SCIREC: DIAGNOSTIC EVALUATION OF MULTIMODAL MULTI-TURN RELATIONAL REASONING WITH ADAPTIVE INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Relational reasoning requires the process of perceptual understanding, comparing, and integrating the underlying relationships between concepts. This ability consists of multiple categories, such as analogical, structural, and cause-effect, each capturing a different aspect of higher-order understanding. To examine the performance of multimodal large language models (MLLM) on these relational inference tasks, we develop SciReC, a model-adaptive multimodal academic dialogue benchmark. As the relational reasoning process involves multiple representations and various factors (visual understanding, exhibiting knowledge, and memory recall), we propose DMRA, a deficit-based diagnostic framework that quantifies the contribution of these components to identify the primary reason of unsuccessful cases. Claude 4.6 achieved the best performance on overall relational score with 73%, followed by GPT 5.4 with 68%. Performance trends indicate that open-source models achieve their lowest scores on spatial relations, while proprietary models struggle more with hierarchical and sequential relations. Additionally, among the evaluated knowledge domains, astronomy consistently yields the lowest scores for most models. The results of DMRA reveal that relational reasoning is the primary source of error across all models, followed by memory limitations.

1 INTRODUCTION

Relational reasoning is a fundamental factor of intelligence (Duncan, 2003), and this capacity sets human cognition apart from the abilities of other animals. (Holyoak & Thagard, 1995; Oden et al., 2001). This ability to extract relevant information between entities is linked to logical thinking and problem-solving in novel scenarios (Cattell, 1987; Halford et al., 1998). Scientific concepts in STEM education and social sciences, such as biology, physics, chemistry, sociology, and psychology, consist of examples of relational concepts that describe different levels of various scenarios (Goldwater & Schalk, 2016). Even basic scientific concepts require learners to develop representations of relational knowledge, and achieving expertise further depends on building interconnected concepts and classifying problems by their underlying relational structure (Goldwater & Schalk, 2016). The acquisition of relational concepts and their combinations, based on scientific relations, supports memory retrieval and problem-solving.

Multi-image relational reasoning in multimodal AI is addressed through diverse tasks and datasets, including analogical reasoning, which focuses on extracting and applying relational rules, (Yilmaz et al., 2025; Bitton et al., 2022; Zhang et al., 2019), single-turn QA (Johnson et al., 2016; Nie et al., 2025; Wang et al., 2024), and multi-turn dialogue (Liu et al., 2024b; Jiang et al., 2024), each capturing different aspects of relational structure. In a multi-turn conversation, scenarios require cross-context integration; however, many of these benchmarks rely on predefined conversation scenarios that are far from real-world dialogue interactions. In contrast, successful real-life conversations rely on the participation of both parties, so the context of the dialogue shifts dynamically in response to what each side says. However, current conversation datasets do not engage with the output of the AI models and continue the conversation with static inputs. Although they evaluate different model capabilities, their non-interactive and one-sided message handling often leads to incoherent and poorly grounded human-AI dialogues. To better evaluate the human-AI interaction abilities along with long-context history and multimodal inputs, double-sided and model-adaptive benchmarks are necessary.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In many evaluation tasks, the atomic abilities or conditions are ignored and not included in the evaluation process, despite their influence on model performance and their potential to mislead results. This limitation is especially critical considering that all reasoning processes begin with perception (Peirce, 2012). Visual understanding is the starting point in multimodal tasks, and without processing visual data, the models are not expected to reliably perform reasoning. This is similar to knowledge-based multimodal relational reasoning tasks, where models are required to integrate visual data with conceptual information from multiple sources. Missing information about any part of the relation prevents the model from forming a complete and accurate inference. Therefore, a diagnostic approach is essential to identify the main failures of model performance and behavior.

In response to these challenges, we introduce SciReC, a multi-image, multi-turn, model-adaptive benchmark for relational reasoning in academic domains, which evaluates the models' performance across eight categories of relational reasoning (comparative, sequential, cause-effect, hierarchical, spatial, structural-functional, analogical, and other), along with their knowledge. These relational questions require visual perception to interpret the context, memory to retain and retrieve relevant details, and the integration of knowledge to infer the underlying relationships.

Our SciReC benchmark consists of the following features: (1) **Academic context**: Instead of daily-life conversations, our benchmark focuses on university-level subject contexts, including physics, biology, and chemistry, shifting dialogue questions from general settings to knowledge-based domains. (2) **Adaptive multimodal dialogue**: Unlike prior datasets, our benchmark adjusts the conversation flow according to the models' evaluated output, which determines the context of the questions. Instead of offline conversations, it provides a dynamic structure that allows flexibility in both the number of images and the number of conversational turns. (3) **Diagnostic error analysis**: Besides evaluating the model's relational reasoning capability in a knowledge-based domain, we also assess the models' knowledge gap, visual understanding, and memory retention abilities in the same concept. These categories support a diagnostic approach by identifying the main failures in relational reasoning tasks.

Our contributions and observations are summarized below:

- We introduce SciReC, a model-adaptive multimodal academic conversation benchmark designed to evaluate the relational reasoning of multimodal large language models across eight categories in diverse domains, incorporating three key factors: visual, knowledge-based, and memory.
- We evaluate state-of-the-art models on SciReC questions, achieving up to 73% accuracy (proprietary) and 56% (open-source). Through detailed analysis, we identify each model's strengths and weaknesses across relational reasoning categories and scientific and social science domains.
- We introduce DMRA, a comprehensive diagnostic evaluation framework that identifies the root of failure cases through deficit-based analysis. The results indicate that relational reasoning is the primary factor across all models, followed by memory.

2 Related Work

2.1 Multimodal Multi-turn Datasets

To evaluate the multimodal dialog capability of current models, many benchmarks have been introduced. While MMDU (Liu et al., 2024b) focuses on understanding the conversation using relations between multiple images, LoCoMo (Bei et al., 2026), and Mem-Gallery (Maharana et al., 2024) address the memory issues of the models. MMCR-Bench (Yan et al., 2025) examines the logical consistency of conversations by checking contextual referencing, and MMRC (Xue et al., 2025) evaluates memory and reasoning tasks in conversations, such as cross-turn reasoning and image management. Mantis-Eval (Jiang et al., 2024) evaluates co-reference, comparison, reasoning, and temporal understanding of models in dialogues. While MultiVerse (Lee et al., 2025) evaluates different aspects of the models using a checklist, ConvBench (Liu et al., 2024a) utilizes a three-level hierarchical evaluation: perception, reasoning, and creation. Like their context, their domains also differ; some of them (Maharana et al., 2024; Bei et al., 2026; Liu et al., 2024b; Jiang et al., 2024; Yan et al., 2025) consist only of real-life conversations, and others (Xue et al., 2025; Lee et al., 2025; Liu et al., 2024a) include both scientific and daily life dialogues. Inter-image benchmarks focus on daily-life contexts with limited relational diversity, while science benchmarks assess knowledge gaps rather than relational reasoning. SciReC bridges this gap by evaluating relational reasoning in multimodal conversations in both scientific and social science domains, as summarized in Figure 1.

	Error Analysis	Relation	Knowledge	Visual	Memory
MMCR	✗	✓	✗	✓	✗
MultiVerse	✗	✗	✓	✓	✓
Mantis-Eval	✗	✓	✗	✓	✗
MMDU	✗	✓	✗	✓	✗
Mem-Gallery	✗	✗	✗	✗	✓
MMRC	Pattern-based	✗	✓	✓	✓
LoCoMo	Qualitative	✗	✓	✗	✓
ConvBench	Hierarchical	✗	✓	✓	✗
SciReC	Deficit-based	✓	✓	✓	✓

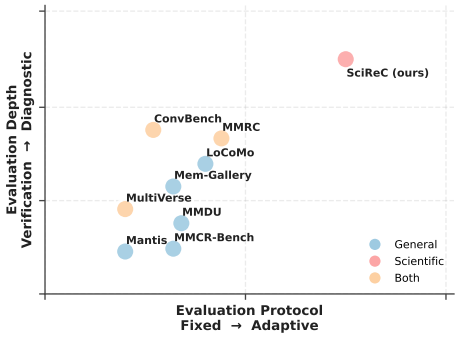


Figure 1: The table compares multimodal conversational datasets in terms of context and error analysis, while the graph illustrates their comparisons across evaluation protocols and depth.

2.2 Evaluation Protocols and Diagnostic Error Analysis

Most of the multimodal multi-turn benchmarks utilized existing datasets as a data source (Yan et al., 2025; Lee et al., 2025; Jiang et al., 2024; Bei et al., 2026; Liu et al., 2024a), a few of them collected conversations from human-AI interaction (Xue et al., 2025) and AI-AI interaction (Maharana et al., 2024). Although interactions make dialogs dynamic during data collection, evaluation is typically conducted in fixed conversational sequences in which the performance of the model does not influence the next turns, resulting in a static interaction flow; see Figure 1. In contrast, SciReC provides adaptive conversational evaluation over a fixed question pool where the subsequent context is determined with a model-dependent approach. Although the question set is predefined, the evaluation process adopts a conditional interaction flow, where the selection of future question types, Knowledge Gap, Visual Perception, and Memory Retrieval, depends on whether the model's performance exceeds a predefined threshold. By implementing this approach, SciReC enables conditionally adaptive multi-turn reasoning evaluation, rather than an offline process.

Most benchmarks in Figure 1 do not provide explicit error analysis of model performance in multimodal conversations. Their evaluation approaches primarily rely on LLM-based scoring. Although MULTIVERSE (Lee et al., 2025) employs a checklist-based metric for evaluation, it does not provide error analysis derived from these criteria. While LoCoMo (Maharana et al., 2024) identifies common errors through manual qualitative analysis, MMRC (Xue et al., 2025) conducts pattern-based analysis. ConvBench (Liu et al., 2024a) adopts hierarchical analysis among perception, reasoning, and creation processes. Although these analyses provide detailed failure patterns, they do not attribute failures to underlying rationales with systematic decomposition. Unlike these analyses, our benchmark employs deficit-based diagnostic analysis, which quantitatively identifies the relative deficit of multiple components to explore the potential error source. The method provides diagnostic insights into model failures by focusing on the weakest-performing component.

3 SciReC Benchmark

3.1 Benchmark Overview

The SciReC benchmark is designed to evaluate the relational reasoning capabilities of current multimodal large language models in multimodal multi-turn academic conversations. This benchmark challenges models to integrate a broad range of complex concepts in two images by exploring the underlying relational structure. The dataset consists of five types of questions: relational reasoning, knowledge-based, visual understanding, memory retrieval, and memory validation, which verify whether memory-based questions are answerable using the images. We categorize relations into eight types: comparative (similarity/difference), spatial (physical location, position), sequential (temporal ordering), cause–effect (process-outcome), structural–functional (how system components function and interact), hierarchical (part–whole structures), analogical (relational mapping across domains), and other (exemplification, explanation, etc.). While relational questions demand information drawn

from multiple images, other question types are specific to a single image, and figures are not provided for relational reasoning and memory questions. Figure 7 displays the dataset creation pipeline.

