

Training-free VQA with Selective Meta-Reasoner for Interpretable Vision-Language Reasoning

Seungyeon Lee
statai3237@knu.ac.kr
Kyungpook National University
Daegu, South Korea

Wonjun Choi
sangju@knu.ac.kr
Kyungpook National University
Daegu, South Korea

Dong-Gyu Lee*
dglee@knu.ac.kr
Kyungpook National University
Daegu, South Korea

Abstract

Visual Question Answering (VQA) is a representative vision-language task that integrates images and questions to generate answers. With the increasing importance of interpretable AI, there has been growing attention in VQA that generates question relevant rationales and leverages them to produce more accurate answers. However, existing approaches remain underexplored in deriving textual rationales from visual rationales and selectively utilizing relevant rationales in reasoning. In this work, we propose a selective meta-reasoner based Training-free VQA named SMR-VQA to improve interpretable vision-language reasoning. We introduce a selective meta reasoner that leverages textual rationales derived from visual rationales. The selective meta reasoner ensures that only relevant rationales contribute to the reasoning process, addressing the problem of applying unnecessary rationales during answer derivation. The SMR-VQA requires no additional training and leverages off-the-shelf models to automatically generate reasoning paths for images and questions, which are then used to predict answers effectively. We conduct extensive experiments using two VQA datasets: VQA-X and A-OKVQA. Experiments demonstrate that the SMR-VQA model outperforms state-of-the-art models, achieving a 5.7% improvement in accuracy on VQA-X and an 5.4% improvement in VQA score on the A-OKVQA dataset.

Keywords

Visual Question Answering, Multi-modal Reasoning, Training-free, Selective Meta-Reasoner, Interpretability

1 Introduction

Visual Question Answering (VQA) is a multimodal task requiring models to answer natural language questions about a given image [9, 12]. The goal of VQA is to assess visual understanding and evaluate the integration of visual and linguistic reasoning [28]. Due to its comprehensive nature, VQA has been widely adopted as a benchmark for measuring progress in vision-language understanding. The field of interpretable artificial intelligence (AI) has gained significant attention, and this trend has extended to VQA [29]. Recent works have incorporated explainable AI (XAI) techniques and chain-of-thought (CoT) reasoning into VQA to enhance transparency and trustworthiness [14, 15]. This approach aims to make the underlying decision-making process more interpretable by providing a human-understandable rationale alongside the answers [3, 5].

Despite these advancements, current research on interpretable VQA still faces important limitations, which can be summarized

into two main issues. First, existing methods often generate textual rationales [8, 21] that rely on incomplete or even incorrect information generated from the image, the question, or the reasoning process. This can lead to explanations that appear plausible but fail to reflect relevant evidence. Second, prior works have explored different approaches, ranging from the use of a single reasoning path to the aggregation of multiple reasoning paths [15, 17]. However, existing approaches fall short in selectively reflecting the most relevant and reliable reasoning. The resulting explanations often contain redundant or misleading information, which compromises the reliability of the final answer.

To address these limitations, we propose a novel **Selective Meta-Reasoner based Training-free VQA (SMR-VQA)** method that introduces a reliable reasoning process for generating answers without requiring additional training. Our method incorporates visual rationales that highlight regions in the image relevant to the given question. The integration of such visual rationales allows the textual reasoning process to be guided toward producing information more closely aligned with the question. Furthermore, our method selectively identifies and uses the reliable reasoning paths instead of relying on all reasoning paths equally. This selective meta-reasoning generation enables the model to construct faithful and trustworthy rationales, which in turn support the derivation of more accurate final answers. We conduct experiments on two VQA benchmarks, VQA-X and A-OKVQA, to evaluate answer generation. The experimental results show that the proposed SMR-VQA improves accuracy on VQA-X from 80.5 to 85.1 and increases the VQA score on A-OKVQA from 53.4 to 56.3, demonstrating the effectiveness of the proposed approach. These findings highlight that our method can perform selective reasoning over visual and textual rationales to generate more accurate answers.

