT insensitive to batch size. We finetune and evaluate the models on NVIDIA V100 or A100 GPUs. The finetuning ranges from 1 hour to about 15 hours, varying by the dataset.

Following the common experiment protocols, we employ a teacher forcing and greedy decoding strategy for finetuning.

A.3 Evaluation Details

As specified, we use the validation or test set multiple choice accuracy as the evaluation metric. In A-OKVQA, we report `val/test` accuracy, and `val` accuracy in e-SNLI-VE, `test` (zero-shot split) accuracy in VSR. For simplicity and consistency, we evaluate ablation experiments on A-OKVQA

[2]BLIP: https://github.com/salesforce/BLIP

[3]OFA: https://huggingface.co/OFA-Sys

validation set. Following the common experiment protocols [27, 67], we report the single run results for performance comparison.

The exemplars at the inference of Cola-Zero are randomly sampled from the training set, i.e., supposedly help the LM learn the input data distribution and output format but do not leak relevant information to the evaluation question.

A.4 A-OKVQA Direct Answer Results

In addition to MC accuracy, we present the direct answer (DA) accuracy of models on the A-OKVQA validation set in Tables 6 and 7.

	FLAN-T5-Small	FLAN-T5-Base	FLAN-T5-XL	FLAN-T5-XXL
Cola-FT	56.5	60.6	64.1	65.4
Cola-Zero (2-shot)	30.3	34.6	57.6	61.0
Cola-Zero (0-shot)	28.6	36.0	55.0	59.3

Table 6: **A-OKVQA validation set DA performance.** Extension of Figure 5.

	1-shot	2-shot	3-shot	4-shot
Cola-Zero	60.2	61.0	60.7	59.2

Table 7: **Cola-Zero in-context few-shot learning DA performance on A-OKVQA validation set.** Extension of Figure 6.

A.5 Qualitative Examples

In this section, we provide more qualitative examples on A-OKVQA (Figure 8), e-SNLI-VE (Figure 9), and VSR (Figure 10) datasets.

Due to the large span of the three figures, for better visibility, we put the detailed description directly in each figure's caption part. We illustrate how Cola-FT and Cola-Zero process the VLMs answers in each example. Overall, in these examples, we can observe that even if BLIP and OFA provide wrong answers, Cola can still present the correct answer based on the captions provided by OFA and BLIP, as well as the choice set. This may illustrate how Cola amazingly accomplishes visual reasoning tasks via coordinating BLIP and OFA.

A.6 Failure Cases

In Figure 11, we provide a few failed cases to analyze the specific behavior of Cola.

The leftmost example's correct answer is *kayaking*, but there are no hints from OFA and BLIP's answers and captions. Therefore Cola-Zero incorrectly provides the answer *OFA* without sufficient information as hints, while surprisingly Cola-FT answered correctly from OFA's *boating* answer.

The left example again has insufficient information from captions. While BLIP answers *no* and OFA answers *yes*, Cola-FT chooses to answer *maybe*, which looks natural but unfortunately picks the wrong choice.

The right example's captions contain enough information this time. But both Cola-FT and Cola-Zero are misled by BLIP's wrong answer *no parking*.

The rightmost example also has insufficient information from captions. In this situation, Cola has no choice but to believe either BLIP or OFA's answer, but it mistakenly prefers BLIP's wrong answer.

A.7 Prompt Templates

Across three datasets, the prompt template is roughly the same, with minor differences mainly in the format of the questions and choices. We list the prompt templates adopted in A-OKVQA and e-SNLI-VE/VSR in Table 8 and Table 9, respectively.

Question	Why might people sit here?	The room can be described as what?	In what type of location are they playing with the body board?	What is in front of the monitor?
OFA caption	colorful umbrellas on the riverwalk	living room layout and decor medium size how to decorate a small living room dining combo mant	person, left, and person look at a painting of a great white shark.	a desk with a computer, a lamp, a laptop, and a plant.
BLIP caption	a colorful umbrella umbrella with colorful umbrellas	a dining room table with a glass table and chairs	a man holding a surfboard while another man is standing next to him	a desk with a computer and a lamp
Choices	['to testify', '<u>to rest</u>', 'to shop', 'get tattoo']	['<u>tidy</u>', 'messy', 'on fire', 'destroyed']	['<u>room</u>', 'beach', 'park', 'store']	['<u>keyboard</u>', 'phone', 'mouse', 'headphones']
OFA answer	to eat	living room	bedroom	a keyboard
BLIP answer	yes	dining room	beach	monitor
Cola-Zero answer	to rest	tidy	beach	keyboard
Cola-FT answer	to rest	tidy	room	keyboard

Figure 8: **A-OKVQA qualitative examples.** Leftmost: LM doesn't use BLIP and OFA's answers, but may observe from captions to derive the correct final answer. Left: As shown on the left, LM does not follow the wrong answers from OFA and BLIP but gets the correct answers from captions. Right: With both OFA and BLIP answering incorrectly, LM derives the correct one from both VLMs' captions and answers. Rightmost: After assessing the questions, answers, and captions, LM goes with OFA's answer and rewrites it to match the expression in the choices. The correct choices are <u>underlined</u>. Cola-Zero answers are given in zero-shot settings.

