CoG-DQA: Chain-of-Guiding Learning with Large Language Models for Diagram Question Answering

Supplementary Material

1. Design of the Guiding Head

As mentioned in the main part, in order to ensure the diversity of prompts, we manually define five different guiding heads for each stage of the LLM guiding component, see Fig. 1 for details.

2. Additional Experiments

2.1. Statistics of Datasets

The detailed statistics of four datasets are shown in Tab. 1.

2.2. Baseline Methods

QA-based models: Antol *et al*. [2] first proposed the task of free-form and open-ended Visual Question Answering (VQA), with some attempts at simple deep models. Ben-Younes *et al*. [3] proposed a multimodal tensor-based decomposition to efficiently parametrize bilinear interactions between visual and textual representations. Gao *et al*. [6] alternatively passed information within and across different modalities based on an inter-modality and intra-modality attention mechanisms. Haurilet *et al*. [8] presented a new model for visual reasoning, aimed at capturing the interplay among individual objects in the image represented as a scene graph. Kembhavi *et al*. [10] introduced Diagram Parse Graphs (DPG) as representation to model the structure of diagrams, and defined syntactic parsing of diagrams as learning to infer DPGs for diagrams. Kim *et al*. [11] introduced a algorithm contains visual and textual information in an abstract and integrated way. Kim *et al*. [12] proposed BAN, which can find bilinear attention distributions to utilize given vision-language information seamlessly. ViLT [13] is a model that simplifies the processing of visual inputs to a convolution-free approach similar to how textual inputs are processed. Li *et al*. [14] has demonstrated that certain attention heads of a visually grounded language model actively ground elements of language to image regions. Lu *et al*. [16] developed a strong baseline, Patch-TRM, which parses the diagram in a pyramid layout and applies cross-modal transformers, with attention mechanism to learn the meaningful joint diagram-question feature. Wang *et al*. [22] introduced the tasks of diagram classification (DC) and diagram question answering (DQA) based on the new dataset, and proposed the Diagram Paring Net (DPN) that focuses on analyzing the topological structure and text information of diagrams. Yu *et al*. [25] developed a Multi-modal Factorized Bilinear (MFB) pooling approach to efficiently and effectively combine multi-modal features, which resulted in superior performance for VQA compared with other bilinear pooling approaches. Yu *et al*. [26] proposed MCAN, which consists of Modular Co-Attention (MCA) layers cascaded in depth. Each MCA layer models the self-attention of questions and images, as well as the guided-attention of images jointly using a modular composition of two basic attention units. Yuan *et al*. [27] proposed a novel structural parsing integrated Hierarchical Multi-Task Learning (HMTL) model for diagram question answering based on a multi-modal transformer framework. Zheng *et al*. [29] designed a novel cross-modality relevance module that is used in an end-to-end framework to learn the relevance representation between components of various input modalities under the supervision of a target task.

Finetune-based models: For the first time, Gomez-Perez *et al*. [7] taped the potential of transformer language models and bottom-up and top-down attention to tackle the language and visual understanding challenges that Textbook Question Answering task entails. Ma *et al*. [18] proposed a Weakly Supervised learning method for TQA (WSTQ), which regards the incompletely accurate results of essential intermediate procedures for this task as supervision to develop Text Matching (TM) and Relation Detection (RD) tasks and then employs the tasks to motivate itself to learn strong text comprehension and excellent diagram semantics respectively. Raffel *et al*. [21] explored the landscape of transfer learning techniques for NLP by introducing a unified framework that converted all text-based language problems into a text-to-text format. Xu *et al*. [23] proposed a novel model named MoCA, which incorporated Multi-stage domain pretraining and Cross-guided multimodal Attention for the TQA task.