2 Methodology

In this section, we present a novel selective meta-reasoner based Training-free VQA framework named SMR-VQA for interpretable vision-language reasoning. As shown in Fig. 1, SMR-VQA identifies visual rationales in an image relevant to a question and generates corresponding textual and global rationales. It then automatically selects the most relevant rationales based on the threshold criteria and integrates them to generate the final answer.

2.1 Visual and Textual Rationale Generation

In this phase, we first generate visual rationales by applying Grounding-DINO [19] to localize object-level bounding boxes corresponding to the question in the image. The visual rationales $v_i \in \{v_1, v_2, \dots, v_5\}$ use the top five bounding boxes ranked by confidence score for question-based object localization, instead of a single bounding box

*Corresponding Author

© 2025 Copyright held by the owner/author(s).

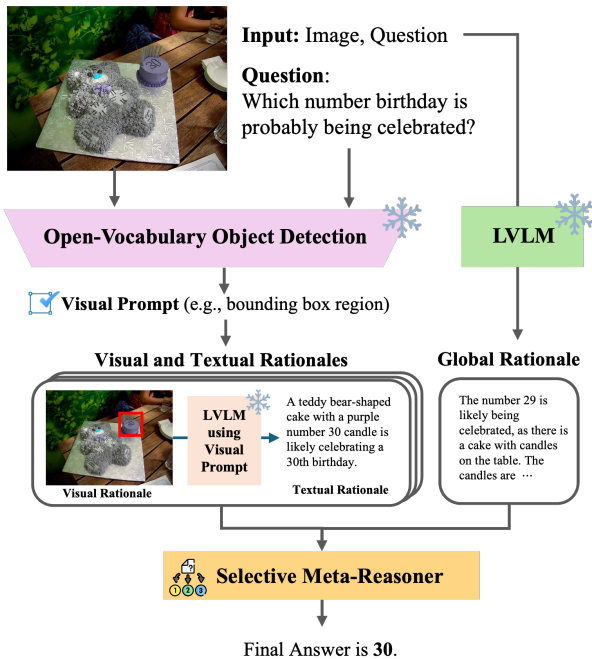


Figure 1: The proposed SMR-VQA model architecture.

to mitigate inaccurate outputs. Bounding boxes from each visual rationale serve as visual prompts over the original image, enabling the ViP-LLaVA [1] model to generate region-focused textual rationales $t_i \in \{t_1, t_2, \dots, t_5\}$ from the image and the visual prompts (e.g., red bounding box). We adopt ViP-LLaVA for this capability of generating fine-grained, region-aware textual descriptions. Specifically, we instruct the ViP-LLaVA model with a text prompt “*Question: {question}. Based on the question, describe only the relevant content in the red box. Do not mention bounding boxes, highlights, or rectangles.*”. Each visual rationale produces a corresponding textual rationale, resulting in a set of textual rationales that align with the input visual rationales. We also generate a global textual rationale t_g for the entire image to capture information that can be missed when focusing solely on specific regions.

2.2 Selective Meta-Reasoner

In this reasoning step, we introduce a selective meta-reasoner, which automatically selects the rationales needed in the reasoning process among the textual rationale candidates and constructs a reasoning path. The selective meta-reasoner uses a VQAScore [18] to compute an alignment score $s_i = \{s_1, s_2, \dots, s_g\}$, which measures the probability of each candidate rationale representing the image, and selects the relevant rationales.

$$s_i = P(\text{“yes”} | I, Q(t_i)), \quad i \in \{1, 2, 3, 4, 5, g\}, \quad (1)$$

where question prompt $Q(t_i)$ is “Dose this figure show t_i ?”. We calculate an image-text alignment score using the given image and textual rationale candidates t_i and global rationale t_g . The selective criteria are set heuristically through experiments. We use rationales with alignment scores greater than 0.6 to guide reasoning.

The selective usage of image-related rationales aims to prevent the reasoning process from using irrelevant information.