VQA Prompt Template

Answer the following multiple-choice question by OFA and BLIP's description and their answers to the visual question. OFA and BLIP are two different vision-language models to provide clues.

OFA's description: `<OFA caption>`
BLIP's description: `<BLIP caption>`

Q: `<Question>`

OFA's answer: `<OFA answer>`
BLIP's answer: `<BLIP answer>`

Choices: `<Choices to the question>`

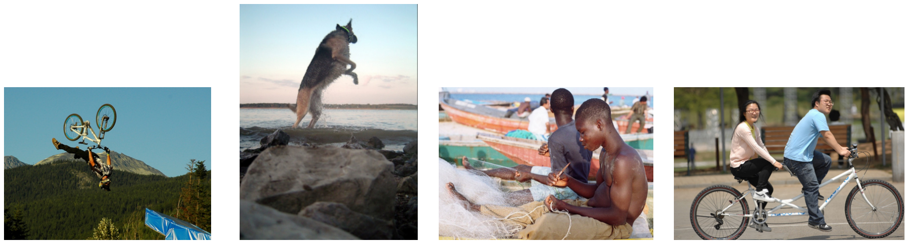
A:

Table 8: **VQA prompt template for the LM, for VQA v2 / OK-VQA / A-OKVQA.** The LM is instructed to coordinate VLMs. Each question set defines *visual context*, *question with choices*, and *plausible answers*.

A.8 Parameter-efficient Finetuning

To further reduce the computation cost in model adaptation, we explored parameter-efficient finetuning (PEFT) techniques to reduce finetuning parameter counts. Specifically, we use $(\mathtt{IA})^3$ [49], which finetunes an overhead of 1 million parameters, equivalent to 0.01% of the full parameters of FLAN-T5-XXL.

Compared to full finetuning, $(\mathtt{IA})^3$ requires more iterations to converge. The performance of a $(\mathtt{IA})^3$ finetuned FLAN-T5-XXL model is on par with a fully finetuned FLAN-T5-Small (80 million



Question	Does the image describe " A professional daredevil "?	Does the image describe " the dog is a shitz " ?	Does this image describe "Two twenty-somethings prepare to catch salmon while other older men catch catfish" ?	Does this image describe "A little girl gets hit by a woman riding a bike" ?
OFA caption	person doing a flip on a mountain bike	a dog jumping out of the water.	men repairing fishing nets on the beach in zanzibar, tanzania	a man and a woman on a tandem bike
BLIP caption	a man doing a trick on a bike in the air	a dog jumping over rocks in the water	a man sitting on a boat with a fishing net net	a man and woman riding a bicycle in a parking lot
Choices	['yes', 'maybe', 'no']	['yes', 'maybe', 'no']	['yes', 'maybe', 'no']	['yes', 'maybe', 'no']
OFA answer	yes	no	yes	yes
BLIP answer	yes	no	yes	no
Cola-Zero answer	yes	no	no	no
Cola-FT answer	maybe	maybe	maybe	no

Figure 9: **e-SNLI-VE qualitative examples.** Leftmost: As the connection to *daredevil* is not obvious in BLIP and OFA's captions, although Cola-Zero is misled, Cola-FT correctly answers *maybe*. Left: Similar to the left example, Cola-FT answer correctly as no obvious connections are seen from the captions to this question. Right: Similar to the left example, the fact of *catch catfish* is not reasonable from the captions, Cola-FT picks the correct answer *maybe*. Rightmost: As *girl gets hit* is not obvious in BLIP and OFA's captions and answers, Cola-Zero and Cola-FT both follow BLIP to choose the correct answer *no*. The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.

parameters) counterpart (Figure 5). Notably, the former is associated with more computation and memory footprint as a consequence of more parameters in the forward pass.

A.9 Extended Ablation Studies

Do caption labels offer useful information to LLM? How would more prompt variations affect the performance of Cola? We tested Cola-Zero with and without caption labels on A-OKVQA validation set, observing a slight decrease in performance when without them (70.39% w/t vs. 69.97% w/o). More ablative experiments showed that removing the VLM's answer labels led to a substantial drop in performance (70.39% w/t vs. 67.62% w/o). Removing the model characteristic descriptions also led to a decrease (70.39% w/t vs. 68.37% w/o).