Prompt-based models: Chen *et al*. [5] proposed a model, which is surprisingly effective for semi-supervised learning on ImageNet, useing of big (deep and wide) networks during pretraining and fine-tuning. OpenAI [19] proposed a Transformer-based model and pretrain it to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. LLaVA-series [15] model represents a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding. Yao *et al*. [24] proposed Graph-of-Thought (GoT) reasoning, which modeled human thought processes not only as a chain but also as a graph. Zhang *et al*. [28] proposed Multimodal-CoT that incorporates language and vision modalities into

LLM Guiding 1

(1) Based on the above contents, What additional knowledge is needed to solve this question?
(2) List what other background knowledge the model needs in order to solve thisquestion.
(3) Enumerate the additional foundational knowledge required to address this question.
(4) Outline the supplementary background information essential for solving this question.
(5) Can you specify the additional foundational knowledge needed to address this question comprehensively?

LLM Guiding 2

(1) Which objects in the diagram should the model focus on in order to solve related question? List by serial number plus object name, for example 1. Object a; 2. Object b.
(2) In order to solve the problem, which objects in the diagram should the model focus on?
(3) To solve this question, the model should focus on Some key objects in the Diagram.
(4) The model needs to identify the specific potential objects within the diagram to answer the question.
(5) What objects in the diagram should the model prioritize when addressing the relevant question?

LLM Guiding 3

(1) To answer the question, what potential relationships between these objects does the model need to be aware of? List in the following format: 1. (Object 1, Object 2, Relation a) ; 2. (Object 4, Object 5, Relation b).
(2) In order to solve this question, what potential relationships between objects does the model need to focus on?
(3) To address this question, the model should identifying potential relationships between these objects.
(4) To tackle this question, it's crucial for the model to identify the potential relationships between these objects.
(5) What are the essential object relationships the model needs to be aware of to get the right answer?

Figure 1. The design of prompts for each stage of the LLM guiding component. In the stages 2 and 3, the specific format of the answer is only shown in the prompt (1) and is not repeated in other prompts.

Dataset	SQA-I		TQA-DMC		CSDQA		AI2D	
	Diagram	Q&A	Diagram	Q&A	Diagram	Q&A	Diagram	Q&A
Train	6,218	6,218	1,499	6,501	713	1,995	2,535	7,829
Val	2,097	2,097	660	3,285	238	664	259	906
Test	2,017	2,017	590	2,781	238	618	308	978
All	10,332	10,332	2,749	12,567	1,189	3,227	3,101	9,708

Table 1. The number of diagrams and questions of SQA-I, TQA-DMC, CSDQA, and AI2D datasets.

a two-stage framework that separates rationale generation and answer inference.

2.3. Detailed Comparison Results

We conducted more fine-grained comparative experiments on the SQA-I and CSDQA datasets. Tab. 2 shows the performances of the baseline models and our model on the SQA or reconstructed SQA-I dataset. Among them, the upper part of Tab. 2 is the experimental results on the SQA dataset in previously published manuscripts, which divides the questions into several categories. Similarly, we also statistics the experimental results of CoG-DQA under the corresponding categories in the lower part of Tab. 2. As mentioned in the main section, the SQA-I dataset filters questions that contain visual content in the SQA dataset. The amount of data is reduced and the question distribution changes accordingly, so our experimental results cannot be compared fairly with the results in the upper part of Tab. 2. It can be seen that under the new division, CoG-DQA is close to the SotA performance (gray background) under the previous full data scale, and can still reach a comparable level with a 51.28% reduction of samples. However, the CoG-DQA model can have obvious disadvantages in certain subjects, such as language sciences. After analysis, our model is less effective when targeting subjects with less visual content correlation, which is also one of our future research contents.

Tab. 3 shows the detailed results of the CSDQA test split. It can be seen that the CoG-DQA model has achieved optimal results in almost all types of questions. For true-or-false questions, CoG-DQA improves accuracy by 9.85% and 5.97% on easy questions and all questions, respectively.

