Probing Logical Reasoning of MLLMs in Scientific Diagrams

Yufei Wang and **Adriana Kovashka**
Department of Computer Science, University of Pittsburgh
yuw384@pitt.edu and kovashka@cs.pitt.edu

Abstract

We examine how multimodal large language models (MLLMs) perform logical inference grounded in visual information. We first construct a dataset of food web/chain images, along with questions that follow seven structured templates with progressively more complex reasoning involved. We show that complex reasoning about entities in the images remains challenging (even with elaborate prompts) and that visual information is underutilized.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015; Goyal et al., 2019) is a widely used probing methodology for assessing Multimodal Large Language Models (MLLMs). Benchmarks are created to uncover weaknesses and improve context comprehension (Kembhavi et al., 2017), spatial reasoning (Johnson et al., 2017; Ranasinghe et al., 2024), and temporal reasoning (Zhang et al., 2024). These reasoning tasks focus on natural photographs or abstract synthetic scenes but leave a critical gap in evaluating MLLMs' ability to reason about scientific diagrams, charts, and figures—formats that inherently encode structured factual knowledge. While some studies have highlighted these knowledge-rich formatted data (Lu et al., 2022; Kembhavi et al., 2016; Kahou et al., 2018; Masry et al., 2022; Hou et al., 2025), they often lack a rigorous assessment of *logical* reasoning within scientific contexts. This omission is significant, as logical reasoning is key to real-world applications like science education, environmental monitoring, and medical diagnostics, where precise, context-sensitive answers are essential. A prime example is MLLMs in pedagogical tools for grading or tutoring, which must assess student answers in context and identify when a response misses requirements or problem logic. To bridge this gap, we construct VQA tasks to probe the logical reasoning capabilities of MLLMs using scientific graphs, pushing beyond surface-level pattern recognition to a deeper understanding of structured information represented in a visual manner.

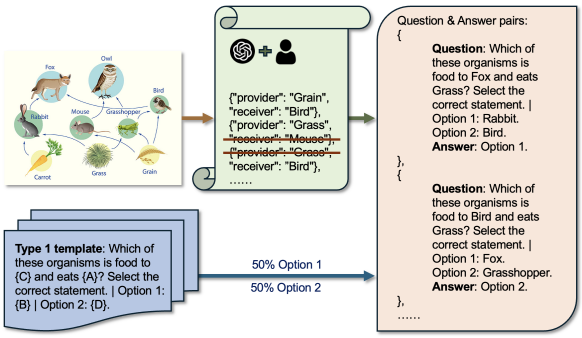


Figure 1: Grounded logical questions generation pipeline. Letters $\{A\}, \{B\}, \{C\}, \{D\}$ serve as placeholders for organism names, filled based on relations extracted from the image.

We consider logical reasoning as understanding the logical relations entailed in the question text and identifying the logical relations exhibited in the image. Previous work approached logical reasoning only from the textual aspect by connecting two questions using logical connectives like "and", "or", and "not" and inferring the final answer following the rules defined by the connective (Gokhale et al., 2020). However, the generated questions ("Is not the sky blue or is there snow on the mountains?") often follow an unnatural style rarely found in human communication. We instead ground the logical questions in science graphs which naturally contain logical relations, and maintain the natural tone of questions and their practicality. We extract the prey-predator relations contained in food web images and make use of logic entailment to create the question-answer pairs.

We create two datasets, one using real and one using synthetic images and relations, to evaluate logical reasoning ability over food webs. We evaluate four recent MLLMs and find they struggle

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10706–10718
November 4-9, 2025

with complex or chained relationships, and when logical reasoning is required that goes beyond regurgitation of pretraining knowledge. We also find suboptimal use of the visual content and of clarification on how to perform the logic tasks.

2 Datasets Creation

We selected food web images as our test bed since they encode directed relations, reflect scientific knowledge, and thus naturally render themselves to logical reasoning. Existing datasets only contain a handful of food web images without annotations of specific relations between organisms (Lu et al., 2022) or the annotations are limited by the parsing strategies and thus not comprehensive enough to enable creating our questions (Kembhavi et al., 2016). We propose two datasets to address this gap.

Real images. We start with the food web images in the aforementioned datasets and complement them with downloaded images via Google web search with resolution from 316x159 to 4000x4000, then manually filtering out the ones that do not portray the relations using arrows.[1] We design the pipeline in Fig. 1 to extract the relations from the image and then apply predefined logically challenging templates for the creation of the questions and answers. We query GPT-4o to generate the captions describing the relation in the form of provider and receiver pairs and later manually filter out the hallucinated relations. We then apply the logical questions templates, resulting in 9,327 logical questions for 146 images. We also investigated directly prompting GPT-4o to generate logical questions and answers, but the results always included hallucinated information not in the image. Our pipeline, on the contrary, ensures the validity and logical soundness of the generated question-answer pairs.