Since relational questions are extracted separately from each chapter, the number of generated dialogues equals the number of chapters. In total, 189 multimodal multi-turn conversations with 656 relational questions are generated along with ground truths extracted from the textbooks; for details, please refer to Appendix A. For each image, at least three knowledge-based and visual perception questions are created, along with one memory question and one memory validation question. The number of turns in each dialog varies depending on the number of relations identified in the chapters and the performance of the models, which can lead to the image questions varying across models.

3.2 Benchmark Construction

Data Extraction and Annotation To evaluate the relational reasoning abilities of current models across academic disciplines, we utilized 12 college-level, open-source textbooks in different domains: Biology (Clark et al., 2018), Chemistry (McMurry, 2023; Flowers et al., 2019), Physics (Ling et al., 2016; Moebs et al., 2016), Astronomy (Fraknoi et al., 2022), Economics (Greenlaw et al., 2022), Psychology (Spielman et al., 2020), Behavioral Neuroscience (Kirby et al., 2024), and Calculus (Strang & Herman, 2016a;b;c). The chapters and images in the textbooks are extracted, and figures are manually filtered to exclude those that do not rely on contextual knowledge. The textbook chapters serve as the source for question creation, providing captions and the surrounding text, including all information about images. During the relation extraction and question generation processes, we employed GPT-5.4, which identified image-related content in chapters by the ”Figure” tag. Relations and source texts were extracted from this textual content, **without access to the images**. Since the source texts play a key role for task evaluation, we ensure their accuracy through carefully designed prompts (see Appendix E) that restrict the extraction process to using only textual content in the textbook, thereby mitigating the risk of hallucinating information not present in the textbook. In the next step, extracted relationships are validated by filtering out unclear ones by the model. Then it categorizes each relation into eight types and converts the source text into ground truths with the relational clause linking two figures. To avoid hallucinations, ground truths are generated to strictly follow the source text without introducing any external information; see Appendix E.

Question Generation. After final validation of the results, the questions are generated utilizing the relation types, ground truths, and question templates. To prevent answer leakage and ensure structural consistency, we employed templates for both visual and relational questions. Since the knowledge-based questions obtain a clear structure and context directly extracted from the texts, we did not apply templates. A variety of questions is ensured through relational templates with several variants per relation type, and visual templates spanning eight categories with multiple question forms in each (see Appendix E.1). The model fills the placeholders (concepts, attributes, figure names) with words or phrases that appear in the ground truths to complete the question generation process. On the other hand, memory and memory validation questions are generated based on the exact facts needed to answer the relational questions. Although they share the same answers, memory validation questions are slightly paraphrased from memory questions and augmented with images to evaluate whether visual context improves model performance. **The model generates questions exclusively from the textual content, without accessing the images.** The textbook captions and surrounding paragraphs are used only during the offline benchmark construction to extract relations, generate ground truths, and create questions. However, during evaluation, models receive only images and questions, making their answers independent of textbook captions and textbook context.

Quality Control with Human Annotators. After verifying relations, category type, and relational questions with GPT 5.4, human annotators validated the quality of the generated relational question sets. We selected relational Q&A pairs: a single question from each chapter of the Biology and Physics textbooks for a multi-round manual review process by professionals. For evaluation, we defined three criteria: (1) relevance for relationships: Does the generated question accurately capture the intended relationship between the figures? (2) accuracy: Is the provided ground truth actually correct? (3) visual grounding: Can the question be answered correctly without looking at the images, relying only on generic or commonly known domain knowledge? For each round, annotators reviewed these three criteria and gave feedback about the correct and failure cases. According to these findings, we modified the prompts, dataset creation pipeline, and templates to mitigate the errors. This

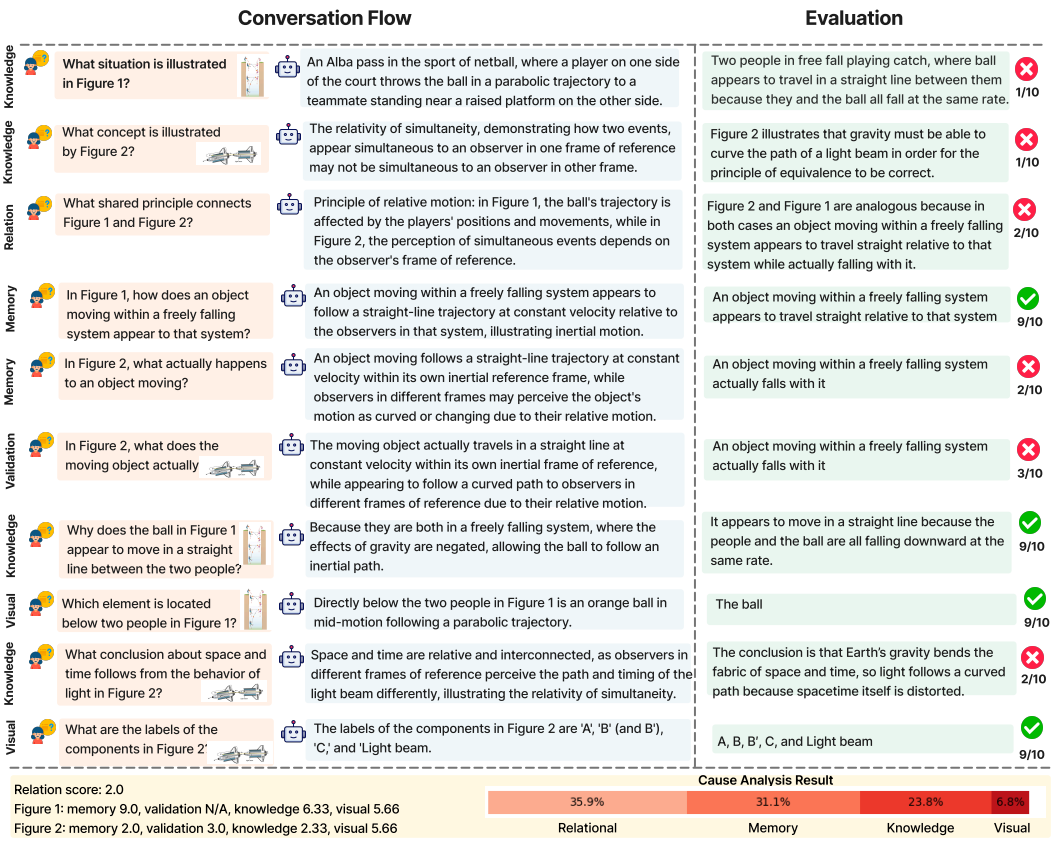


Figure 2: Example conversation flow and evaluation for an analogical relational question in astronomy. The panels show the question, model response, and ground truth with scores, respectively. The next question type is adjusted based on the scores. The last row shows the average score across all question types and the main failure rates. While relational reasoning is the primary reason (35.9%), the other three factors in the bar chart are associated with Figure 2, and the remaining portion (2.4%) corresponds to Figure 1. For better visualization of figures and detailed calculations, see Appendix C.

process ensures that the relational questions in the dataset are constructed using categorically accurate templates, grounded in reliable ground truth, and depend on visual information. Once the reviews met the quality standards, we finalized the pipeline and created the dataset. Then we **manually reviewed each relational question** alongside the corresponding figures, ground truths, and textbook source, and removed any inconsistent or unclear instances. This elimination process is applied only to relational questions, as they require integrating multiple concepts across figures, making them more error-prone, while other types can be directly and reliably extracted from textbooks.

Adaptive Dialogue Creation. The dialogue, illustrated in Figure 2, starts with a knowledge-based question that focuses on understanding the main concept depicted in the images. The next turn presents a relational reasoning question between two figures without displaying the images. The model's response is immediately evaluated using a scoring model (Claude 4.6), and subsequent question types are determined based on whether the score exceeds a predefined threshold (7/10). If the score falls below the threshold, image-specific memory retrieval questions are asked in the next turns without showing the images. If the model's performance remains below the threshold, a memory validation question is then posed for the failed image, using that image as input. Success on the validation question, despite failure on the memory question, indicates a memory retrieval issue. If both are answered incorrectly, this suggests a knowledge gap or a visual perception error. Finally, knowledge-based and visual questions are asked for both images; however, if already covered in earlier turns, they are skipped, while retaining them for diagnostic analysis. The process repeats for the next relations, making the conversation adaptive and model-dependent based on performance.

3.3 EVALUATION

To evaluate the response of models, we utilized two scoring rubrics: relational and other image-specific tasks. Although the rubrics differ, they share an identical scoring scheme, with responses evaluated on a 1–10 scale: 1–3 incorrect, 4–6 partial, 7–8 mostly correct, and 9–10 complete and accurate. A score of 7 is used as the threshold for both rubrics to classify an answer as correct, with scores below 7 indicating partially correct or incorrect responses. We evaluated two state-of-the-art models: Claude 4.6 (Anthropic, 2026) and GPT 5.4 (OpenAI, 2026) for scoring the model answers. **Their evaluations reached 85.4% consensus across 199 model responses, indicating that the interaction flow is largely consistent regardless of the evaluator model selected.** Disagreement cases were further analyzed to determine which evaluator better aligned with human judgments. After evaluation, we selected to use Claude 4.6, which demonstrates strong alignment with human reviews. For detailed evaluation results and prompts, please refer to Appendix B and Appendix E.

4 DEFICIT-BASED MULTIMODAL RELATIONAL ANALYSIS (DMRA)

To diagnose whether the failed responses are caused by cross-image reasoning or other factors like visual encoding, knowledge gap, and memory retention, we implemented a controlled deficit-based diagnostic framework, DMRA. As the success of relational reasoning depends on the integration of these factors, we employed a two-stage error decomposition, distinguishing between upstream and relational causes. While the upstream stage assesses whether the relational reasoning is weak because of figure-level deficits, the relational stage captures cross-figure reasoning errors such as integration and abstraction. For each figure i, scores: memory M_i, knowledge K_i, visual V_i, and memory validation MV_i are used to compute upstream deficits $D_x^{(i)}$ with threshold $T = 7$, (eq. 1).

$$D_x^{(i)} = \max(0,\, T - X_i), \quad x \in \{M, K, V\}. \tag{1}$$

We employ memory validation questions to isolate memory errors from perception and knowledge errors. The score difference between memory and validation questions, g_i, dynamically adjusts the weights of each factor; see Figure 3. We utilize the softmax function to change the weights, use equal base weights for each task $b_M = b_K = b_V = 0.33$, and select a validation effect parameter $\lambda = 0.2$. In case the memory scores equal the validation scores, we define a bias parameter $\tau = 0.1$, which slightly reduces the memory weight, as the score does not improve with images; see Equation 2. As shown in Equation 3. The importance score of the memory task $z_M^{(i)}$ is updated based on its base weight, validation gap, and bias. Using the importance scores, the softmax function normalizes the weights of each task $w_x^{(i)}$.