Model	Learning	Format	NAT	SOC	LAN	NO	G1-6	G7-12	IMG/AVG
MCAN [26]	train set	QCM-A	56.08	46.23	58.09	55.40	51.65	59.72	51.17
Top-Down [1]	train set	QCM-A	59.50	54.33	61.82	62.90	59.79	57.27	54.88
BAN [12]	train set	QCM-A	60.88	46.57	66.64	65.51	56.83	63.94	52.60
DFAF [6]	train set	QCM-A	64.03	48.82	63.55	64.11	57.12	67.17	54.49
ViLT [13]	train set	QCM-A	60.48	63.89	60.27	57.00	60.72	61.90	61.38
Patch-TRM [16]	train set	QCM-A	65.19	46.79	65.55	64.95	58.04	67.50	55.28
VisualBERT [14]	train set	QCM-A	59.33	69.18	61.18	58.54	62.96	59.92	62.17
UnifiedQAbase [21]	zore-shot	QCM-A	68.16	69.18	74.91	77.84	72.98	65.00	61.38
UnifiedQAbase [21]	train set	QCM-A	71.00	76.04	78.91	81.81	77.06	68.82	66.53
GPT-3.5 [5]	zero-shot	QCM-A	74.64	69.74	76.00	77.42	76.80	68.89	67.28
GPT-3.5 [5] w/ CoT	2-shot	QCM-AE	75.44	70.87	78.09	79.93	78.23	69.68	67.43
GPT-4 [19] w/ CoT	few-shot	QCM-AE	85.48	72.44	90.27	92.89	86.66	79.04	71.49
MM-CoT-large [28]	train set	QCM-LE-A	95.91	82.00	90.82	92.89	92.44	90.31	88.80
GoT-T5-large [24]	train set	QCM-LE-A	96.51	82.26	93.61	94.29	93.83	90.86	89.56
Published results on SQA ⇑									
Our results on SQA-I ⇓									
CoG-DQA	train set	QCM-A	76.10	79.45	65.91	76.45	79.71	70.92	78.85
CoG-DQA	train set	QCM-LE-A	**92.06**	**79.83**	**74.73**	**91.65**	**89.78**	**83.08**	**89.32**

Table 2. Accuracy (%) on test split of SQA (above) and SQA-I (below) datasets, along with accuracy scores for different question types, including natural, social, and language sciences, and no context, as well as grades 1-6 and 7-12. Format names: Q = question, C = context, M = multiple options, A = answer, E = explanation, L = lecture.

Models	TF(E)	TF(C)	TF(ALL)	MC(E)	MC(C)	MC(ALL)	ALL
Random	50.00	50.00	50.00	25.00	25.00	25.00	37.50
MFB [25]	53.14	52.08	56.51	34.72	33.33	30.21	43.36
BAN [12]	52.08	52.07	57.29	33.33	28.13	27.34	42.32
MACN [26]	56.60	54.17	59.64	34.03	32.29	29.17	44.41
MUTAN [3]	51.39	57.29	54.43	28.47	27.08	27.86	41.15
CMR [29]	51.56	51.33	51.21	30.20	32.69	30.70	40.91
ISAAQ [7]	59.79	58.62	60.32	41.34	34.48	41.07	50.70
DPN-QA [22]	57.29	59.38	58.85	35.07	33.33	31.77	45.31
WSTQ [17]	59.28	58.62	58.62	41.04	31.03	38.49	48.55
LLaVA-1.5 (7b)[15]	63.47	61.84	62.65	28.97	28.13	28.64	47.31
LLaVA-1.5 (13b)[15]	65.54	63.96	64.43	29.76	28.16	29.36	48.65
GPT-3.5 [5] 0-shot	54.12	51.33	53.21	26.24	25.53	25.68	45.77
GPT-3.5 [5] 2-shot	64.68	65.53	65.02	27.93	26.52	27.04	46.83
GPT-4.0 [19] 2-shot	65.74	**67.37**	66.85	27.58	27.24	27.45	48.03
MM-CoT-large [28]	64.40	57.63	63.11	53.60	42.37	51.46	57.28
CoG-DQA	**75.60**	61.02	**72.82**	**65.60**	**55.93**	**63.75**	**68.28**

Table 3. Accuracy (%) on test split of CSDQA dataset. The best performance is bolded and the second-best performance is underlined. TF denotes True-or-False Questions, and MC denotes Multiple Choice Questions. E indicates easy difficulty questions, and C indicates complex difficulty questions.