Synthetic images. To reduce the chance of MLLMs leveraging memorized biological knowledge, we generate synthetic food web images. We employ Stable Diffusion 3 (Esser et al., 2024) to produce 10 white-background images for each of the 92 animal categories in the LVIS dataset (Gupta et al., 2019). For each web, we randomly sample $N \in \{5, 6, 7, 8\}$ animals and select one image per animal. We then create $N-2$ (up to 5) chain relations (e.g., A eats B, B eats C) and $N-3$ (up to 4) multiple-prey relations (e.g., A eats B and C), followed by randomly adding remaining edges with probability 0.2 to simulate sparsity. Graphviz (Gansner and North, 2000) is used to visualize the food web. We generate 400 synthetic images of reslution 1025*1025 and 24,206 associated questions. The pipeline is fully automated and can be easily scaled. Due to the randomized structure, the resulting food webs are unlikely to align with real-world biology, reducing shortcut opportunities for pretrained models. We assess 20 examples for real and synthetic images across each type: 50% of the questions are inconsistent with real-world knowledge, compared to only 4% for real images. We will release the data.[2]

Type	Relation	Template
1	$A \rightarrow B$	Does A eat B?
2	$A \nrightarrow B$	Does B eat A?
3	$A \nrightarrow B$	Does B contain matter that was once part of A?
4	$A \rightarrow B \rightarrow C$	Does C contain matter that was once part of A?
5	$A \rightarrow B \rightarrow C$	Which of these organisms is food to C and eats A? Select the correct statement. | Option 1: B Option 2: D.
6	$A \rightarrow B \leftarrow C$	If A does not exist, will B still survive?
7	$A \rightarrow B, C \nrightarrow B, E \nrightarrow D, G \nrightarrow F$	Select the correct statement. | Option 1: B eats A or B eats C. Option 2: D eats E or F eats G.

Table 1: Each capital letter denotes an organism. An arrow $A \rightarrow B$ (B eats A) means A is the provider and B the consumer. A negated arrow $\nrightarrow$ signifies no direct relationship exists between A and B in the diagram. In Type 5, D in Option 2 must be present in the image but not simultaneously related to both A and C.

Questions. We compose the relations from the real/synthetic images into binary questions using the logical relations "and", "or" and "not" while preserving the natural tone and practicality. For example, we ask "If the food source does not exist, will this animal still survive?" which could be practical to the development of an AI assistant supporting students learning about ecosystems. Tab. 1 presents all types of logical questions we generated.

We start with questions that only concern single arrows. Type 1 questions are generated by flipping arrow directions in the image, asking whether the

[1] The process is straightforward: look at the GPT4o generated relations and the image together, check if there is an arrow pointing from the prey to the predator for each relation, and then remove the relation entries that are invalid, i.e. not present on the image. We removed roughly 40% for images that have more than 5 relations. Hallucinations increase as the food web/chain images become more complex.

[2] https://github.com/elffiona/Probing_Logical_Reasoning

Model	Image	Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
LLaVA	Real	63.08%	77.13%	73.15%	55.56%	66.24%	76.36%	45.26%	52.64%
	White (Real QA)	61.03%	89.15%	71.16%	54.23%	62.38%	71.49%	43.76%	51.25%
	Synthetic	51.60%	88.06%	50.09%	47.57%	47.80%	52.09%	49.83%	50.81%
InternVL	Real	74.23%	95.74%	78.12%	65.89%	76.29%	86.10%	65.63%	71.29%
	White (Real QA)	63.67%	94.19%	64.53%	53.14%	60.16%	82.27%	54.83%	65.26%
	Synthetic	53.91%	78.21%	51.07%	52.16%	51.07%	59.12%	51.78%	51.55%
Qwen-VL	Real	77.62%	98.26%	84.28%	66.21%	88.55%	91.31%	61.03%	75.03%
	White (Real QA)	62.25%	96.90%	70.99%	54.92%	71.73%	85.40%	62.71%	70.04%
	Synthetic	59.31%	77.30%	53.95%	63.38%	62.65%	63.47%	52.13%	53.20%
GPT	Real	70.12%	96.12%	79.61%	55.92%	55.96%	83.66%	67.85%	69.00%
	White (Real QA)	69.21%	91.09%	79.18%	55.92%	56.89%	85.17%	64.48%	66.78%
	Synthetic	51.69%	66.52%	50.19%	55.60%	50.52%	51.45%	50.50%	46.38%

Table 2: Zero-shot accuracy for LLaVA-NeXT (Liu et al., 2024), InternVL2.5 (Chen et al., 2024), Qwen2.5-VL (Bai et al., 2025) and GPT-4o-mini (OpenAI, 2024) on question-answer pairs from real food web images, completely white images using the same questions from real images, and synthetic images with their corresponding questions.

prey eats the *predator*. We flip 50% of arrows randomly to balance positive and negative pairs. Type 2 questions follow the same textual format as Type 1, but the negative question-answer pairs ask about two organisms that are not linked by arrows in the diagram. Each Type 2 is also rephrased in a less direct format to create Type 3 questions.