Do longer image captions improve reasoning performance? On A-OKVQA validation set, we tested longer image descriptions (>50 tokens) but found no gain compared to Cola or single VLMs. Longer captions decreased FLAN-T5+OFA's accuracy by 0.61% and FLAN-T5 with BLIP by 0.69% on the A-OKVQA validation set. Cola (captions <30 tokens) reached 77.73%, outperforming individual VLMs. Longer captions lacked meaningful visual context, possibly due to short text and image pairs in their training datasets. This experiment reaffirms Cola's effectiveness in aggregating individual VLM functionalities.

B Extended Related Works

B.1 Finetuning Large Language Models

Large language models [7, 64, 5] pretrained on massive amounts of unstructured data have gradually demonstrated great performance by finetuning on additional task-specific instances. Finetuning

Question	Does this image describe "The truck contains the elephant" ?	Does this image describe "The bed is under the handbag" ?	Does this image describe "The couch is behind the hot dog" ?	Does this image describe "The bowl contains the banana" ?
OFA caption	an elephant being transported on a truck in sri lanka	a black and white tuxedo cat with a white nose, yellow eyes, and white	person enjoying a meal by the fire	bananas and mangoes in a bowl
BLIP caption	a truck with a large elephant in the back of it	a black cat laying on a bed with a pillow	a man sitting on a couch with a plate of food	a bowl of fruit is shown in this bowl
Choices	['yes', 'no']	['yes', 'no']	['yes', 'no']	['yes', 'no']
OFA answer	yes	no	yes	yes
BLIP answer	no	no	yes	no
Cola-Zero answer	no	no	no	yes
Cola-FT answer	yes	no	no	yes

Figure 10: **VSR qualitative examples.** Leftmost: As OFA caption mentioned *elephant being transported* and OFA provides the correct answer, Cola-FT follows OFA's choice. Left: As OFA and BLIP provide the same answer, Cola-Zero and Cola-FT follow the choice. Right: As the captions do not provide obvious information, even BLIP and OFA provide the same answer, Cola-Zero and Cola-FT are not misled to the wrong choice. Rightmost: As the captions provide strong clue *bananas in a bowl*, although BLIP's answer is incorrect, Cola-Zero and Cola-FT still choose the correct answer. The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.

Question	What are the people doing in the water?	Does the image describe " The man is making a vase"?	What kind of zone is this bike parked in?	Does this image describe "The motorcycle is beside the truck" ?
OFA caption	black and white photo of a man on a bike looking at a canoe in the river	person on the potter's wheel	a city made by people bucharest	men walking past a truck in kabul, afghanistan.
BLIP caption	a man and woman on a bike in a park	a man is sitting on a chair and is using a wheel	a bicycle parked next to a pedestrian crossing sign	a man walking down the street in a city
Choices	['surfing', 'fishing', 'kayaking', 'swimming']	['yes', 'maybe', 'no']	['temporary', 'pedestrian', 'no parking', 'handicap']	['yes', 'no']
OFA answer	boating	yes	pedestrian	yes
BLIP answer	swimming	no	no parking	no
Cola-Zero answer	OFA	no	no parking	no
Cola-FT answer	kayaking	maybe	no parking	no

Figure 11: **Failed cases.** The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.

a large language model can be considerably more sample efficient than re-training from scratch, although acceptable performance may still require a considerable quantity of data [88]. Recent

e-SNLI-VE / VSR Prompt Template

Answer the following multiple-choice question by OFA and BLIP's description and their answers to the visual question. OFA and BLIP are two different vision-language models to provide clues.

OFA's description: `<OFA caption>`
BLIP's description: `<BLIP caption>`

Q: does the image describe `<hypothesis>` ?

OFA's answer: `<OFA answer>`
BLIP's answer: `<BLIP answer>`

e-SNLI-VE Choices: [yes, no, maybe]
VSR Choices: [yes, no]

A:

Table 9: **e-SNLI-VE/VSR prompt template for the LM.** The LM is instructed to coordinate VLMs. Each question set defines *visual context*, *hypothesis*, and *plausible answers*.

	Accuracy	# Finetuning Params
Finetuning	77.73	11B (100%)
PEFT, $(\texttt{IA})^3$	63.76	1M (0.01%)

Table 10: **$(\texttt{IA})^3$ [49] parameter-efficient tuning (PEFT) performance.** We finetune a FLAN-T5-XXL model on the A-OKVQA training set and evaluate it on the A-OKVQA validation set.

works have finetuned task-specific models that demonstrate amazing capabilities in many real-world applications, such as Copilot for program synthesis [9].