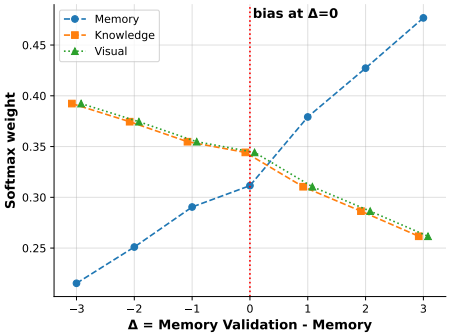


Figure 3: Softmax weights vs Δ

$$g_i = \begin{cases} MV_i - M_i, & MV_i \neq \varnothing \\ 0, & MV_i = \varnothing \end{cases}, \qquad \beta_i = \begin{cases} \tau, & MV_i \neq \varnothing \text{ and } MV_i = M_i \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$z_M^{(i)} = b_M + \lambda g_i - \beta_i, \quad z_K^{(i)} = b_K, \quad z_V^{(i)} = b_V, \qquad w_x^{(i)} = \frac{e^{z_x^{(i)}}}{e^{z_M^{(i)}} + e^{z_K^{(i)}} + e^{z_V^{(i)}}}. \tag{3}$$

The weight of each figure α_i is calculated by normalizing each figure's total deficit $D^{(i)}$ (with a smoothing term $\lambda_{\text{fig}} = 0.5$) by the sum of all smoothed deficits of all figures $D^{(j)}$, ensuring balanced and non-zero contributions. The contribution of each upstream task $R_x^{(i)}$ is computed by multiplying the figure weight α_i, the final task weight $w_x^{(i)}$, and task deficit score $D_x^{(i)}$ associated with figure i. The total upstream failures U are computed by aggregating contributions of upstream tasks across all figures; see Equation 4.

$$\alpha_i = \frac{D^{(i)} + \lambda_{\text{fig}}}{\sum_j \left(D^{(j)} + \lambda_{\text{fig}}\right)}, \qquad R_x^{(i)} = \alpha_i\, w_x^{(i)}\, D_x^{(i)}, \qquad U = \sum_x \sum_i R_x^{(i)}. \tag{4}$$

For each question, the relation score R is used to quantify the deviation of the answer from the threshold, where F represents total failure. To find the unexplained contribution C_{rel} caused by relational reasoning, we subtract the explainable part from the total failure. If $U \geq F$, then the outcome C_{rel} can be attributed to limitations in the upstream task rather than to a relational error. In contrast, if the upstream score is lower than the total failure, the difference between them, C_{rel}, corresponds to the relational error score. To share the failures proportionally within the total failure, we scale the upstream errors $s = \frac{\min(F,U)}{U}$ when $U > 0$, which prevents them from exceeding the total failure; otherwise, $s = 0$. The scaled contribution of upstream tasks $C_x^{(i)}$ is computed by multiplying their unscaled contributions by the scale parameter; refer to Equation 5.

$$F = \max(0, T - R), \quad C_{\text{rel}} = F - \min(F, U), \quad C_x^{(i)} = s\, R_x^{(i)}. \tag{5}$$

The total amount of attributed failure $Total$ is calculated by summing up the upstream scores with the relational error score. To represent each task's percentage contribution to the total failure, we normalize the scores by dividing each scaled contribution of tasks to the total failure; see Equation 6. This provides a quantitative analysis of the error underlying the model's incorrect relational answers.

$$Total = \sum_i \sum_x C_x^{(i)} + C_{\text{rel}}, \quad C \in \{C_x^{(i)}\} \cup \{C_{\text{rel}}\}, \quad P = \frac{C}{Total} \times 100. \tag{6}$$

After normalization, we aggregate figure-level scores into relational, memory, knowledge, and visual categories. For each failed relational question, categories are ranked by percentage, and the top two are recorded as primary and secondary causes; their frequencies are then reported as percentages across all failed questions. This deficit-based approach offers a diagnostic evaluation of the models' weaknesses and the primary failure in the composite task.

5 Evaluation Results

We evaluate open-sourced and proprietary models on SciReC: GPT-5.4 (OpenAI, 2026), Qwen3.5 (Team, 2026) and Claude 4.6 (Anthropic, 2026), Gemma-3-27B (Team et al., 2025), Mistral3 (Liu et al., 2026), MiniCPM-V 4.5 (Yu et al., 2025) and InternVL3 (Zhu et al., 2025).

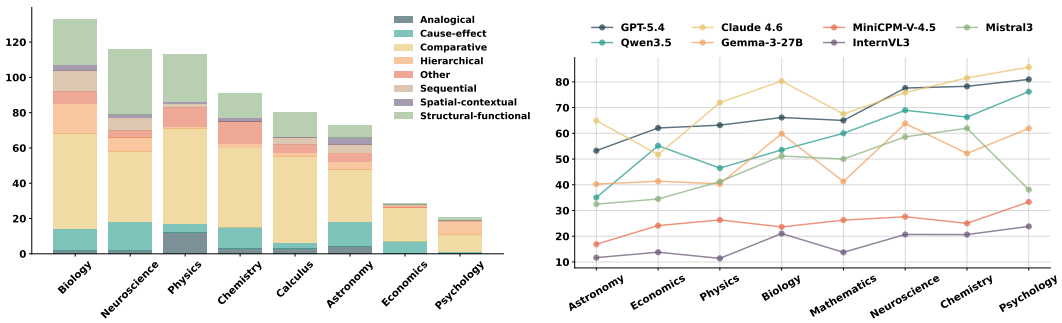
Table 1: Experiment results of models across relational categories. The models are ordered by their overall performance, and the color scale reflects their relative performance within each model: red indicates the lowest scores, while green indicates the highest.

Models	Comparative	Analogical	Sequential	Structural	Cause-Effect	Spatial	Hierarchical	Other	Overall
Claude4.6	70.10	80.77	64.52	82.81	79.71	75.00	70.00	69.39	73.78
GPT-5.4	65.12	61.54	64.52	78.91	78.26	66.67	55.00	59.18	67.99
Qwen-3.5-9B	51.83	76.92	51.61	60.00	66.67	41.67	50.00	59.18	56.25
Gemma3-27B	45.51	50.00	41.94	61.71	59.42	41.67	45.00	55.10	50.76
Mistral3	45.85	57.69	45.16	50.78	57.97	33.33	47.50	51.02	48.78
MiniCPM-V	22.26	38.46	19.35	24.80	26.9	33.33	25.00	34.69	24.85
InternVL3	15.61	3.85	19.35	18.75	21.74	25.00	22.50	14.29	17.07

5.1 Main Results

We evaluated the relational reasoning abilities of current MLLMs on the SciReC benchmark. Table 1 presents the accuracy of the models on each relational task and overall. Our evaluation shows that Claude 4.6 is the best-performing model with 73.78% overall accuracy, achieving the highest scores for all categories. GPT-5.4 and Qwen-3.5 follow Claude4.6, yielding 68% and 56.25%, respectively. While Gemma3 and Mistral3 reach the closest overall performance to Qwen-3.5 with 50.77% and 48.78%, in order, other open-source models score below 24%. The results indicate that although the gap in relational reasoning between some open-source and proprietary models is narrowing, open-source models are also separating into distinct performance tiers. For example, Qwen-3.5 performs better than GPT-5.4 on analogical relational questions and shows the same achievement for the category "Other," and it achieved over twice the overall score of MiniCPM-V and InternVL3.

Model performance varies across relational categories, which shows their weaknesses and strengths for relational reasoning. While Claude 4.6, GPT-5.4, and Gemma-3 achieve their highest score on



(a) Distribution of relation types across domains.

(b) Performance across knowledge domains.

Figure 4: Overall comparison across domains and relation distributions.

structural relations, Claude 4.6 struggles in sequential relations (64%), GPT-5.4 in hierarchical relations (55%), and Gemma-3 in spatial relations (41%). Both proprietary models excel in understanding and explaining the structure of the concept and its functions. However, GPT-5.4 has difficulty with reasoning across multiple levels of organization, and Claude4.6 shows limitations in understanding order and step-by-step processes, same as MiniCPM-V. On the other hand, spatial relations remain challenging for many open-source models, suggesting that they lack an understanding of positional configuration within concepts. While Qwen3-5 demonstrates strong performance in analogical relations with 77%, which transfer knowledge across different concepts, other high-performing open-source models, Gemma3 and Mistral3, achieve their highest score on cause and effect relations, which identify how changes in one factor influence another. Although analogical relations represent the highest-performing class for MiniCPM-V (38%), they are the lowest score for InternVL (3.8%).

5.2 Performance Across Domains

SciReC consists of various relational reasoning questions across eight academic domains; see Figure 4. While domains like Biology and Behavioral Neuroscience primarily involve sequential and hierarchical relations, Physics and Chemistry do not exhibit the same inherent relational categories. Comparative relations dominate across all domains, while structural–functional and ”other” relations are present in most domains but appear less frequently in Economics and Psychology in the dataset. Analogical relations in SciReC are mostly present in Physics and Astronomy, and spatial relations are mainly derived from Astronomy. Calculus and Psychology questions are not mainly focused on cause-and-effect relations in SciReC. Relational question examples are provided in Appendix D.

As illustrated in Figure 4, most of the models struggle to correctly answer relational reasoning questions in the Astronomy domain, except Claude 4.6, which shows weaker performance in Economics (52%), where relations are mostly represented in graphs. In contrast, the Psychology domain obtains the highest accuracy for most of the models, except Mistral3, which shows the lowest performance (38%) after InternVL3 and MiniCPM-V-4.5. Claude 4.6 obtains the highest accuracy across all domains except Economics and Behavioral Neuroscience, where GPT-5.4 outperforms it with approximately 62% and 78%, respectively. While Gemma-3 and InternVL3 show better performance on life sciences (Biology, Behavioral Neuroscience, and Psychology), at approximately 62% and 23%, their performance decreases with quantitative sciences (Calculus and Physics) to 40% and 14%, respectively. MiniCPM-V-4.5 shows relatively consistent performance across all domains, ranging between 18% and 33%. While the success of Qwen3-5 in Economics, Calculus, Behavioral Neuroscience, and Psychology is close to the proprietary models up to a 10% difference, its Astronomy, Physics, Biology, and Chemistry performances indicate a substantial gap.

5.3 Diagnostic Analysis of Model Errors

To analyze failures in relational questions, we use the DMRA framework with a two-stage deficit-based approach. As illustrated in Figure 5, the result shows that relational reasoning, which requires integration of the underlying relationship between two concepts, causes primarily performance breakdowns for all models, except MiniCPM-V-4.5. The high-performance models, GPT-5.4, Claude

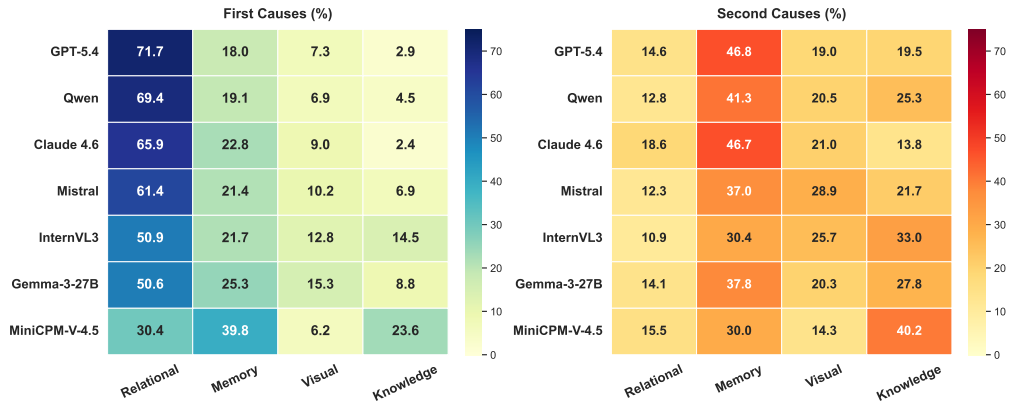


Figure 5: Results of diagnostic analysis per model showing first and second reasons

4.6, and Qwen-3.5, show high relational reasoning error between 66% and 71%, and low visual and knowledge causes below 9% and 4.5%, respectively. This reveals that these models obtain domain-specific knowledge and understand the visual inputs; however, they struggle to connect concepts. While Mistral 3 shows similar relational reasoning (61%) and memory-related (21%) causality with these models, its visual and knowledge-based error rates increase to 10% and 7%. Relational failure drops to 51% for InternVL3 and Gemma-3, but their knowledge gap and visual limitation increase. Memory accounts for 18–25% as a first reason across models, except MiniCPM-v-4.5, where memory is the main challenge (39%), followed by relational reasoning (30%) and knowledge gap (23%).