We increase reasoning difficulty by introducing three-organism chain relations, forming Type 4 questions to examine logical reasoning ability over the transitivity of the directed prey-predator relations. As matter moves from prey to predators, we ask "Does [final predator] contain matter that was once part of [first prey]?" The corresponding negative counterpart asks "Does [first prey] contain matter that was once part of [final predator]?"

A chained relation composes two separate relations with the relation "and." Type 5 questions challenge the model to identify the organism in the middle of the chain. To keep the question binary, we ask the model to select between two options. The correct option is linked to both mentioned organisms; the wrong option is present in the image but lacks the required relations. We balance answers for Type 5 by alternatively putting the correct option in option 1 or 2 in the template.

We then created Type 6 questions which require identifying that there are other potential food resources in the image to prevent the extinction of the target organism. This is an "or" relation involving parallel arrows that point at the same organism (predator B requires prey A or prey C). The negative cases for Type 6 involve predators that only have one prey shown in the image.

Type 7 questions also implement the logical relation "or". They ask for a selection between two statements, each of which consists of two relations connected through "or." The correct option statement has one correct relation and one incorrect relation (linked by "or", thus correct overall). The wrong option has two incorrect relations. We use the same balance strategy as for Type 5.

3 Results and Analysis

We evaluated four strong MLLMs (Liu et al., 2024; OpenAI, 2024; Bai et al., 2025; Chen et al., 2024) on our datasets using the same prompt and instructed them to provide one-word yes/no answers, or the option number for selection questions. Accuracy is calculated via string matching. Tab. 2 reveals the zero-shot accuracy of all models across real, white, and synthetic images. Tab. 3 reports accuracy for Qwen2.5-VL with added reasoning strategies. Tab. 4 presents the error analysis in terms of false positive and false negative rates across question types. As a preview, we highlight several patterns across models and datasets: (1) Models demonstrate good performance on simpler question types (Types 1/2/5) of real images that can be answered by leveraging prior knowledge shortcuts, as verified by the minimal drop in accuracy when real images are replaced with white images. However, they consistently struggle with more complex question types. (2) Accuracy drops consistently across all models when evaluated on synthetic food webs, suggesting that when we remove biological priors using synthetic images, models struggle with composed logical reasoning VQA task. (3) For Type 1 questions, all models show a significant drop in performance on synthetic images compared to real images except LLaVA, which improves in this setting, suggesting it is in-

fluenced by shortcut patterns from pretraining. (4) Including chain-of-thought (CoT) inspired strategies in the prompt does not help with enhancing the logical reasoning ability.

We highlight the broader contrast between simpler and more complex question types. Across both real and synthetic images, models perform substantially better on Type 1 questions than on other types, with margins of roughly 15–30% (Tab. 2). Since all questions share the same instruction-following requirement, this factor is controlled for, allowing us to attribute the observed performance gap primarily to logical complexity. This confirms that regardless of biological priors, models continue to struggle with logical reasoning.

3.1 Performance on real images

In Tab. 2 (Real rows) we see that Qwen2.5-VL achieves the highest accuracy on real images (77.62%), followed by InternVL2.5 (74.23%), GPT-4o-mini (70.12%) and LLaVA (63.08%). All models demonstrated good performance for question Type 1, 2, and 5, which directly ask for the prey-predator relationship between two organisms. However, models do not perform well on Type 3 and 4 questions. These ask for the relationship from the perspective of movement of matter; Type 3 focuses on length-2 relations and Type 4 on length-3 relations. Even though the visual information required to answer Type 2 and 3 questions correctly is the same, the indirect format of Type 3 hinders MLLMs. Models also encounter difficulties in Type 6 and 7 questions, which require two-step reasoning to understand the logic in the text and looking for corresponding information in the image. To illustrate, the logic entailed in the question "If rabbit does not exist, will lion still survive?" is an "or" relation that the model needs to identify, regarding whether lion has the same prey-predator relation with another organism in the image as the one it has with rabbit. For Type 7, "or" logic is more obvious by construction, but the models need to first identify the four relations in the options and reason based on image evidence.