2.3 Answer Generation

In answer generation, textual rationales related to the image obtained through the selective meta-reasoner are used to infer the final answer. Our proposed SMR-VQA constructs reasoning paths automatically through a selective meta-reasoner. We use a frozen Llama3-Instruct (8B) model [7] as the large language model (LLM) for reasoning to generate answers based on the selective rationales. We utilize the Llama3-Instruct model due to its strong instruction-following ability, which leads to improved natural language understanding and generation. We apply a text prompt to generate the final answer as shown in Table 1.

Table 1: Prompt input example of answer generation.

Prompt	<p>Instruction: Answer the following question using the provided text. Answer Format: - Respond with a single short answer without using ‘a’ or ‘an’. Answer in a single short word. Do not provide any explanation or reasoning. - Pay attention to spacing. Provided text: {rationale} Question: {question} Answer:</p>
---------------	--

3 Experiments

3.1 Experimental Setting

3.1.1 Datasets. We evaluate the performance of our proposed SMR-VQA against other baseline models on two VQA benchmarks: VQA-X [20] and A-OKVQA [23]. VQA-X provides human-annotated textual explanations along with question-answer pairs, enabling evaluation of both answer prediction and rationale generation. A-OKVQA is a crowdsourced dataset that requires commonsense and world knowledge for reasoning and answering questions. In this work, we focus on VQA answer prediction by leveraging selective rationales into the reasoning process using the proposed method.

3.1.2 Baselines. For the VQA-X dataset, we compare the proposed SMR-VQA with six baselines, including VQA-E [16], PJ-X [20], e-UG [11], DMRFNet [31], VCIN [27], and MRVQA-C [14]. The evaluation of predicted answers uses accuracy as a metric. We use seven baselines on A-OKVQA, including PICa [30], Img2LLM [10], LAMOC [6], L2A [26], VCTP [2], DIETCOKE [15], and Brote-IM-XL [25], to evaluate VQA performance. We compare our method with several baselines using VQA score: $\min(\frac{\#human \text{ that provided that answer}}{3}, 1)$.

3.2 Experimental Results

3.2.1 Main Results. Table 2 presents the results of our method compared to various baselines on the VQA-X dataset. The previous state-of-the-art [13] achieved 80.5 accuracy, while our approach improves the performance to 85.1, demonstrating a significant gain over existing methods. Table 3 shows the results on A-OKVQA dataset,

Table 2: Comparison results on VQA-X with state-of-the-art models. Value in bold indicates the best performance, and the second best result is underlined.

Baselines	Any Training ?	Overall Accuracy (↑)
VQA-E [16]	✓	70.2
PJ-X [20]	✓	76.4
e-UG [11]	✓	<u>80.5</u>
DMRFNet [31]	✓	72.6
VCIN [27]	✓	77.7
MRVQA-C [14]	✓	78.8
SMR-VQA (ours)	✗	85.1

Table 3: Comparison results on A-OKVQA with state-of-the-art models. Value in bold indicates the best performance, and the second best result is underlined.

Baselines	Any Training ?	VQA Score (↑)
PICa [30]	✗	42.4
Img2LLM [10]	✗	42.9
LAMOC [6]	✓	37.9
L2A [26]	✓	48.5
VCTP [2]	✗	53.2
DIETCOKE [15]	✗	47.5
Brote-IM-XL [25]	✓	<u>53.4</u>
SMR-VQA (ours)	✗	56.3

evaluated using the VQA score. The previous baseline achieved a score of 53.4, whereas our approach improves the performance to 56.3, representing a relative increase of approximately 5.4%. These results demonstrate a significant improvement over existing methods in zero-shot VQA evaluation. The significant performance improvements observed across both benchmarks demonstrate that the proposed SMR-VQA method effectively performs reasoning in VQA to generate more accurate answers.

Table 4: Performance comparison of selective criteria using alignment scores: Top-K and Threshold value selection.