Across most models, memory-related issues rank as the second main failure (37–46%); however, for InternVL3 and MiniCPM-V-4.5, knowledge gaps instead emerge as the second leading cause (up to 40%). While knowledge gaps and limited visual understanding often emerge as secondary reasons, their impact varies across models. For example, Claude 4.6 shows the lowest knowledge gap (13%) as the second factor, whereas Qwen 3.5 exhibits a higher contribution (25%), and GPT-5.4 shows a balanced impact (19%) between visual and knowledge-related causes. Across other open-source models, knowledge-based deficiencies exceed visual limitations, ranging from 7% to 26%, except for Mistral 3, where visual limitations are higher by 7%. Overall, the results show that strong models primarily fail at relational reasoning, whereas other models struggle more with upstream tasks.

Although strong models address the knowledge gap and visual perception errors, they struggle with memory bottlenecks, just like other models. Limitations in retrieving the relevant information from chat history and integrating it into the reasoning process act as a barrier for models to reach higher performance. These memory limitations extend to the agent-based systems that rely on MLLMs for tracking chat history, reasoning across multiple steps, and integrating past information. In agent-based settings, memory is utilized continuously across multiple interaction steps, and memory-related errors can accumulate over time, which leads to cascading failures and incorrect reasoning processes.

6 Conclusion

We introduce SciReC, a model-adaptive benchmark for multimodal academic dialogue, designed to assess the relational reasoning ability of MLLMs in domain-specific concepts. We include upstream tasks covering memory, knowledge, and visual reasoning, and introduce the DMRA framework, which quantitatively analyzes the main failures of relational questions by a two-stage deficit-based approach. Performance trends across models indicate that structural relations achieve the highest scores among proprietary models (78-82%), whereas sequential (64%) and hierarchical relations(55-70%) show the lowest. In contrast, most open-source models struggle with spatial relations (33-41%). Model performance varies across domains, with Astronomy being the most challenging and Psychology yielding the highest performance. Among incorrect responses, relational reasoning is the primary reason, ranging from 50% to 71%, followed by memory as the secondary failure (37–46%). Knowledge and visual gaps are less prominent in proprietary models but remain major error sources for most open-source models. SciReC highlights relational reasoning limitations in MLLMs, and DMRA identifies the underlying reasons; together, they aim to improve model performance.

AI USE STATEMENT

In this work, we used generative AI tools (LLMs) to generate synthetic data sets, help to develop the theoretical framework, formulate mathematical claims, and implement methods. We have not used generative AI tools to interpret results, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, and support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, brainstorming, sourcing/searching for information, editing a research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. A subset of the LLM-generated synthetic data was manually reviewed for consistency with the defined generation rules and source materials, although some generation errors may remain. LLM-assisted theoretical framework and mathematical formulations were reviewed by one author for conceptual and mathematical correctness. LLM-generated code was tested in the experimental pipeline and validated against expected outputs. Information and literature identified with LLM assistance were manually verified against the original sources. All LLM-assisted manuscript edits were reviewed by the authors to ensure accuracy and consistency with the intended meaning. Brainstormed ideas were critically assessed by the authors before inclusion in the work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research poses no ethical issues because it utilizes open-source textbooks and pre-existing, publicly accessible models without requiring any subjective assessments. This work does not involve collecting personal or sensitive information. The study adheres fully to the ethical standards established by the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The prompt instructions, a detailed explanation of the dataset generation pipeline, the calculation of error decomposition, and an example calculation case are provided in Section 3, Section 4, and Appendix C. The study's execution can be facilitated by referring to the details in Appendix E. Additionally, we plan to make the data and code freely available under an open-access license, along with detailed instructions that allow the reproducibility of the experiment results, upon acceptance of the paper. The research strictly follows the ICLR Reproducibility Requirements.

REFERENCES

Anthropic. Claude 4.6 (opus & sonnet). https://www.anthropic.com/news/claude-sonnet-4-6, 2026. Proprietary model.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents, 2026. URL https://arxiv.org/abs/2601.03515.

Yonatan Bitton, Ron Yosef, Eli Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. Vasr: Visual analogies of situation recognition, 2022. URL https://arxiv.org/abs/2212.04542.

Raymond Bernard Cattell. *Intelligence: Its structure, growth and action*, volume 35. Elsevier, 1987.

Mary Ann Clark, Matthew Douglas, and Jung Choi. *Biology 2e*. OpenStax, Houston, Texas, March 2018. URL https://openstax.org/books/biology-2e/pages/1-introduction.

John Duncan. Intelligence tests predict brain response to demanding task events. *Nature Neuroscience*, 6(3):207–208, 2003.

Paul Flowers, Klaus Theopold, Richard Langley, and William R. Robinson. *Chemistry 2e*. OpenStax, Houston, Texas, February 2019. URL https://openstax.org/books/chemistry-2e/pages/1-introduction.

Andrew Fraknoi, David Morrison, and Sidney Wolff. *Astronomy 2e*. OpenStax, Houston, Texas, March 2022. URL https://openstax.org/books/astronomy-2e/pages/1-introduction. Accessed: 2026-05-05.

Micah B. Goldwater and Lisa Schalk. Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, 142(7):729–757, 2016. doi: 10.1037/bul0000043.

Steven A. Greenlaw, David Shapiro, and Daniel MacDonald. *Principles of Economics 3e*. OpenStax, Houston, Texas, December 2022. URL https://openstax.org/books/principles-economics-3e/pages/1-introduction.

Graeme S Halford, William H Wilson, and Steven Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences*, 21(6):803–831, 1998.

Keith J. Holyoak and Paul Thagard. *Mental leaps: Analogy in creative thought*. MIT Press, Cambridge, MA, 1995.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning, 2024. URL https://arxiv.org/abs/2405.01483.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Elizabeth D. Kirby, Melissa J. Glenn, Noah J. Sandstrom, and Christina L. Williams. *Introduction to Behavioral Neuroscience*. OpenStax, Houston, Texas, November 2024. URL https://openstax.org/books/introduction-behavioral-neuroscience/pages/1-introduction.

Young-Jun Lee, Byung-Kwan Lee, Jianshu Zhang, Yechan Hwang, Byungsoo Ko, Han-Gyu Kim, Dongyu Yao, Xuankun Rong, Eojin Joo, Seung-Ho Han, Bowon Ko, and Ho-Jin Choi. Multiverse: A multi-turn conversation benchmark for evaluating large vision and language models, 2025. URL https://arxiv.org/abs/2510.16641.

Samuel J. Ling, William Moebs, and Jeff Sanny. *University Physics Volume 2*. OpenStax, Houston, Texas, October 2016. URL https://openstax.org/books/university-physics-volume-2/pages/1-introduction.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas

Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models, 2024a. URL https://arxiv.org/abs/2403.20194.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms, 2024b. URL https://arxiv.org/abs/2406.11833.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents, 2024. URL https://arxiv.org/abs/2402.17753.

John McMurry. *Organic Chemistry*. OpenStax, Rice University, Houston, Texas, September 2023. URL https://openstax.org/books/organic-chemistry/pages/1-why-this-chapter.

William Moebs, Samuel J. Ling, and Jeff Sanny. *University Physics Volume 1*. OpenStax, Houston, Texas, September 2016. URL https://openstax.org/books/university-physics-volume-1/pages/1-introduction.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yun Xing, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. Mmrel: Benchmarking relation understanding in multi-modal large language models, 2025. URL https://arxiv.org/abs/2406.09121.

David L. Oden, Roger K. R. Thompson, and David Premack. Can an ape reason analogically? comprehension and production of analogical problems by sarah, a chimpanzee (pan troglodytes). In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov (eds.), *The analogical mind: Perspectives from cognitive science*, pp. 471–497. MIT Press, Cambridge, MA, 2001.

OpenAI. Gpt-5.4. https://platform.openai.com/docs/models/gpt-5.4, 2026. Large language model.

Charles S Peirce. *Philosophical writings of Peirce*. Courier Corporation, 2012.

Rose M. Spielman, William J. Jenkins, and Marilyn D. Lovett. *Psychology 2e*. OpenStax, Houston, Texas, April 2020. URL https://openstax.org/books/psychology-2e/pages/1-introduction.

Gilbert Strang and Edwin "Jed" Herman. *Calculus Volume 1*. OpenStax, Houston, Texas, March 2016a. URL https://openstax.org/books/calculus-volume-1/pages/1-introduction.

Gilbert Strang and Edwin "Jed" Herman. *Calculus Volume 2*. OpenStax, Houston, Texas, March 2016b. URL https://openstax.org/books/calculus-volume-2/pages/1-introduction.

Gilbert Strang and Edwin "Jed" Herman. *Calculus Volume 3*. OpenStax, Houston, Texas, March 2016c. URL https://openstax.org/books/calculus-volume-3/pages/1-introduction.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman,

Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding, 2024. URL https://arxiv.org/abs/2406.09411.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He, and Yu Qiao. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation, 2025. URL https://arxiv.org/abs/2502.11903.

Dawei Yan, Yang Li, Qing-Guo Chen, Weihua Luo, Peng Wang, Haokui Zhang, and Chunhua Shen. Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning, 2025. URL https://arxiv.org/abs/2503.18533.

Nilay Yilmaz, Maitreya Patel, Lawrence Luo, Tejas Gokhale, Chitta Baral, Suren Jayasuriya, and Yezhou Yang. Voila: Evaluation of mllms for perceptual understanding and analogical reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 69763–69788, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms

via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning, 2019. URL https://arxiv.org/abs/1903.02741.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Dataset Details

In total, SciReC encompasses 656 relations across eight distinct subject categories. These relations are handled in 189 conversations for these subjects. The number of turns in each dialogue is flexible, as the conversational flow depends on the model's performance. The minimum number of turns is three in the dataset, which consists of caption questions and one relational question.

The number of conversations is directly related to the number of chapters in textbooks. As relations are extracted from chapters, the connections between figures and topics are preserved in one conversation. The distribution of dialogues across subjects is provided in Figure 6. The SciReC consists of 8 scientific and social science domains. While Biology covers the most conversations in the dataset with 43 (22.6%), the number of dialogues in Psychology is the lowest with 5 (3.2%), followed by Calculus with 13. The combination of Organic Chemistry and Chemistry subjects encompasses 36 conversations. The count of dialogues in the Physics and Astronomy domains is similar, 29 and 26, respectively. The subjects on Behavioral Neuroscience and Economics constitute around 10% of the total dialogues, with similar numbers, 19 and 17.