3.2 Importance of the image

To better understand model behavior, we design an ablation by replacing food web images with white images of the same size, while keeping all questions and answers unchanged. Thus no visual information is provided and we evaluate the model's performance relying solely on the language model component. Tab. 2 (White rows) shows the performance of models with white images does not drop significantly (in some rare cases, it even improves). Thus these MLLMs are largely leveraging the prior in language model to answer instead of getting information from the image. This observation leads to future work in understanding and improving the visual representations in MLLMs while reducing bias introduced in the language model.

3.3 Performance on synthetic images, and impact of pretraining bias

Tab. 2 (Synthetic rows) shows that all four models experience consistent trends. Qwen2.5-VL achieves the highest accuracy on synthetic images (59.31%), followed by InternVL2.5 (53.91%), LLaVA (51.60%) and GPT-4o-mini (51.54%). We observe a substantial drop in accuracy with synthetic images, with performance approaching random guessing (questions are binary). We find a significant drop in Type 1 questions for all models except LLaVA, which performed poorly on real images but improves in the synthetic setting. We attribute this to the fact that the other models likely rely on shortcut patterns from pretraining, as the wording in Type 1 questions can be commonly seen during training. In contrast, LLaVA appears less dependent on such shortcuts and benefits slightly when this bias is removed. We also observe big drops on Type 2 across all models; Type 2 shares the same wording as Type 1 but introduces harder negatives—relations that are not displayed in the image. For Type 3, which utilize the same relations as Type 2 but with a rarer word format, the drop is smaller, likely because the unusual phrasing makes it less common in pretraining data, reducing the impact of shortcut exploitation. In Type 4, the performance drops again. Although Type 3 and 4 share the same phrasing, Type 4 involves indirect relations where the predator and prey are separated by multiple intermediate nodes as they are from more distinct food web levels, leading to more plausible chains in biology. Thus in real images, models that rely on pretraining bias perform better on Type 4 than Type 3. This trend disappears in the synthetic setting, where such chains are randomized. Overall, our findings show that when we remove biological priors using synthetic images, models struggle with composed logical reasoning VQA task. Note that we verify the drop is not due to recognition failures by querying Qwen2.5-VL with balanced recognition questions ("Is *animal* in

Model	Image	Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
Qwen-VL	Real	77.62%	98.26%	84.28%	66.21%	88.55%	91.31%	61.03%	75.03%
	Real + S	76.21%	98.45%	74.81%	79.60%	83.18%	81.00%	54.83%	75.10%
	Synthetic	59.31%	77.30%	53.95%	63.38%	62.65%	63.47%	52.13%	53.20%
	Synthetic + S	56.88%	79.49%	52.43%	61.72%	51.30%	57.19%	56.39%	54.55%

Table 3: Zero-shot accuracy for Qwen2.5-VL(Bai et al., 2025) with pure QA-style prompts and prompts that include strategies for each question. + S denotes the addition of strategies in the prompt.

		Synthetic					Real				
Model	Metric	Type 1	Type 2	Type 3	Type 4	Type 6	Type 1	Type 2	Type 3	Type 4	Type 6
LLaVA	False Pos.	4.47%	3.15%	1.79%	2.07%	47.38%	22.87%	19.97%	38.29%	27.57%	54.38%
	False Neg.	7.47%	46.76%	50.64%	50.13%	2.79%	0.00%	6.88%	6.15%	6.19%	0.35%
InternVL	False Pos.	14.95%	6.28%	16.23%	16.01%	34.54%	4.07%	7.56%	21.13%	10.40%	18.95%
	False Neg.	6.84%	42.65%	31.61%	32.92%	13.75%	0.19%	14.32%	12.98%	13.32%	15.41%
Qwen-VL	False Pos.	15.95%	4.90%	29.39%	25.10%	1.67%	1.74%	5.73%	32.56%	9.35%	0.89%
	False Neg.	6.75%	41.14%	7.23%	12.25%	46.20%	0.00%	9.98%	1.23%	2.10%	38.09%
GPT	False Pos.	8.04%	5.34%	32.62%	33.92%	40.47%	3.88%	10.41%	43.81%	43.46%	13.37%
	False Neg.	25.45%	44.47%	11.77%	15.56%	9.03%	0.00%	9.98%	0.27%	0.58%	18.78%

Table 4: False positive and false negative rates for LLaVA-NeXT (Liu et al., 2024), InternVL2.5 (Chen et al., 2024), Qwen2.5-VL (Bai et al., 2025) and GPT-4o-mini (OpenAI, 2024) across Yes/No questions (Types 1, 2, 3, 4, and 6).

the image?"), achieving above 98% accuracy for both real and synthetic images.