B.2 Instruction-based Learning

Recent advances in the capabilities of language models have piqued researchers' curiosity in the field of instruction-based learning [22, 58, 80, 20]. The core of instruction-based learning is to explore the knowledge of the language model itself. In contrast to prompt learning to stimulate the language model's ability to complete blanks, instruction tuning more focuses on activating the language model's comprehension by giving obvious instructions to models and expecting correct feedback. Earlier work [61] finetune BART [40] using instructions and few-shot examplars in question answering, text classification, and text modification. Their findings suggest that few-shot instruction tuning improves performance on unseen tasks. [60] finetunes GPT-2 Large and also observes that few-shot examplar instruction tuning could improve performance. [78] finetunes T5-11B with more diverse instruction templates and observe similar improvements in zero-shot learning. More recent work [99] performs large-scale experiments with a 137B FLAN-T5 model and instruction-tune it on over 60 datasets verbalized via instruction templates. They observe FLAN-T5 substantially improves over zero-shot GPT-3 (175B) on 20 of 25 evaluation datasets. OpenAI also releases InstructGPT [64] based on GPT-3 [7], it makes use of human annotations to steer desired model behavior through both instruction tuning and reinforcement learning of human feedback. They discover that InstructGPT is favored by humans over unmodified GPT-3.

B.3 Visual Reasoning

Beyond the uni-modal reasoning tasks such as question answering (QA) [93, 35, 11, 73, 72, 19, 68, 14, 89, 21, 116, 109, 6], visual reasoning requires models to not only understand and interpret visual information but also to apply high-level cognition to derive rational solutions [34, 29, 3, 53, 54, 77, 110]. Several tasks have been introduced to address visual reasoning, such as visual question answering (VQA) [1], in which models are expected to provide answers to questions related to an image and visual entailment (VE) [103], where the model is required to determine the similarity or relationship between a given image and a description. Classic visual reasoning methods have employed an image encoder and a text encoder, along with a reasoning block that utilizes attention mechanisms [111, 65, 112, 96], neuro-symbolic methods [107, 55, 106], or external knowledge [56, 24, 10] to perform reasoning.

Recent progress in large pre-trained models has led to the development of language models (LMs) that possess exceptional commonsense reasoning capabilities [71, 13, 12, 70]. These models can potentially replace the reasoning block in visual reasoning tasks, and LMs' lack of perception can be compensated by incorporating multiple vision-language models (VLMs) trained on different domains [69, 95, 44]. For example, PICa [105] converts the image into captions that GPT-3 [7] can understand, and adapts GPT-3 to solve the VQA task in a few-shot manner by providing a few in-context VQA examples. However, there is still a lack of research on how to harness the collective power of these complementary VLMs for visual reasoning tasks.

B.4 Model Ensembling

Model ensembling is a powerful machine learning technique that combines the predictions of multiple models to improve the overall performance of a given task [16]. Classic model ensembling methods include simple averaging, weighting the predictions based on model performance, and stacking the models. By combining the predictions of multiple models, ensembling can reduce the variance and bias of the final predictions, resulting in a more robust and accurate model [76]. Ensemble methods have been shown to perform well in a wide range of tasks, including image classification, natural language processing, and time series forecasting. However, when it turns to multimodal tasks such as visual reasoning, a simple combination is not applicable to heterogeneous models as their inputs and outputs vary.

The Mixture-of-Experts (MoE) [84, 75, 115, 39, 41] can be conceptualized as a model ensemble strategy implemented at the level of network architecture. MoE-based multi-modal models [28] excel in leveraging the specific strengths of each expert, thereby delivering the performance that often outstrips that of any individual expert. In these networks, the credibility of each expert's output is dynamically weighted, facilitating a comprehensive and nuanced response to multimodal tasks.

However, even within this sophisticated framework, challenges can arise, particularly when managing heterogeneous pre-trained multimodal models. To address this problem, an innovative approach known as Socratic Models (SMs) [113] has been proposed. SMs employ prompt engineering to guide these diverse models through multimodal discussions, effectively combining their varied knowledge. This method promotes a more harmonious and effective integration of different models, enhancing the ensemble's ability to handle complex tasks.

With a similar goal, [46] proposes a closed-loop iterative consensus optimization method to utilize the strengths of individual models. However, previous methods do not fully explore the potential of a centralized solution or adapt to the separate functionalities of different models, particularly in the visual reasoning scenario. Recent studies, such as CICERO [59], have shown that language models possess strong capabilities in coordinating multiple agents, which inspires us to reorganize pre-trained multimodal models with a focus on the language models.

Broader Impact

This study inherits ethical risks of biases from pretrained VLMs and LMs, depending on their training data. We suggest the users consider the possible biases in reasoning and prompt the model to interpret its predictions in natural languages when necessary.