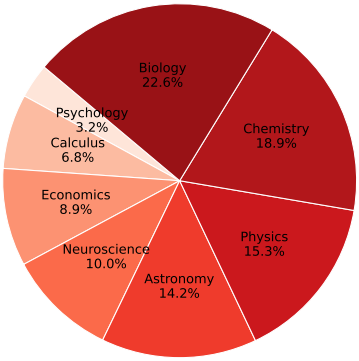


Figure 6: Distribution of dialogues

The dataset creation pipeline of the SciReC is illustrated in Figure 7 and explained in Section 3. For evaluating the open-source models, we utilized an A100 GPU. The time of execution for each model takes approximately 15-18 hours, with a 4096 model context length.

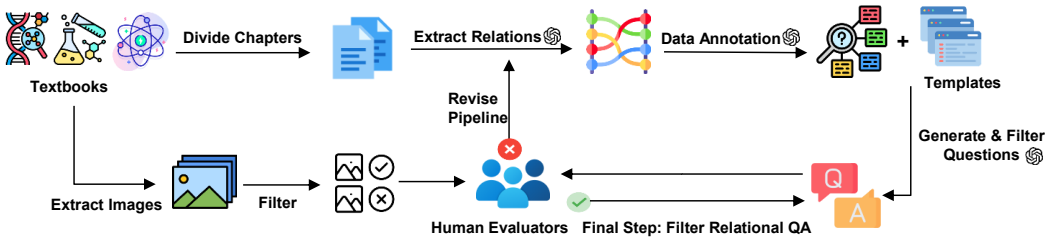


Figure 7: The benchmark creation pipeline, which accepts textbooks as input and processes them by extracting chapters and images. After extracting the relations and ground truths, questions are generated: relational and visual with templates, visual, memory, and memory validation without templates. A curated collection of filtered Q&A pairs is evaluated in multiple rounds by human raters for three criteria (relational category match, accuracy, and visual grounding) until the pairs satisfy the established quality criteria.

Data License. The data in SciReC is derived from OpenStax textbooks (CC BY-NC-SA 4.0). We provide proper attribution to the original authors and release our dataset under the same license.

Societal Impacts. This paper presents a relational reasoning dataset in scientific and social science concepts, designed to assess the model's relational reasoning ability along with diagnostic analysis. Potential positive societal consequences include the improvement of more reliable AI tools in education, providing transparency on models' performance with reasoning failures, and providing a road map for models based on comprehensive diagnostic results. While efficient, integration into automated pipelines like tutoring requires human oversight, as relying solely on automation may lead to misleading information.

B Evaluation Model Selection

To select the scoring model aligned with the human assessment, we evaluated Claude 4.6 and GPT 5.4 with the same subset of Physics questions. By including the constraints in the prompt, we optimize the scoring for both models. As illustrated in Figure 8, we evaluate the Qwen-3.5 model responses for conversational flow and score their results online with both models. The question numbers for each question type varies, as the flow is conditionally dependent on the model's performance. In total, 199 model responses are evaluated: 82 knowledge-based, 63 visual, 22 relational and memory-related, and 10 memory validation. As seen in the patterns in Table 2, **most questions resulted in consensus between the two candidate judges, with agreement exceeding 90% for knowledge and visual questions and reaching 81.8% for memory questions, while relational questions exhibited substantially more disagreement, reflecting their greater complexity for scoring.** Figure 8 shows that for memory and validation questions, GPT-5.4 avoids using the 4-6 score range in contrast to Claude 4.6. As illustrated in relational questions, GPT-5.4 tends to give higher scores, mostly above the threshold. While the average scoring difference between models in knowledge, memory validation, and memory questions is around 1 point, for relational questions, it is 1.68.

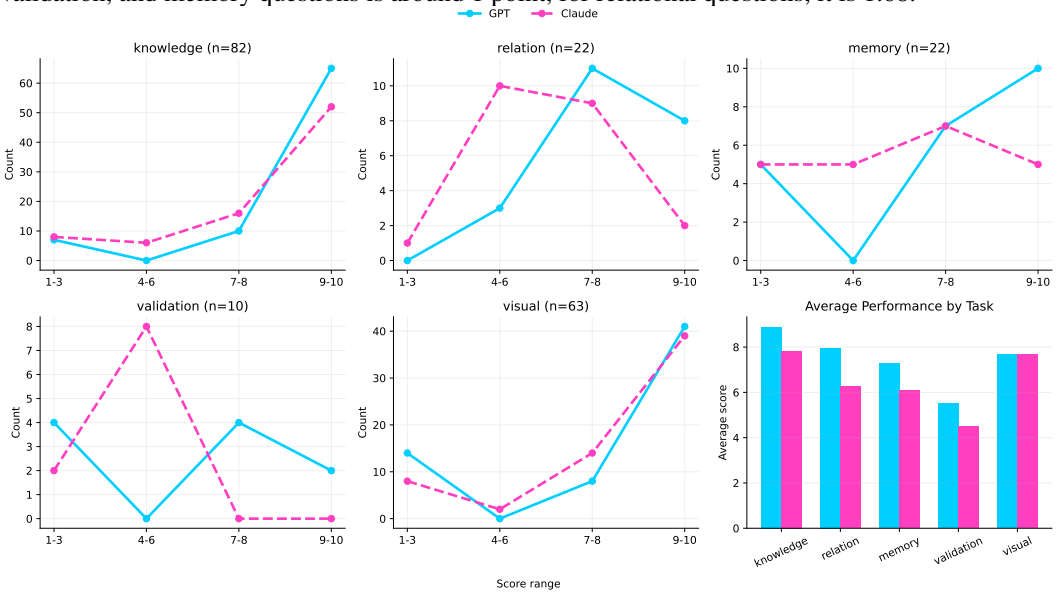


Figure 8: Distribution of evaluation scores for GPT and Claude across different task types. Line plots illustrate the number of records within each score range, and the bar chart summarizes average performance across tasks.

After observing these score differences, we selected 29 conflicting cases (13 relational, 8 knowledge-based, 4 visual, and 4 memory retention questions) from the same set of questions with all types where one model scores above the threshold while the other scores below it. These cases represent the only instances in which the choice of judge could alter the subsequent evaluation path. The remaining 85.4% of the sampled responses received consistent threshold decisions from both judges and therefore would have resulted in identical dialogue trajectories regardless of which judge was used. Consequently, conflict cases constitute the most informative subset for judging selection, as

Table 2: Agreement analysis across question categories. Total Questions denotes the number of evaluated questions in each category. Consensus reports cases where Claude and GPT produced the same evaluation, while Conflict Cases indicates disagreements between the two models. For conflict cases, Claude Agreement and GPT Agreement report the number and percentage of conflict cases in which human evaluation agreed with each model's evaluation.

Category	Total Questions	Consensus	Conflict Cases	Claude Agreement	GPT Agreement
Knowledge	82	74 (**90.2%**)	8 (9.8%)	5/8 (62.5%)	3/8 (37.5%)
Relational	22	9 (40.9%)	13 (59.1%)	12/13 (**92.3%**)	1/13 (7.7%)
Memory	22	18 (81.8%)	4 (18.2%)	4/4 (**100%**)	0/4 (0%)
Visual	63	59 (93.7%)	4 (6.3%)	4/4 (**100%**)	0/4 (0%)
Memory Validation	10	10 (**100%**)	0 (0%)	-	-

they directly determine the adaptive evaluation. The aim is to select the most human-aligned scoring model for evaluating the model responses. We request human annotators to determine which score is more accurate by reviewing the question, ground truth, images, and two different scores, along with their justifications. As shown in Table 2, human experts agreed with Claude 4.6 in 100% of the visual and memory conflict cases, 92.3% of the relational conflict cases, and 62.5% of the knowledge conflict cases, leading us to select Claude 4.6 as the judge. The expert feedback also shows that for relational questions, GPT-5.4 provides higher scores despite the missing contextual details in the answer, although it is constrained in the prompt. **Claude 4.6 not just analyzes the correctness of the relationship but also evaluates the concept integration required to explain the underlying relation between them.** Based on the human review, we selected Claude 4.6 as the scoring model, which shows better alignment with human scoring.

C Example of DMRA Calculations

This section illustrates the steps involved in performing DMRA calculations on the given Figure 2. The higher-resolution figures in the relational question is provided in Figure 9. The illustrated part of the dialog for one relational reasoning consists of one relational, two memory-based, one memory validation, three knowledge-based, and three visual questions for each figure. The threshold for each type of question is 7, and the scores below 7 create a deficit. The score of the relational reasoning question is 2.0, and the average scores of other question types are provided in Table 3. As the memory score of Figure 1 is above the threshold, a memory validation question is not required.

Table 3: Scores by question type for each figure (out of 10)

	Memory	Memory Validation	Knowledge	Visual
Figure 1	9.0	N/A	6.333	5.666
Figure 2	2.0	3.0	2.333	5.666

The computation of the DMRA framework consists of multiple steps:

1) As the first step, the deficit scores for each task in each figure are computed.

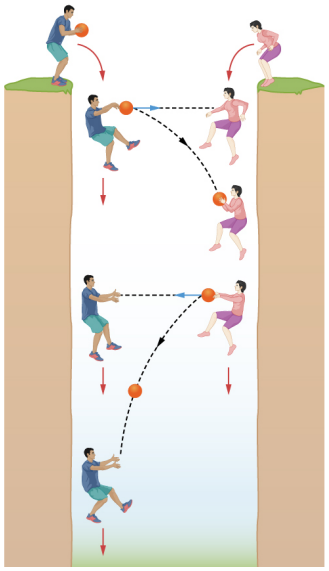
Figure 1 deficits:

- memory: max(0, 7 - 9.0) = 0
- knowledge: max(0, 7 - 6.333) = 0.666
- visual: max(0, 7 - 5.663) = 1.333

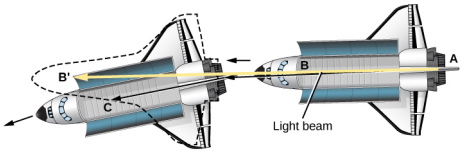
Total deficit for Figure 1 = 2

Figure 2 deficits:

- memory: max(0, 7 - 2.0) = 5
- knowledge: max(0, 7 - 2.333) = 4.666
- visual: max(0, 7 - 5.666) = 1.333



(a) Figure 1



(b) Figure 2

Figure 9: High resolution of visual images shown in Figure 2

Total deficit for Figure 2 = 11

2) The default weights for all tasks are 0.33. However, each task contributes differently to the result. Therefore, we adjust each task's weight for each figure based on its deficit scores using softmax.

For Figure 1: There is no validation gap since its memory score is above the threshold. Therefore, the final weights after the softmax process are the same for all three tasks: 0.33.

- memory weight = 0.333
- knowledge weight = 0.333
- visual weight = 0.33

For Figure 2: Validation gap = 3.0 - 2.0 = 1.0 which changes memory logit to $0.33 + 0.2 \times 1.0 = 0.53$. The logits of knowledge and visual remain the same, 0.33, as they influence the result equally, in contrast to memory. After applying softmax to these logits, we reach the adjusted weight for each task:

- memory weight = 0.3792
- knowledge weight = 0.3104
- visual weight = 0.3104

3) As the deficit of each figure is not the same amount, we calculate the figure weights with $\lambda = 0.5$, which prevents ignoring the contribution of other figures. The results of the calculation show that the deficit of Figure 2 is dominant, with 82.1% of the total upstream deficits.