However, the performances of the GPT-based models are significant on complex multiple-choice questions but fail on other questions. For multiple choice questions, CoG-DQA improves accuracy by 12%, 13.56%, and 12.29% on easy questions, complex questions, and all questions, respectively. For all the questions, CoG-DQA improves the

Model	Dataset	
	SQA-I	CSDQA
w/ ResNet	78.85	68.28
w/ CLIP	78.26	67.97
w/ DETR	78.82	68.40

Table 4. Accuracy (%) of using different vision encoders on SQA-I and CSDQA datasets.

LLM	Dataset	
	SQA-I	CSDQA
GPT 3.0	77.26	65.84
GPT 3.5	78.85	68.28
text-davinci	79.14	68.03

Table 5. Accuracy (%) of using different LLMs on SQA-I and CSDQA datasets.

accuracy by 11%. This demonstrates the effectiveness of our method.

2.4. Impact of Different Encoders

Visual features from different encoders can impact model performance. Our study assesses three prominent types of visual features: ResNet [9], CLIP [20], and DETR [4]. Both CLIP and DETR represent patch-like features, with DETR primarily rooted in object detection. For a fair comparison, we uniformly set the objects dimension of all visual feature to 50. Tab. 4 presents a comparative analysis of these visual features. Our findings demonstrate that the integration of visual features consistently results in enhanced performance, as opposed to models that depend exclusively on linguistic data. Moreover, different encoders do not produce major differences in experimental performance. Consequently, we default to using ResNet in CoG-DQA due to its good generalization performance and ease of use.

2.5. Impact of LLM

The CoG module mainly uses the guiding chains to transfer the effective knowledge of LLM to special areas, and the knowledge possessed by different types of LLM is also different. In order to explore the effect of different LLMs on performance, we selected three popular LLMs: GPT-3, GPT3.5, and text-davinci-003 and conducted experiments on the SQA-I and CSDQA datasets. Tab. 5 presents a comparative analysis of these LLMs. According to the experimental results, GPT-3.5 has similar performance to text-davinci and is significantly better than the GPT-3.0 model. Due to the assisting and guiding role of LLMs in the CoG-DQA framework, it can be flexibly adjusted to the latest language models. Since GPT-4V has not yet fully opened the interface during our experiments, the combination with the latest models of this type can be one of the future research contents.

2.6. Case Study

In this section, we visualize positive and negative cases of the CoG-DQA model on four datasets. In these cases, we mainly visualize the response of LLMs to manually designed guiding heads, as well as some key feature interaction processes. Fig. 2 shows the positive case on the AI2D dataset, where LLM gives relevant and effective responses at each stage, assists in diagram parsing to construct key visual and textual graphs, and obtains the correct answer. Fig. 3 shows the positive case on the SQA-I dataset. In the responses at different stages of LLM, there is some background knowledge (red part) that is contrary to the question, but the small amount of noise can be ignored compared to the acquisition of key features, and it still provides positive help for answer reasoning. Fig. 4 shows the positive case on the CSDQA dataset. As can be seen from this example, there are certain scenarios where relying solely on the object parsing and other capabilities of DPTs is enough to reason about the correct answer.