3.4 Strategies for improvement

As a preliminary effort, we experimented with including strategies, presented in Tab. 7 (appendix) that explains in the prompt the steps to solve for the logical difficulties specific to each question type. For example, for the logically complex Type 6 questions phrased as "If A does not exist, will B still survive?", we instruct the model: "Find out if B eats any other organisms other than A. If there are other food sources for B, answer yes." This approach follows the spirit of CoT prompting, guiding the model through the reasoning process needed to reach the correct answer. Tab. 3 shows limited benefit overall, with only a few types improving under strategy prompts. We hypothesize that the elongated prompt might reduce the importance of the visual tokens and reduce the bias from biological priors.

3.5 Error analysis

To further examine the sources of model errors, we conducted an error analysis in Tab. 4 focusing on both logical failures and systematic error rates across question types. For Type 7 questions, which involve combining two clauses with an "or," we observe that Qwen2.5-VL achieves an accuracy of 94.05% on questions whose correct option contains two correct clauses and only 73.46% on questions whose correct option contains one correct clause and one incorrect clause. This difference of 20.59% suggests that models often ignore the semantics of disjunction: for the overall statement to be true, it suffices that only one clause is true. Instead, models frequently treated both clauses as if they were required, leading to systematic misclassification.

We also report the distribution of false positive and false negative rates for Yes/No question types (Types 1-4 and 6). Results are summarized for both synthetic and real images in Tab. 4.[3] Several patterns emerge. False negatives dominate for Type 2 in the synthetic setting in contrast to real images, highlighting reliance on pretraining priors. False positives are also high for Type 6 across models, reflecting persistent difficulty in rejecting incorrect relations.

4 Conclusion

We showed that reasoning about chained and indirect relationships in food web/chain diagrams remains challenging for modern MLLMs. The problem is especially evident when models cannot rely on pretraining knowledge. We hope our work inspires work targeting precise visual processing of structured information in images, and better image-aligned structured instruction following.

[3]For Option 1/2 question types (Types 5 and 7), the option indices are randomly assigned, so error rates are not directly comparable and are excluded.

5 Limitations

Our study aims to probe the potential challenges MLLMs face in logical reasoning VQA tasks through constructing question-answer pairs from the chained and indirect relationships in food web/chain diagrams and evaluating them with LLaVA-NeXT, InternVL and GPT-4o-mini. First, we acknowledge that this is not an exhaustive list of MLLMs, but via the observed similar trends in their results, we hope to provide a demonstration of utilizing grounded logical questions to highlight the potential improvement directions for MLLMs. We also recognize that the dataset scale is limited in our study. More images can be processed through our pipeline to extract relations and generate question-answer pairs, ultimately enhancing the generalization of the results. Secondly, we note that there are still challenges in parsing the real food web/chain images. The prey-predator relations extraction by querying GPT4o is not guaranteed to capture all relations existing on the image. Resolving the parsing difficulties here can also increase our dataset scale and make the pipeline fully automated without the need of manual filtering on the GPT4o output. Additionally, as shown in Tab. 8 in the Appendix, our preliminary trial to include strategies in the spirit of CoT reveals that models exhibit distinct behaviors across each question type, suggesting that manipulating prompt text may not provide a generalizable solution for improving performance on the logical reasoning task we propose. Meanwhile, consider the white rows in Tab. 2 and our insights on visual information being underutilized, we hypothesize that focusing on improving logical components in the image representations might be a more broadly applicable direction.

Acknowledgment

This work was supported by National Science Foundation Grant No. 2329992. This research was also supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Emden R Gansner and Stephen C North. 2000. An open graph visualization system and its applications to software engineering. *Software: practice and experience*, 30(11):1203–1233.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020*, page 379–396, Berlin, Heidelberg. Springer-Verlag.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vision*, 127(4):398–414.

Agrim Gupta, Piotr Dollár, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5359.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. 2025. Do vision-language models really understand visual language? In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR) Workshop Track*. Workshop track, ICLR 2018.

Aniruddha Kembhavi, Matteo Salvato, Erik Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016*, volume 9908 of *Lecture Notes in Computer Science*. Springer, Cham.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. 2024. Learning to localize objects improves spatial reasoning in visual-llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987.

Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. 2024. Can Vision-Language Models be a Good Guesser? Exploring VLMs for Times and Location Reasoning . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 625–634, Los Alamitos, CA, USA. IEEE Computer Society.

A Appendix

We show the prompts used to extract prey-predator relations in A.1; an example to demonstrate the relation between Type 1/2 and 5 questions in A.2; sample sizes of our datasets in A.3; question-specific strategies given to models in A.4; a model consistency exploration on real images in A.5; full results with strategy prompt in A.6; human performance study in A.7; and qualitative examples in A.8.

A.1 Relation Extraction Prompts

In our automated logical questions generation pipeline, we queried GPT4-o to analyze the food web/chain image and provide a list of provider-receiver relations of the organisms before manually filtering out the hallucinated pairs. Tab. 5 demonstrates the prompt we use for this generation.