Figure 1 weight = (2 + 0.5) / [(2 + 0.5) + (11 + 0.5)] = 0.178

Figure 2 weight = (11 + 0.5) / 14 = 0.821

4) After computing the figure and task weights, the next step is to calculate the total deficit contribution of each figure and all upstream tasks.

Figure 1 total deficits:

- memory: 0.178 * 0.333 * 0 = 0
- knowledge: 0.178 * 0.333 * 0.666 = 0.039
- visual: 0.178 * 0.333 * 1.333 = 0.079

Total : 0 + 0.039 + 0.079 = 0.119

Figure 2 total deficits:

- memory: 0.821 * 0.379 * 5 = 1.557
- knowledge: 0.821 * 0.31 * 4.666 = 1.19
- visual: 0.821 * 0.31 * 1.333 = 0.34

Total : 1.557 + 1.190 + 0.34 = 3.087

Upstream deficits:

- memory: 0 + 1.557 = 1.557
- knowledge: 0.039 + 1.19 = 1.229
- visual: 0.079 + 0.34 = 0.419

Total: 1.557 + 1.229 + 0.419 = 3.206

5) Total upstream deficit represents the explained error score by knowledge gap, limited visual understanding, and memory issues. To find the unexplained, rational part, we need to use the score (2.0) from the relational reasoning question.

Total failure = 7 - 2 = 5

Upstream total = min (5 , 3.206) = 3.206 Relational = 5 - 3.206 = 1.793

6) As the upstream total is equal to the upstream deficits, they do not require normalization.

- Figure 1 Memory = 0
- Figure 1 Knowledge = 0.039
- Figure 1 Visual = 0.079
- Figure 2 Memory = 1.557
- Figure 2 Knowledge = 1.19
- Figure 2 Visual = 0.34
- Relational = 1.793

By dividing each deficit contribution to the total failure, we reach the final percentages of each factor.

- Figure 1 Memory = (0 / 5) * 100 = 0%
- Figure 1 Knowledge = (0.039 / 5) * 100 = 0.8%
- Figure 1 Visual = (0.079 / 5) * 100 = 1.6%
- Figure 2 Memory = (1.557 / 5) * 100 = 31.14%
- Figure 2 Knowledge = (1.19 /5) * 100 = 23.8%
- Figure 2 Visual = (0.34 / 5) * 100 = 6.8%
- Relational = (1.793 / 5) * 100 = 35.9%

The results show that most of the failures were caused by relational reasoning, followed by Figure 2 memory and Figure 2 knowledge. The tasks in Figure 1 have minimal impact on the results due to their low deficits and low figure weight.

D EXAMPLES OF RELATIONAL QUESTIONS

The SciReC varies in terms of relation types and scientific and social subjects. This diversity contributes to a richer contextual representation within the dataset. The combination of relations and knowledge domains provides a comprehensive and diverse evaluation setting. Some examples from the SciReC dataset are shown in Figures 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19.

Domain: Psychology
Relation Type: Hierarchical

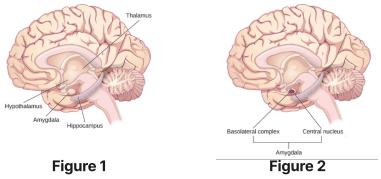


Figure 1 **Figure 2**

Question:

How does Figure 1 provide a top-level view that is refined by Figure 2?

Ground Truth:

Figure 1 shows the limbic system, which includes the amygdala, and Figure 2 breaks down the amygdala into its subnuclei, including the basolateral complex and central nucleus.

Figure 10: Example of a hierarchical relational question in the Psychology domain

Domain: Physics
Relation Type: Other

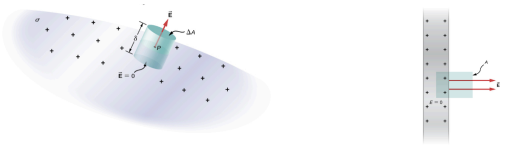


Figure 1 **Figure 2**

Question:

How does the structure in Figure 1 support what is described in Figure 2?

Ground Truth:

Figure 1 and Figure 2 are linked by the use of Gauss's law for conducting surfaces: Figure 1 derives the surface field relation E = σ/ε0 with a Gaussian cylinder at a conductor's surface, and Figure 2 applies that same relation to find the electric field outside an infinite conducting plate with surface charge density σ on each face.

Figure 11: Example of "Other" type of relational question in the Physics domain

Domain: Economy
Relation Type: Comparative



Figure 1

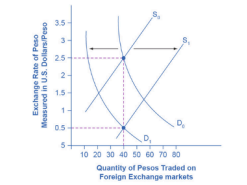


Figure 2

Question:

What key contrast in the economic drivers of exchange-rate shifts can be identified between Figure 1 and Figure 2?

Ground Truth:

Figure 1 and Figure 2 comparatively show how exchange-rate shifts arise from different economic drivers: in Figure 1, higher relative interest rates increase demand and reduce supply for a currency, causing appreciation, whereas in Figure 2, high inflation in Mexico decreases demand and increases supply for the peso, causing depreciation.

Figure 12: Example of a comparative relational question in the Economics domain

Domain: Astronomy
Relation Type: Cause-Effect

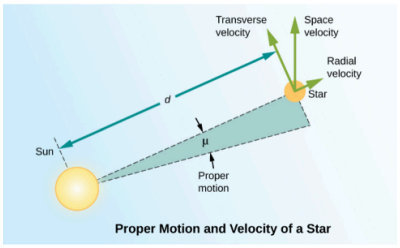


Figure 1

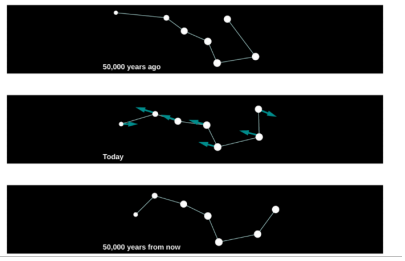


Figure 2

Question:

How does Figure 2 explain the effect shown in Figure 1?

Ground Truth:

Figure 2 explains that a star's space velocity determines how its position and distance from the Sun change over time, and Figure 1 shows this effect as the Big Dipper's handle appears more bent 50,000 years from now.

Figure 13: Example of a cause and effect relational question in the Astronomy domain

Domain: Calculus
Relation Type: Comparative

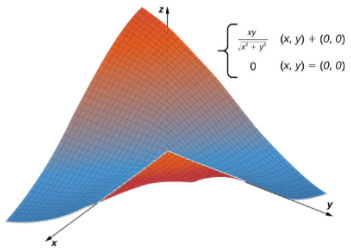


Figure 1

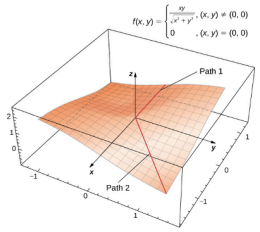


Figure 2

Question:

What does Figure 1emphasize that Figure 2 does not, and vice versa, for the function f(x, y)=xy/√(x^2+y^2) at the origin?

Ground Truth:

Figure 1 and Figure 2 both depict the same function f(x, y)=xy/√(x^2+y^2) with emphasis on its failure at the origin: Figure 1 shows that it does not have a tangent plane there, while Figure 2 shows that it is not differentiable at the origin.

Figure 14: Example of a comparative relational question in the Calculus domain

Domain: Organic Chemistry
Relation Type: Other

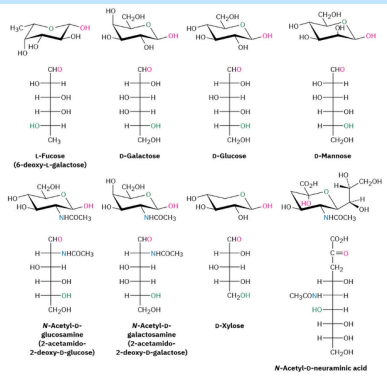


Figure 1

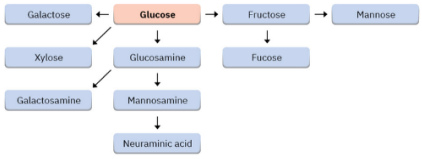


Figure 2

Question:

How do Figure 1 and Figure 2 relate?

Ground Truth:

FIGURE 1 identifies the eight essential human monosaccharides, and FIGURE 2 summarizes how all of these essential monosaccharides arise from glucose by conversion.

Figure 15: Example of a ”Other” type of relational question in the Organic Chemistry domain

Domain: Behavioral Neuroscience
Relation Type: Spatial-contextual

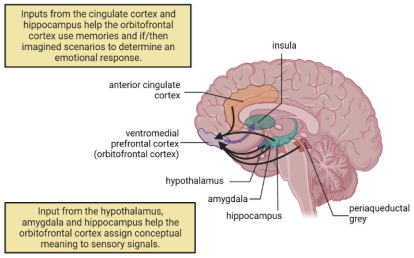


Figure 1

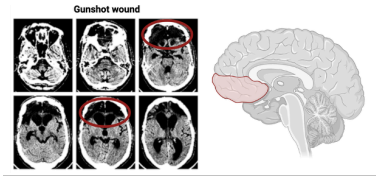


Figure 2

Question:

How does Figure 2 provide spatial detail that complements the region identified in Figure 1?

Ground Truth:

Figure 2 shows the location of the gunshot wound within the ventromedial prefrontal cortex identified in Figure 1 as the orbitofrontal region just above the eye sockets.

Figure 16: Example of a spatial-contextual relational question in the Behavioral Neuroscience domain

Domain: Biology
Relation Type: Structural-functional

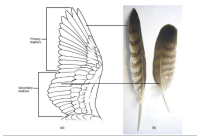


Figure 1

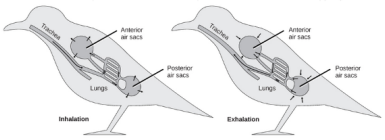


Figure 2

Question:

How does the structure in Figure 1 support what is described in Figure 2?

Ground Truth:

Figure 1 and Figure 2 are linked functionally because Figure 1 concerns bird flight feathers on the wings, while Figure 2 shows the unidirectional airflow in bird lungs that is adapted to support the high metabolic rate required for flight.

Figure 17: Example of a structural-functional relational question in the Biology domain

Domain: Physics
Relation Type: Cause-effect

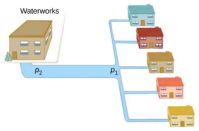


Figure 1

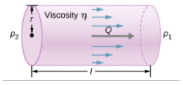


Figure 2

Question:

How does Figure 2 use system shown in Figure 1 to explain water pressure difference in heavy use?

Ground Truth:

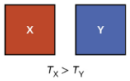
Figure 1 applies the laminar-flow equation, Q = (p2 − p1)/R, shown in Figure 2 to explain that increased flow through the water main lowers the home pressure p1 because flow rate depends on the pressure difference and tube resistance.