Dataset	Selective Criteria	Value	Score (↑)
VQA-X	Top-K	K=1	83.0
		K=2	84.4
		K=3	84.7
	Threshold	0.6	85.1
		0.7	84.6
		0.8	84.6
A-OKVQA	Top-K	K=1	54.1
		K=2	55.8
		K=3	57.4
	Threshold	0.6	56.3
		0.7	56.3
		0.8	56.1

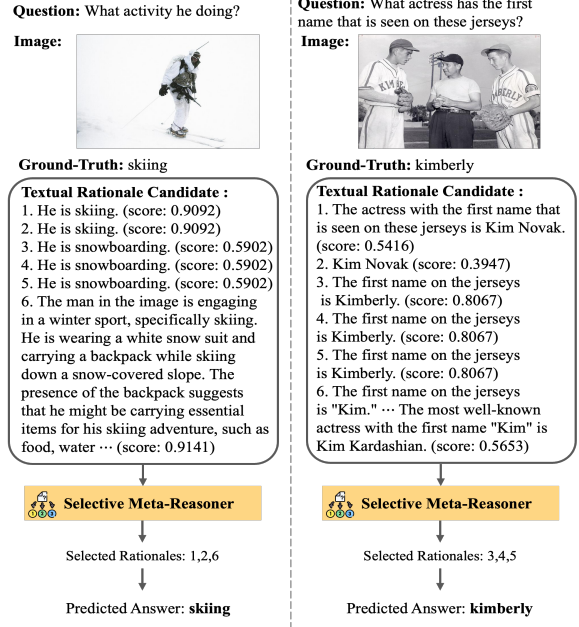


Figure 2: Effect of using selective meta-reasoner to construct reasoning paths for answer generation.

3.2.2 Ablation study. In Table 4, we conduct the ablation study to evaluate the criteria for selecting rationales. We compared the top-k selection method with a threshold-based approach, which does not predefine the number of rationales. The results demonstrate that using the threshold of 0.6 during the reasoning process yields the best performance. These results demonstrate that textual rationales, derived from image regions relevant to the question, contribute to explaining the content of the image. By selectively incorporating the necessary information during reasoning, the quality of inference improves, leading to more accurate answer generation.

Table 5: Comparison of performance across two visual prompt based models in generating visual rationales.

Dataset	Visual Rationale	Score (↑)
VQA-X	ViP-LLaVA [1]	85.1
	Qwen-2.5-VL [22]	83.8
A-OKVQA	ViP-LLaVA [1]	56.3
	Qwen-2.5-VL [22]	56.1

Table 5 shows the performance comparison of models used for generating textual rationales with visual prompts. For visual prompt based text generation model, ViP-LLaVA model shows relative improvements over Qwen-2.5-VL [22] of approximately 1.6% on the VQA-X, and also exhibits performance gains on the A-OKVQA. These results highlight the difference in generating information from image regions relevant to the question by leveraging visual prompts. Table 6 demonstrates that the Llama3-Instruct model yields superior reasoning ability in answer generation compared to using either the Llama2-Chat [24] or Vicuna-1.5 [4] models.