User:
You are a sophisticated scientist. Given a food web image or a food chain image, extract all the provider and receiver relations listed using arrows between organisms in the image. Return in JSON format following ["provider": "Carrot", "receiver": "Rabbit", "provider": "Grass", "receiver": "Rabbit", "provider": "Grass", "receiver": "Grasshopper"].

Table 5: Prompt utilized in question generation pipeline.

A.2 Example of the relation between Type 1/2 and 5 questions

For the Type 5 question "Which of these organisms is food to wolf and eats grass? Select the correct statement. | Option 1: Rabbit Option 2: Grasshopper.", we can decompose it into 2 Type 1 questions with rabbit: "Does wolf eat rabbit?" and "Does rabbit eat grass?" and 2 Type 2 questions with grasshopper: "Does wolf eat grasshopper?" and "Does grasshopper eat grass?". This arises from the following fact. For the correct option rabbit, the answer to both of the two sub-questions is "yes" with existing arrows in the image, which makes them the positive cases in Type 1. And for the wrong option grasshopper, the answer will be determined by whether there is an arrow between the organisms in the image, which is exactly the

positive case and negative case in Type 2. In addition, by the definition of the wrong option, one of the two questions must have "no" as its answer which means there is no arrow.

A.3 Sample size for each question type

See Tab. 6.

Type	Real Size	Synthetic Size
1	516	2968
2	2354	5590
3	2196	5590
4	856	2803
5	863	2803
6	1129	1484
7	1442	2968
Total	9327	24206

Table 6: Sample size for each question type.

A.4 Strategies given to models to answer the questions

See Tab. 7.

A.5 Model consistency (real images)

After comparing the zero-shot performance across different logical question types, we switched gears to analyze the consistency of the MLLMs as the third aspect of logical reasoning. Type 5 questions can be viewed as the composition of a subset of Type 1 questions and Type 2 questions (see appendix for an example). If the model behaves consistently, answering the two relevant Type 1 and two Type 2 questions correctly should be enough evidence for the model to answer the target Type 5 question correctly with logically understanding the relation “and” and elimination of choices. Additionally, answering the two Type 2 correctly should also give the model enough evidence to do elimination on the choices. We find that MLLMs are fairly consistent: if all of the decomposed, simpler (Type 1/2) questions are answered correctly, the models do have a better chance to answer the composed questions correctly using the process of elimination. When both of the two Type 2 decomposed questions have “no” as the correct answer, LLaVA-NeXT achieves the composed question accuracy of 78.16% and InternVL achieves 87.9% , while if one of the Type 2 questions has “yes” as the correct answer, LLaVA-NeXT can only achieve 65.87% and InternVL achieves 74.6%. We hypothesize that this is because the organism in the Type 2 questions is completely unrelated with the two organisms mentioned in the Type 5 question text, which makes the question less confounded. Furthermore, we also noticed that the overall accuracy for the decomposed questions accuracy for LLaVA-NeXT and InternVL are 72.72% and 77.22% respectively due to the difficulty of Type 2 questions as it relies on the model understanding the absence of arrows. This is confirmed by looking at the accuracy of Type 1 and Type 2 decomposed questions separately: LLaVA-NeXT can achieve 86.17% for Type 1 but only 59.28% for Type 2.

A.6 Full model performances with strategy prompt

See Tab. 8.

A.7 Human performance

To validate that our tasks are well-defined and solvable by humans, we conducted a small-scale human study. We randomly selected 5 questions for each question type, using one synthetic and one real image, resulting in 70 questions in total. One human subject answered 68 correctly (97% accuracy). The two errors (3%) were attributable to annotation issues in the GPT-generated labels, caused by ambiguous or hard-to-see arrows in the diagrams. This negligible error rate contrasts with the much larger performance gaps observed across question types (15–30%) and between real and synthetic data (20–30%). These findings confirm that logical inference based on the images is straightforward for humans, but remains challenging for current MLLMs.

A.8 Qualitative examples

See Figures 2, 3, 4, 5.

Type	Strategy
1	For a question of 'Does A eat B?', identify A and B in the graph and check if there is an arrow pointing from B to A.
2	For a question of 'Does A eat B?', identify A and B in the graph and check if there is an arrow pointing from B to A.
3	For a question of 'Does A contain matter that was once part of B?', find out if A eats B.
4	For a question of 'Does A contain matter that was once part of B?', answer 'Yes' if A consumes matter that comes from B. For example, 'A directly eats B' or 'A eats C and C eats B'.
5	For a question of 'Which of these organisms is food to A and eats B?', first find out all the organisms that A eats, then check if any of the organisms eats B.
6	For a question of 'If A does not exist, will B still survive?', find out if B eats any other organisms other than A. If there are other food sources for B, answer yes.
7	For an option of 'A eats B or C eats D', first split the option into two parts 'A eats B' and 'C eats D'. If one of the statements is correct, the entire option is correct.