Fig. 5 shows the negative case on the SQA-I dataset. This common error is due to a large number of irrelevant or incorrect responses from LLMs. The root cause is that there are errors or inaccuracies in the automatically generated diagram captions that interfere with LLM's understanding. This phenomenon may be effectively improved in the future by replacing it with more advanced LLMs, especially LLMs that support multi-modal inputs. Fig. 6 shows the negative case on the TQA-DMC dataset. This failure is caused by LLMs overextended knowledge and redundant objects that rarely overlap with objects in the diagram parsing results. In future work, we consider solving this type of problem by better balancing the interaction between LLMs and DPTs.

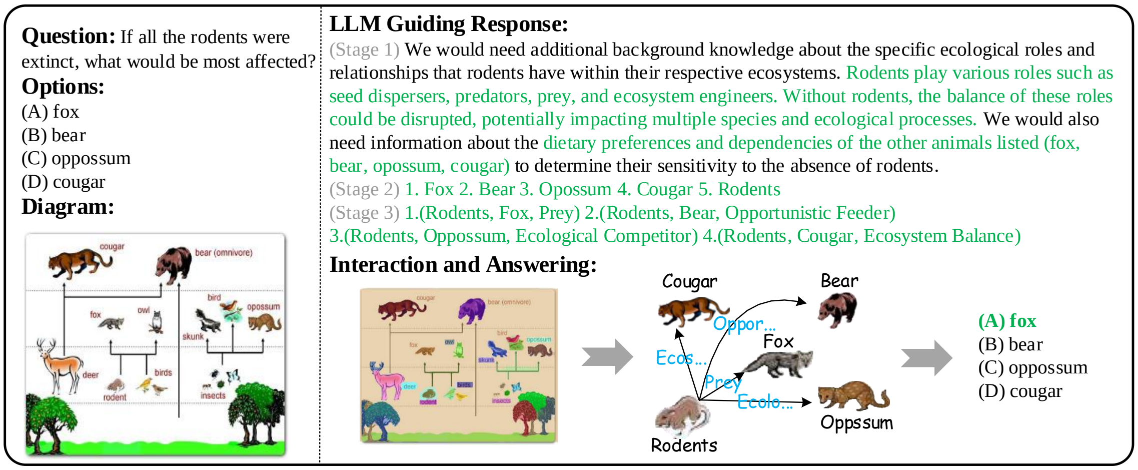


Figure 2. Positive case of CoG-DQA model on AI2D dataset.

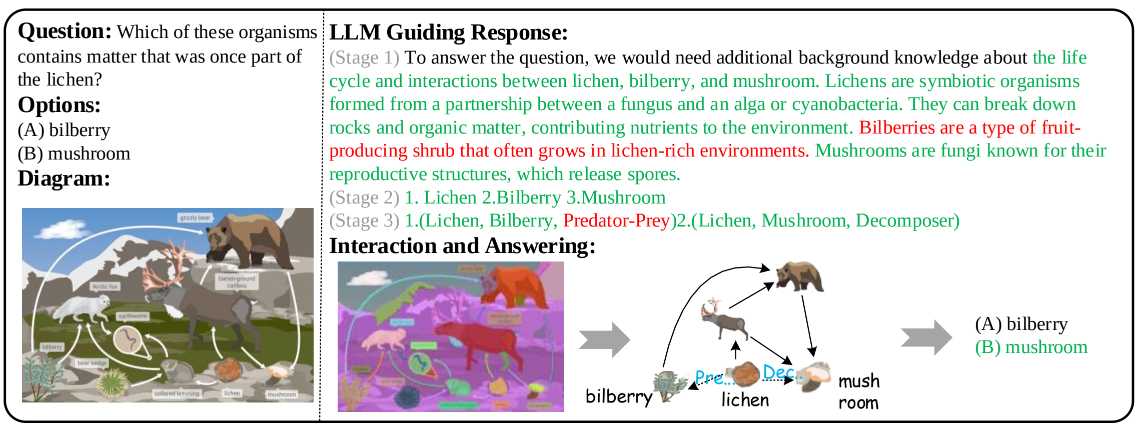


Figure 3. Positive case of CoG-DQA model on SQA-I dataset.