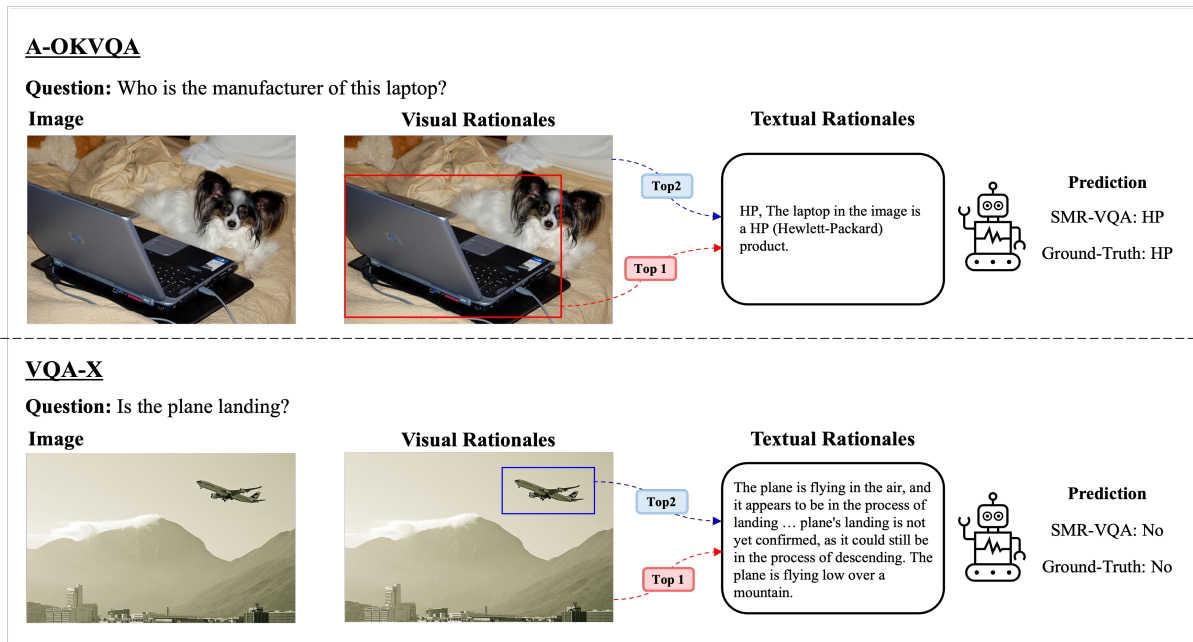


Figure 3: Examples of the VQA-X and A-OKVQA using SMR-VQA, which generates answers based on textual rationales derived from visual rationales.

Table 6: Performance comparison of LLM used in answer generation.

Dataset	LLM	Score (↑)
VQA-X	Llama3-Instruct [7]	85.1
	Llama2-Chat [24]	74.8
	Vicuna-1.5 [4]	81.7
A-OKVQA	Llama3-Instruct [7]	56.3
	Llama2-Chat [24]	54.0
	Vicuna-1.5 [4]	43.5

3.2.3 *Qualitative Analysis.* Fig. 2 shows a qualitative evaluation of the proposed selective meta-reasoner. As illustrated in the examples, relying on incorrect evidence often introduces irrelevant information, leading to performance degradation. In contrast, our method uses a selective meta-reasoner to effectively select the rationales directly relevant to the given question during the reasoning. These findings demonstrate that selective meta-reasoner improves the precision of rationale selection and enhances the overall reliability of reasoning. Fig. 3 shows examples of VQA-X and A-OKVQA using SMR-VQA. The examples demonstrate that generating question-related textual rationales through visual rationales enhances reasoning.

4 Discussion

In this work, we proposed SMR-VQA, which automatically selects image relevant rationales for generating an answer. Across two benchmarks, SMR-VQA consistently outperformed previous state-of-the-art, incorporating visual and textual rationales. Nevertheless,

the current rationale generation strategy can exhibit biases across certain object categories or question types, potentially affecting fairness in human-centered AI applications. We plan to evaluate selective rationales with broader benchmarks and employ uncertainty-aware rationale selection. In addition, future studies will extend answer generation beyond single-word outputs by extending diverse or free-form prompting and conducting human-in-the-loop evaluations to assess clarity, completeness, and user trust.

5 Conclusion

In this study, we propose a novel SMR-VQA method and introduce a selective meta-reasoner that automatically generates the reasoning process required to answer image-based questions. We leverage visual rationales as visual prompts to generate textual rationales, enabling the generation of evidence from image regions relevant to the question. The proposed method evaluates VQA answer prediction performance on two benchmarks. Experimental results demonstrate that our approach selectively and generates the reasoning necessary for answering, achieving superior performance compared to existing methods.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2025-02214941, 50%) and the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2020-II201808, 50%).