Figure 18: Example of a cause and effect relational question in the Physics domain

Domain: Chemistry
Relation Type: Analogical



Figure 1



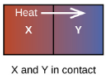


Figure 2

Question:

What shared principle connects Figure 1 and Figure 2?

Ground Truth:

Figure 1 and Figure 2 are analogous because both depict spontaneous processes driven by a tendency toward more uniform dispersal, with Figure 1 showing matter spreading as a gas expands into both flasks and Figure 2 showing energy spreading as heat flows from object X to object Y.

Figure 19: Example of an analogical relational question in the Chemistry domain

E PROMPT AND TEMPLATES

The prompts used in the study are provided in Figures 20, 21, 22, 23, 24 , 25, 26, 27 , 28 , 29. The templates utilized to generate visual and relational questions are provided in Figures 30, 31, 32, and 33.

Finding Relations

```
You are given a textbook containing various images numbered, each accompanied by
descriptive text.

Your task is to:
- Identify meaningful relations between pairs of distinct main images in the textbook.
- Each pair MUST contain two distinct main image names (figure1_name ≠ figure2_name).
- The relationship MUST be derived from Figure 1 and Figure 2 as a whole, rather than
from sub-images within a single figure.
- The relations MUST be determined based on the combined meaning of the evidence from
both figures.
- Provide textual evidence that supports the relationship. This may come from the same
sentence, or nearby sentences/paragraphs that clearly establish the connection.
- Provide the source text of the relation, extracted from the textbook. The source text
MUST explicitly include BOTH figure identifiers and information.
- For each extracted relation, provide a short justification must directly quote or
closely paraphrase the exact part of the text that establishes the relationship.
- If the connection is unclear, implicit, or requires interpretation, discard the pair.
- Systematically evaluate all possible different figure pairs, but include ONLY those
with explicit textual relationships. If no such relationship exists, do NOT include the
pair.
- Double-check all figure pairs and ensure no valid explicit relation is missed.
```

Figure 20: Prompt for extracting the relations from textbooks.

Creating Ground Truths and Categorizing Relations

```
You are creating a ground-truth relational statement between two figures based on a
source_text.

Step 1: Feasibility Check
Determine whether the source_text provides enough explicit evidence of a meaningful
relationship between the two figures.

A valid relationship must:
- Be explicitly supported or strongly implied by the source_text
- Involve both figures in a shared concept, process, or comparison
If NO clear, evidence-supported relationship exists, return null and do NOT attempt to
construct a relationship

Step 2:  Generate Ground Truth
If a valid relationship exists, rewrite the source_text into a single, clear ground_truth
statement that explicitly links the two figures.

Requirements:
- Must explicitly mention BOTH figure identifiers exactly as they appear in the
source_text. Otherwise, return null.
- Must contain ONE clear relational clause connecting the two figures
- Must be grounded in the source_text, but may paraphrase for clarity and fluency. Do NOT
mention or refer to "source_text", "the text", "the passage", or any meta-source.
- Must describe a specific, content-rich relationship, including: key concepts,
mechanisms, or variables

Relational Categories (Select ONE):
Choose the category based on meaning (NOT wording). Prefer the categories below when
applicable.

* Spatial-contextual : If the relationship is based on physical location, position, or
spatial arrangement
* Sequential : If Figure 1 and Figure 2 represent different steps in a process over time.
MUST reflect order between figures, not within a single figure
* Cause-effect : If one figure shows a process, mechanism, or condition that directly
produces the outcome shown in the other figure. Does one figure produce the outcome in
the other? If yes → cause-effect.
* Structural-functional : If one figure identifies the physical arrangement or internal
components of a specific system, while the other figure illustrates the operational
processes or regulatory mechanisms that those same components perform. Do both refer to
the same system, with one explaining how it works? If yes → structural-functional
* Hierarchical : If one figure shows a whole system and the other breaks down a part of
it.
* Analogical : If both figures show two different systems or contexts but same underlying
principle/mechanism. “this works like that”
* Conditional : If the relationship is explicitly conditional: “if X happens → then Y
occurs”.
```

Figure 21: Prompt for relationship validation, ground truth creation, and category selection for relations.

Creating Ground Truths and Categorizing Relations (Continued)

```
* Comparative : If both figures depict the same underlying concept, process, or
mechanism, but differ in context, system, scale, or conditions. Do both figures represent
the same concept or process, but under different contexts, systems, conditions, or levels
of detail? If yes → comparative.
* Other: Use ONLY if none of the above categories accurately capture the relationship.
This includes cases such as contextual linkage without clear category.

IMPORTANT:
- Do NOT force a relationship into one of the predefined categories if it does not fit
well.
- If selecting "Other", ensure the relationship is still clearly defined and specific
(not vague or generic).

Relational Expression Rules:
- Use a clear relational phrase appropriate to the category.
- The relationship must be logically valid and text-supported.
- Do NOT force a category if evidence is weak → return null instead.

Strict Constraints:
- Do NOT introduce new concepts not grounded in source_text
- Do NOT describe figures independently without linking them
- Do NOT rely on figure numbering alone as evidence of relationship

Final Validation:
Before output, ensure:
- Both figures are explicitly mentioned in ground_truth
- The relationship is meaningful and not superficial
- The category matches the actual semantic relationship
- The statement is natural, clear, and logically correct
If any condition fails → return null and do NOT attempt to construct a ground_truth


DATA:
source_text: {source_text}

OUTPUT FORMAT:
- Output ONLY a JSON object and it MUST contain EXACTLY "ground_truth".
- No extra keys. No extra text before or after the JSON array.
```

Figure 22: Prompt for relationship validation, ground truth creation, and category selection for relations (continued).

Visual Question Generation

```
You are a question-generation assistant for a vision+text dataset.

You are given:
1) a chapter from a textbook.
2) A QUESTION TEMPLATE SET (below)

Your task:
For the given chapter, generate strictly visual grounded questions related only to the
figures and sub-figures in the chapter.
- Questions must be strictly follow the template by replacing its placeholders, without
introducing new question styles.
- Identify the figure ID referenced in the chapter and include the figure ID in every
question but not mention about the caption.
- Questions must NOT require domain knowledge, interpretation, multi-step reasoning,
inference or conceptual understanding.
- For each question, extract the ground truth answer derived directly from the text (give
page number as source) related to the figure.
- For each question, assign a category corresponding to the template title.
- Focus only on figure-related content and avoid generic questions that can be answered
without the figure.
- Do NOT hallucinate or infer any information not explicitly written.
- Do NOT include any phrases at the questions that refer to explanations, discussions,
caption, description or information outside the figure (e.g., 'according to', 'based on',
'as described', 'from the discussion', etc.).
- Simple visual-text mappings explicitly stated in the text (e.g., a visual features) are
allowed and do not count as inference. Do not map to abstract concepts, functions, real-
world applications, relationships, or meanings.
- For graph-based figures, identifying explicitly stated increases, decreases, or changes
is allowed and does not count as interpretation.
- For every figure, MUST generate at least one question. If necessary, use simple visual
identification and direct visual-text association questions (e.g., labels, elements,
positions but NOT sub-figure labels like a, b, C) even if information is minimal.
- For each figure, attempt to generate at least three independent vision-based questions
using different templates. Only generate fewer than three if the figure genuinely lacks
enough distinct visual information.
- If no template fits perfectly, choose the closest matching template and adapt it
conservatively without introducing new question styles and asking knowledge-based
questions.

PLACEHOLDER MAPPING:
When templates include placeholders like {Figure X}, {concept}, {attribute}, {process},
etc.:
- Only fill them with words/phrases that appear in text.

<VISUAL  QUESTION TEMPLATE>
```

Figure 23: Prompt for creating visual questions with template

Knowledge-Based Question Generation

```
You are given a chapter from a textbook. Your task is to analyze the content and extract
educational Q&A pairs that are specifically related to figures/sub-figures within
chapters. For each figure:
1 - Identify the figure ID as referenced in the chapter, and include the figure ID in
every question but not mention about the caption or text.
2- Analyze the surrounding paragraphs, any explanatory text that clarifies the figure's
purpose, data, or concept.
3-  Focus only on figure-related content and avoid generic questions that can be answered
without the figure.
4-  Ensure the questions test understanding of the visual and contextual information
provided by the figure.
5-  Generate questions which must require conceptual understanding of the process,
mechanism, function, or significance illustrated in the figure.
6 - For figures that include numerical values, formulas, or worked examples, it is
allowed to generate calculation-based questions where all required values are explicitly
given in the text.
7-  MUST generate the first question as a general knowledge/comprehension question about
what the figure represents, ensuring it tests understanding of the main idea illustrated
in the figure. The answer must be derived directly from the caption or associated text.
Form a natural educational question without giving too much details. (e.g., "What process
is illustrated in Figure 3.2?")
8-  For each question, extract the ground truth answer derived directly from the text
(give page number as source). The ground truths MUST be complete, grammatically correct
sentences, not fragments or partial phrases.
9-  Questions MUST avoid giving unnecessary descriptive or explanatory details.
10- Do NOT include any phrases at the questions that refer to explanations, discussions,
caption, description or information outside the figure (e.g., 'according to', 'based on',
'as described', 'from the discussion', etc.).
11-    Do NOT generate counting, labeling, location-based, or purely visual questions.
12- Do NOT hallucinate details not present in the caption or text.
13-    For every figure, MUST generate at least two questions and MUST attempt to
generate at least three independent knowledge-based questions. Only generate two if the
text genuinely lacks enough information.
```

Figure 24: Prompt for creating knowledge-based questions

Relational Question Generation

```
You are a question-generation assistant for a vision+text dataset.
You will be given:
1) A QUESTION TEMPLATE SET (below),where example questions are grouped by relation_type.
2) One RELATIONAL JSON object contains: figure1_name, figure2_name, relation_type AND
ground_truth

Your task:
Step 1: For EACH given relational JSON object, generate ONE question that matches its
relation_type and is grounded in the ground_truth.
Guidelines:
- Use the corresponding template as a base, but you may slightly revise it to better
align with the ground_truth.
- The question must accurately reflect the ground_truth, even if minor template
adjustments are needed.
- The question MUST explicitly include BOTH figure names exactly as they appear in
ground_truth. Both {{F1}} and {{F2}} placeholders MUST be filled with their corresponding
figure names.
- If relation_type conflicts with the relationship expressed in ground_truth, IGNORE
relation_type and follow ground_truth for question generation. In that case reselect a
new relation_type from the list <Comparative | Spatial-contextual | Sequential | Cause-
effect | Structural-functional | Hierarchical | Analogical | Conditional | Other>
- Keep the question concise and natural and avoid unnecessary elaboration, examples, or
explanations in the question. Avoid answer leakage!

Constraints:
- The question must preserve the SAME relationship type as ground_truth.
- Do NOT use a template if it forces the wrong relationship. Templates are optional and
must not override ground_truth.
- Do NOT introduce new facts, entities, attributes, relationships or explanations that
are not present in ground_truth.
- Avoid unnecessary elaboration, examples, or explanations in the question.
- The question must be answerable using only the given ground truth and do not rely on
external facts, assumptions, or prior knowledge.

Step 2: After finalizing the candidate question, verify that every required concept,
entity, and relation in the question is explicitly present in ground_truth. Verification
steps must be performed internally.
- If the question cannot be fully and explicitly answered using only ground_truth, output
exactly: {{ "question": null }}.
- If the question does not include BOTH figure names exactly as they appear in
ground_truth, output exactly: {{ "question": null }}.
- If the question passes verification, output exactly:  {{ "question": "<the concise
question string>" }} and nothing else.

Step 3: After validating and finalizing relational questions, MUST create two independent
memory questions — one for Figure1 one for Figure 2 — based on the exact facts needed to
answer the relational question.
- Each question must refer to only one figure ID in the question and do not mention the
other figure.
```

Figure 25: Prompt for creating relational reasoning questions with template.