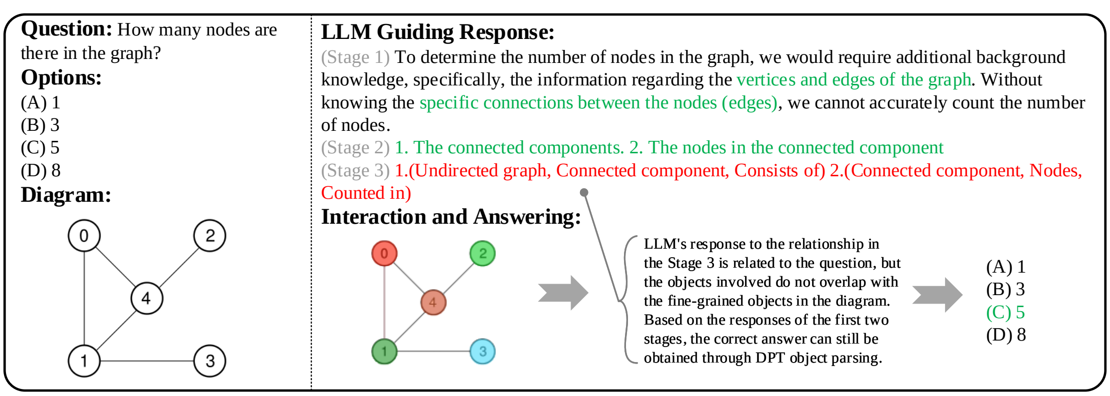


Figure 4. Positive case of CoG-DQA model on CSDQA dataset.

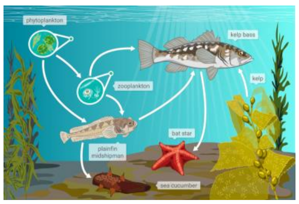
Question: Based on the arrows, which of the following living things is an omnivore?

Options:

(A) kelp bass

(B) zooplankton

Diagram:



LLM Guiding Response:

(Stage 1) We would need additional background knowledge about the specific diet preferences of the depicted penguin and cartoon character. Without this information, we cannot confidently identify the omnivore solely based on the arrows. However, assuming the penguin is representative of real-life penguins, they are primarily carnivorous, feeding mainly on fish and other marine animals. If the cartoon character is not explicitly shown consuming any specific type of food, we cannot make any assumptions about its dietary habits. Zooplankton, on the other hand, refers to microscopic animals in the water and is typically consumed by filter-feeding organisms such as whales and some fish species.

(Stage 2) 1. Kelp bass 2.Zooplankton

(Stage 3) 1.(Kelp Bass, Zooplankton, Predation) 2.(Zooplankton, Phytoplankton, Consumption)

Interaction and Answering:

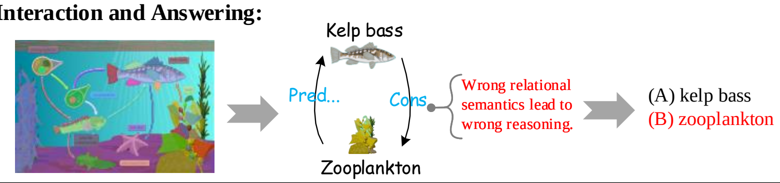


Figure 5. Negative case of CoG-DQA model on SQA-I dataset.

Question: How many animals do the eagle consume?

Options:

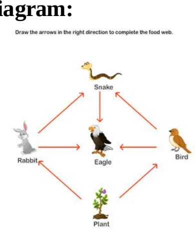
(A) 4

(B) 3

(C) 1

(D) 2

Diagram:



LLM Guiding Response:

(Stage 1) To solve this problem, additional background knowledge would be required, such as the diet and feeding behavior of eagles in the specific ecosystem depicted in the image. Additionally, understanding the trophic levels and interconnections among the different organisms in the food chain would be helpful. Without this information, we cannot accurately determine the number of animals the eagle consumes.