References

- [1] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12914–12923.
- [2] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 1254–1262.
- [3] Yu Cheng, Arushi Goel, and Hakan Bilen. 2025. Visually Interpretable Subtask Reasoning for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2760–2780.
- [4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. <https://lmsys.org/blog/2023-03-30-vicuna/>
- [5] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104* (2024).
- [6] Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Zero-shot Visual Question Answering with Language Model Feedback. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9268–9281. doi:10.18653/v1/2023.findings-acl.590
- [7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv-2407.
- [8] Deepanway Ghosal, Navonil Majumder, Roy Ka-Wei Lee, Rada Mihalcea, and Soujanya Poria. 2023. Language Guided Visual Question Answering: Elevate Your Multimodal Language Model Using Knowledge-Enriched Prompts. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. <https://openreview.net/forum?id=FLSQjYmzlp>
- [9] Jingliang Gu and Zhixin Li. 2024. Beyond Language Bias: Overcoming Multimodal Shortcut and Distribution Biases for Robust Visual Question Answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) (CIKM '24). Association for Computing Machinery, New York, NY, USA, 3767–3771. doi:10.1145/3627673.3679880
- [10] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10867–10877. doi:10.1109/CVPR52729.2023.01046
- [11] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. 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*. 1244–1254.
- [12] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *Comput. Surveys* 57, 8 (2025), 1–36.
- [13] Chengen Lai, Shengli Song, Shiqi Meng, Jingyang Li, Sitong Yan, and Guangneng Hu. 2024. Towards more faithful natural language explanation using multi-level contrastive learning in vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2849–2857.
- [14] Kun Li, George Vosselman, and Michael Ying Yang. 2025. Multimodal Rationales for Explainable Visual Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 191–201.
- [15] Miaoyu Li, Haoxin Li, Zilin Du, and Boyang Li. 2024. Diversify, Rationalize, and Combine: Ensembling Multiple QA Strategies for Zero-shot Knowledge-based VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1552–1566. doi:10.18653/v1/2024.findings-emnlp.84
- [16] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 552–567.
- [17] Tao Li, Linjun Shou, and Xuejun Liu. 2025. Mixture of Rationale: Multi-modal Reasoning Mixture for Visual Question Answering. In *Neural Information Processing*, Mufti Mahmud, Maryam Dobarjeh, Kevin Wong, Andrew Chi Sing Leung, Zohreh Dobarjeh, and M. Tanveer (Eds.). Springer Nature Singapore, Singapore, 105–122.
- [18] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Günter Varol (Eds.). Springer Nature Switzerland, Cham, 366–384.
- [19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*. Springer, 38–55.
- [20] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8779–8788.
- [21] Chen Qiu, Zhiqiang Xie, Maofu Liu, and Huijun Hu. 2024. Explainable Knowledge reasoning via thought chains for knowledge-based visual question answering. *Information Processing Management* 61, 4 (2024), 103726. doi:10.1016/j.ipm.2024.103726
- [22] Qwen, ., An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] <https://arxiv.org/abs/2412.15115>
- [23] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*. Springer, 146–162.
- [24] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [25] Ziyue Wang, Chi Chen, Yiqi Zhu, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. 2024. Browse and Concentrate: Comprehending Multimodal Content via Prior-LLM Context Fusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11229–11245. doi:10.18653/v1/2024.acl-long.605
- [26] Xiaoying Xing, Peixi Xiong, Lei Fan, Yunxuan Li, and Ying Wu. 2024. Learning to Ask Denotative and Connotative Questions for Knowledge-based VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8301–8315. doi:10.18653/v1/2024.findings-emnlp.487
- [27] Dizhan Xue, Shengsheng Qian, and Changsheng Xu. 2023. Variational causal inference network for explanatory visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2515–2525.
- [28] Shuo Yang, Caren Han, Siwen Luo, and Eduard Hovy. 2025. MAGIC-VQA: Multimodal And Grounded Inference with Commonsense Knowledge for Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 16967–16986. doi:10.18653/v1/2025.findings-acl.872
- [29] Xingyi Yang and Xinchao Wang. 2024. Language model as visual explainer. *Advances in Neural Information Processing Systems* 37 (2024), 135094–135128.
- [30] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. In *AAAI*.
- [31] Weifeng Zhang, Jing Yu, Wenhong Zhao, and Chuan Ran. 2021. DMRFNet: deep multimodal reasoning and fusion for visual question answering and explanation generation. *Information Fusion* 72 (2021), 70–79.