Table 7: Strategies we designed to include in the prompt for each question type.

Model	Image	Overall	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
LLaVA	Real	63.08%	77.13%	73.15%	58.56%	66.24%	76.36%	45.26%	52.64%
	Real + S	59.59%	86.70%	66.70%	56.60%	61.33%	62.60%	50.82%	50.17%
	Synthetic	51.58%	88.06%	50.09%	47.57%	47.80%	51.96%	49.83%	50.81%
	Synthetic + S	51.74%	92.25%	50.02%	47.06%	48.03%	51.87%	50.24%	51.01%
InternVL	Real	74.23%	95.74%	78.12%	65.89%	76.29%	86.10%	65.63%	71.29%
	Real + S	67.13%	92.64%	59.26%	73.36%	79.56%	78.45%	50.66%	60.12%
	Synthetic	53.88%	78.21%	51.07%	52.16%	51.07%	58.95%	51.78%	51.55%
	Synthetic + S	54.40%	80.95%	51.36%	56.39%	49.76%	58.26%	48.85%	50.54%
Qwen-VL	Real	77.62%	98.26%	84.28%	66.21%	88.55%	91.31%	61.03%	75.03%
	Real + S	76.21%	98.45%	74.81%	79.60%	83.18%	81.00%	54.83%	75.10%
	Synthetic	59.31%	77.30%	53.95%	63.38%	62.65%	63.45%	52.13%	53.20%
	Synthetic + S	56.88%	79.49%	52.43%	61.72%	51.30%	57.19%	56.39%	54.55%
GPT	Real	70.12%	96.12%	79.61%	55.92%	55.96%	83.66%	67.85%	69.00%
	Real + S	70.37%	93.70%	73.29%	69.22%	67.28%	82.55%	56.13%	64.57%
	Synthetic	51.72%	66.52%	50.19%	55.60%	50.52%	51.64%	50.50%	46.38%
	Synthetic + S	50.07%	64.29%	50.00%	48.37%	47.20%	50.48%	49.50%	49.53%

Table 8: Zero-shot accuracy for LLaVA-NeXT (Liu et al., 2024), InternVL2.5 (Chen et al., 2024), Qwen2.5-VL (Bai et al., 2025) and GPT-4o-mini (OpenAI, 2024) with pure QA-style prompts and prompts that include strategies for each question. + S is indicating the addition of strategies in the prompt.

Example Real Image of 5 Relations

Type 1
Q: Does Snake eat Hawk? A: No.

LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: No.

Type 2
Q: Does Grasshopper eat Grass? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 3
Q: Does Grasshopper contain matter that was once part of Grass? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 4
Q: Does Frog contain matter that was once part of from Hawk? A: No.

LLaVA: Yes.
InternVL: No.
Qwen-VL: No.
GPT: Yes.

Type 5
Q: Which of these organisms is food to Frog and eats Grass? Select the correct statement. | Option 1: Grasshopper Option 2: Snake. A: Option 1.

LLaVA: Option 1.
InternVL: Option 1.
Qwen-VL: Option 1.
GPT: Option 1.

Type 6
Q: If Grass does not exist, will Grasshopper still survive? A: No.

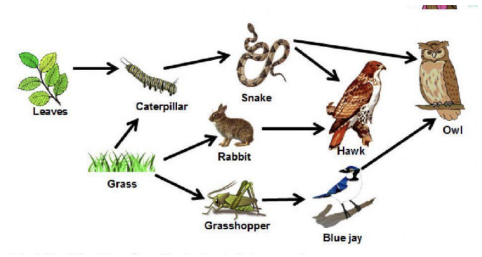
LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: No.

Type 7
Q: Select the correct statement. | Option 1: Grasshopper eats grass or grasshopper eats hawk. Option 2: Hawk eats snake or hawk eats fungi. A: Option 2.

LLaVA: Option 1.
InternVL: Option 2.
Qwen-VL: Option 1.
GPT: Option 1.

Figure 2: Examples of a real food chain image with 5 relations in the image and the partial results of the generated question-answer pairs. Dark green coloring means that the answer given by the model is correct, while red coloring means that the answer is wrong.

Example Real Image of 10 Relations



Type 1
Q: Does Caterpillar eat Snake? A: No.

LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: No.

Type 2
Q: Does Caterpillar eat Leaves? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 3
Q: Does Caterpillar contain matter that was once part of Leaves? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 4
Q: Does Snake contain matter that was once part of Grass? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 5
Q: Which of these organisms is food to Snake and eats Leaves? Select the correct statement. | Option 1: Grass Option 2: Caterpillar. A: Option 2.