(Stage 2) 1. Frog 2.Cricket 3.Decomposers 4.Insectivorous birds 5.Hawks and owls

(Stage 3) 1.(Insectivorous birds, Hawks/Owls, Predation)

Interaction and Answering:

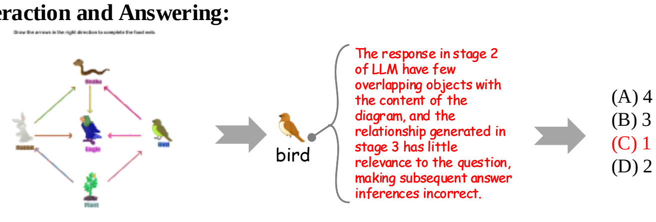


Figure 6. Negative case of CoG-DQA model on TQA-DMC dataset.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015. 1

[3] Hédi Ben-Younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. MUTAN: multimodal tucker fusion for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2631–2639. IEEE Computer Society, 2017. 1, 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229, 2020. 4

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 1, 3

[6] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648, 2019. 1, 3

[7] José Manuél Gómez-Pérez and Raúl Ortega. ISAAQ - mastering textbook questions with pre-trained transformers and bottom-up and top-down attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5469–5479. Association for Computational Linguistics, 2020. 1, 3

[8] Monica Haurilet, Alina Roitberg, and Rainer Stiefelhagen. It's not about the journey; it's about the destination: Following soft paths under question-guidance for visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1930–1939. Computer Vision Foundation / IEEE, 2019. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 4

[10] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 235–251. Springer, 2016. 1

[11] Daesik Kim, Youngjoon Yoo, Jeesoo Kim, Sangkuk Lee, and Nojun Kwak. Dynamic graph generation network: Generating relational knowledge from diagrams. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4167–4175, 2018. 1

[12] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1571–1581, 2018. 1, 3

[13] 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. 1, 3

[14] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, 2020. 1, 3

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3

[16] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 1, 3

[17] Jie Ma, Qi Chai, Jingyue Huang, Jun Liu, Yang You, and Qinghua Zheng. Weakly supervised learning for textbook question answering. *IEEE Trans. Image Process.*, 31:7378–7388, 2022. 3

[18] Jie Ma, Qi Chai, Jingyue Huang, Jun Liu, Yang You, and Qinghua Zheng. Weakly supervised learning for textbook question answering. *IEEE Transactions on Image Processing*, 31:7378–7388, 2022. 1

[19] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2023. 1, 3

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763, 2021. 4

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67, 2020. 1, 3

[22] Shaowei Wang, Lingling Zhang, Xuan Luo, Yi Yang, Xin Hu, Tao Qin, and Jun Liu. Computer science diagram understanding with topology parsing. *ACM Trans. Knowl. Discov. Data*, 16(6):114:1–114:20, 2022. 1, 3

[23] Fangzhi Xu, Qika Lin, Jun Liu, Lingling Zhang, Tianzhe Zhao, Qi Chai, Yudai Pan, Yi Huang, and Qianying Wang. Moca: Incorporating domain pretraining and cross attention for textbook question answering. *Pattern Recognition*, 140: 109588, 2023. 1

[24] Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *arXiv preprint arXiv:2305.16582*, 2023. 1, 3

[25] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1839–1848. IEEE Computer Society, 2017. 1, 3

[26] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6281–6290. Computer Vision Foundation / IEEE, 2019. 1, 3

[27] Zhaoquan Yuan, Xiao Peng, Xiao Wu, and Changsheng Xu. Hierarchical multi-task learning for diagram question answering with multi-modal transformer. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1313–1321. ACM, 2021. 1

[28] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1, 3

[29] Chen Zheng, Quan Guo, and Parisa Kordjamshidi. Cross-modality relevance for reasoning on language and vision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7642–7651. Association for Computational Linguistics, 2020. 1, 3