LLaVA: Option 2.
InternVL: Option 2.
Qwen-VL: Option 2.
GPT: Option 1.

Type 6
Q: If Leaves do not exist, will Caterpillar still survive? A: Yes.

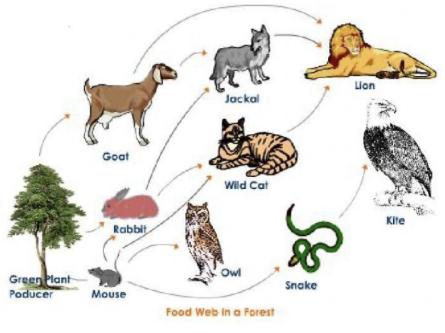
LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: No.

Type 7
Q: Select the correct statement. | Option 1: Caterpillar eats leaves or caterpillar eats snake. Option 2: Grasshopper eats rabbit or grasshopper eats blue jay. A: Option 1.

LLaVA: Option 1.
InternVL: Option 1.
Qwen-VL: Option 1.
GPT: Option 1.

Figure 3: Examples of a real food web image with 10 relations in the image and the partial results of the generated question-answer pairs. Dark green coloring means that the answer given by the model is correct, while red coloring means that the answer is wrong.

Example Real Image of more than 10 Relations



Type 1
Q: Does Wild Cat eat Jackal? A: No.

LLaVA: Yes.
InternVL: No.
Qwen-VL: No.
GPT: No.

Type 2
Q: Does Kite eat Owl? A: Yes.

LLaVA: No.
InternVL: No.
Qwen-VL: Yes.
GPT: No.

Type 3
Q: Does Owl contain matter that was once part of Green Plant? A: No.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 4
Q: Does Wild Cat contain matter that was once part of Green Plant? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 5
Q: Which of these organisms is food to Jackal and eats Green Plant? Select the correct statement. | Option 1: Goat Option 2: Mouse. A: Option 1.

LLaVA: Option 1.
InternVL: Option 2.
Qwen-VL: Option 1.
GPT: Option 1.

Type 6
Q: If Green Plant does not exist, will Goat still survive? A: No.

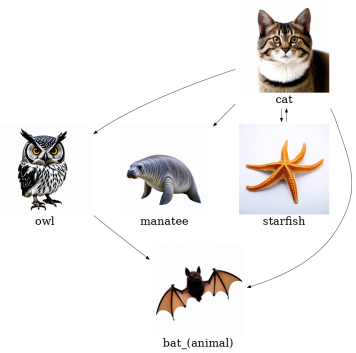
LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: Yes.

Type 7
Q: Select the correct statement. | Option 1: Goat eats Green Plant or Jackal eats Goat. Option 2: Rabbit eats Wild Cat or Mouse eats Snake. A: Option 1.

LLaVA: Option 1.
InternVL: Option 1.
Qwen-VL: Option 1.
GPT: Option 1.

Figure 4: Examples of a real food web image with more than 10 relations in the image and the partial results of the generated question-answer pairs. Dark green coloring means that the answer given by the model is correct, while red coloring means that the answer is wrong.

Example Synthetic Image of 5 animals



Type 1
Q: Does starfish eat cat? A: No.

LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: No.

Type 2
Q: Does cat eat starfish? A: Yes.

LLaVA: No.
InternVL: Yes.
Qwen-VL: Yes.
GPT: No.

Type 3
Q: Does cat contain matter that was once part of starfish? A: Yes.

LLaVA: No.
InternVL: Yes.
Qwen-VL: Yes.
GPT: Yes.

Type 4
Q: Does owl contain matter that was once part of starfish? A: Yes.

LLaVA: No.
InternVL: No.
Qwen-VL: No.
GPT: Yes.

Type 5
Q: Which of these organisms is food to bat$_{(}$animal$_{)}$ and eats starfish? Select the correct statement. | Option 1: cat Option 2: owl. A: Option 1.

LLaVA: Option 2.
InternVL: Option 2.
Qwen-VL: Option 2.
GPT: Option 2.

Type 6
Q: If owl does not exist, will bat$_{(}$animal$_{)}$ still survive? A: Yes.

LLaVA: Yes.
InternVL: Yes.
Qwen-VL: No.
GPT: Yes.

Type 7
Q: Select the correct statement. | Option 1: owl eats bat$_{(}$animal$_{)}$ or manatee eats starfish. Option 2: cat eats starfish or cat eats bat$_{(}$animal$_{)}$. A: Option 2.

LLaVA: Option 1.
InternVL: Option 1.
Qwen-VL: Option 2.
GPT: Option 1.

Figure 5: Examples of a synthetic food web image with 5 animals in the image and the partial results of the generated question-answer pairs. Dark green coloring means that the answer given by the model is correct, while red coloring means that the answer is wrong.