Relational Question Generation (Continued)

```
- Ask for the specific fact, visual cue, or piece of evidence from that figure that is
necessary to answer the relational question. Do NOT ask for the relation itself.
- Questions must be concise. Do not include explanatory phrases or extended descriptions.

- The memory questions must be answerable using only the given ground truth do not rely
on external facts, assumptions, or prior knowledge.

If you cannot generate two valid memory questions:
- Re-extract one atomic fact for each figure from the ground_truth
- Then regenerate the memory questions

Step 4: After creating memory questions, MUST create two additional questions as
validation versions of those memory questions.
- Each validation question must include the figure ID, target exactly the same
information and have the same correct answer as its corresponding memory question.
- The validation question MUST differ only in surface form (e.g., slight rewording, word
order changes, or minor phrasing adjustments).
- Do not add, remove, or replace any details. Not introduce new descriptive details,
explanations, qualifiers, or constraints.

OUTPUT FORMAT:
- Output ONLY a JSON object and it MUST contain EXACTLY "question".
- No extra keys. No extra text before or after the JSON array.

DATA:
relation_type: {relation_type}
ground_truth: {ground_truth}
figure1_name: {figure1_name}
figure2_name: {figure2_name}

PLACEHOLDER MAPPING:
- Use figure names dynamically:
  {{F1}} = figure1_name
  {{F2}} = figure2_name
- When templates include placeholders like {{concept}}, {{attribute}}, {{process}},
etc.:
  - Only fill them with words/phrases that appear in ground_truth.

QUESTION TEMPLATES (choose ONE template that best matches relation_type and fill with
{{F1}}/{{F2}}):
<RELATIONAL QUESTION TEMPLATES>
```

Figure 26: Prompt for creating relational reasoning questions with template (continued).

Model Evaluation

```
You are an expert academic tutor designed to answer college-level textbook questions
clearly, accurately, and rigorously.
 Provide your answer in a couple of concise sentences.
 Return your answer in JSON format using the key 'answer'.
<CHAT_HISTORY>
<QUESTION> (with or without <IMAGE>)
```

Figure 27: Prompt for model evaluation, including chat history and question-specific images.

Scoring Relational Questions

```
You are an evaluator assessing the accuracy of an 'actual output' compared to an
'expected output' for relational reasoning tasks.

Follow these evaluation steps:
- "Identify the key facts and the core relationship described in the expected output
(ground truth).",
- "Compare the actual output to determine whether it captures these key facts.",
- "Evaluate whether the actual output correctly reflects the relationship described in
the expected output.",
- "If the 'actual output' includes additional details beyond what is required, do not
increase or decrease the score, as long as those details do not contradict the expected
output."
- "Assign a final score based on alignment with the expected output in terms of both
factual accuracy and relational reasoning using rubric below."

Use the following scoring rubric:
- 1–3 : Response is incorrect or contradicts the expected output; fails to capture the
key facts or the core relationship described in the ground truth.
- 4–6 : The response includes some relevant concepts or facts from the expected output,
but the main relationship is missing, incorrect, contradicted, or poorly explained, and
it contains major omissions or lacks important contextual details.
- 7–8 : Most key facts and the main relationship are correctly captured, but there are
minor omissions, limited specificity or lack of figure-specific examples.
- 9–10 : The response fully aligns with the expected output, including all key facts, the
correct relational reasoning, and all important contextual details (e.g., specific
examples). There are no omissions, no contradictions, and the explanation is clear and
consistent.

Scoring constraints (must be followed strictly):
- If the core relationship and mechanism are correct, the score MUST be at least 7.
- If important contextual details (such as figure-specific examples) are missing, the
score MUST NOT exceed 8.
- Missing contextual details alone should NOT reduce the score below 7 if the core
relationship is correct.
- If the response includes incorrect, misleading, or unsupported information about any
figure (e.g., wrong components or incorrect interpretation), the score MUST NOT exceed
6.

Actual Output:
<actual_output>

Expected Output (Ground Truth):
<expected_output>
```

Figure 28: Prompt for scoring model answers in relational reasoning questions by comparing with ground truth using a rubric, scoring from 1 to 10.

Scoring Other Question Types

```
You are an evaluator assessing the accuracy of an "actual output" compared to an
"expected output" (ground truth).
Your task is to evaluate how well the actual output semantically aligns with the expected
output in terms of correctness, completeness, and factual consistency.

Follow these steps:
- Identify the key facts, concepts, or conclusions in the expected output.
- Compare the actual output to determine whether it captures these key facts.
- Evaluate whether the actual output correctly reflects the facts in the expected
output.
- If the expected output includes image-derived information, treat it as standard
reference facts.
- If the 'actual output' includes additional details beyond what is required, do not
decrease the score, as long as those details do not contradict the expected output."
- Assign a final score based on how well the actual output semantically aligns with the
expected output using the rubric below.

Use the following scoring rubric:
- 1–3: Incorrect; contradicts or fails to capture key facts from the expected output.
- 4–6: Partial; includes some correct elements but misses important facts or contains
notable inaccuracies.
- 7–8: Mostly correct; captures most key facts with minor omissions or slight
inaccuracies.
- 9–10: Complete and accurate; fully contains the expected facts and has no
contradictions.

Scoring constraints (must be followed strictly):
- If the main facts and concepts are correct, the score MUST be at least 7.
- Missing contextual details alone should NOT reduce the score below 7 if the core facts
are correct.
- If the response mentions the correct answer but does not clearly identify it as the
answer, or provides a different incorrect answer as primary, it should be treated as
incorrect (score ≤ 3).
Actual Output:
<actual_output>

Expected Output (Ground Truth):
<expected_output>

Return your evaluation in JSON format:
{{
  "score": <integer from 0 to 10>,
  "justification": "<brief explanation of the score>"
}}
```

Figure 29: Prompt for scoring model answers in other question types by comparing with ground truth using a rubric, scoring from 1 to 10.

E.1 QUESTION TEMPLATES

Visual Question Templates

```
1) Structure and Quantitative
    - "According to the labels in {Figure X}, how many distinct components make up the
{System Name}?"
    - "How many different categories are represented in the legend of {Figure X}"
    - "Based on the numbering in {Figure X}, which step occurs immediately after {Step
Name}?"
    - "Which element in {Figure X} is highlighted or circled to indicate its
importance?"
    - "How many {visible elements} are shown in {Figure X}?"
    - "How many times does the term “{X}” appear in {Figure X}?"

2) Relational and Directional
    - "Looking at the flow of {Figure X}, does the process move in a linear direction or
is it a repeating cycle?"
    - "Which two components are connected by the central bridge shown in the illustration
in {Figure X}?"
    - "In {Figure X}, is {Element A} positioned as a subset of {Element B}, or are they
shown as separate entities?"
    - "Which elements are directly connected by arrows in {Figure X}?"
    - "What connects {element A} to {element B} in {Figure X}?"
    - "Which components share a direct line connection in {Figure X}?"
    - "What structure appears to the left of the {X} in {Figure X}?"

3) Interpretation of Labels in Graphs
    - “What variable is represented on the {x-axis/y-axis} in {Figure X}?”
    - “What unit is used on the {x-axis/y-axis} in {Figure X}?”
    - “What quantity is being measured in {Figure X}?”
    - “What change in {quantity} is described in {Figure X}?”
    - “Which quantity {increase/decrease} in {Figure X}?”
    - “Where does {increase/decrease} occur in {Figure X}?”
    - "At which labeled point does the graph in {Figure X} reach its highest value?"
    - “Where is the {highest/lowest} value described in {Figure X}?”
    - "What change occurs in {quantity} across {component} in {Figure X}?"

4) Identification
    - "What are the labels of the {components} in {Figure X}?"
    - "What text appears inside the {specific shape or box} in {Figure X}?"
    - "What information is provided in the legend of {Figure X}?"
    - "What is the maximum value shown on the vertical axis in {Figure X}?"
    - "What is the minimum labeled value on the horizontal axis in {Figure X}?"
    - "Which column has the highest numerical value in {Figure X}?"

5) Location/ Position
    - "Which component is placed at the very center of {Figure X}, suggesting it is the
focal point?"
    - "In the hierarchy shown in {Figure X}, which items are located at the 'bottom' or
foundational level?"
```

Figure 30: The template for visual questions

Visual Question Templates (Continued)

```
    - "Is the movement in this process in {Figure X} depicted as moving from left-to-
right or top-to-bottom?"
    - "Which component is positioned at the {top/bottom/center} of {Figure X}?"
    - "Which element appears on the left/right side of {Figure X}?"
    - "What text appears in the central {Figure X}?"
    - "Which element is located directly below {X} in {Figure X}?"

6) Sequence/ Order
`   - "In {Figure X}, between {Point A} and {Point B} on the horizontal axis, does the
trend line move upward, downward, or remain flat?"
    - "Which section of the pie chart represents the largest percentage of the whole in
{Figure X}?"
    - "What sequence is indicated by the numbered labels in {Figure X}?"
    - "Which element appears first according to numbering in {Figure X}?"
    - "Which element appears last according to numbering in {Figure X}?"
    - "According to the arrows, which step follows {Step X} in {Figure X}?"
    - "What connects the {A} to the {B} in {Figure X}?
    - "How many total steps are shown in {Figure X}?"
    - "Which step appears to have multiple inputs in {Figure X}?"
    - "Which labeled component disappears in the final stage in {Figure X}?"
    - "What new labeled structure appears in the second panel in {Figure X}?"

7) Categorization and Grouping
    - "How many items are grouped together under the {item/box} in {Figure X}?"
    - “Which elements are described as belonging to the same group in {Figure X}?”
    - “What label is used to describe a group of elements in {Figure X}?”
    - "Are {Item A} and {Item B} in {Figure X} placed within the same labeled group or
different groups?"
    - “What grouping of elements is described on the [left/right/top/bottom] side of
{Figure X}?”
    - "How many categories are shown in the legend in {Figure X}?"

8) Hierarchical
    - “Which category is described as the top level in {Figure X}?”
    - “Which categories are described as sharing the same parent in {Figure X}?”
    - “What hierarchical relationship is described between {Item A} and {Item B} in
{Figure X}?”
    - "Which two categories share the same immediate parent in {Figure X}?"
    - "Which node has the most direct subcategories in {Figure X}?"
    - “How many subcategories are described under {Category} in {Figure X}?”
```

Figure 31: The template for visual questions (continued)

Relational Question Templates

```
1) Comparative
- "How does {{F1}} differ from {{F2}} with respect to {{attribute/concept/function}}?"
- "What is one key similarity between {{F1}} and {{F2}} regarding {{attribute/concept/
function}}?"
- "What does {{F1}} emphasize that {{F2}} does not, and vice versa?"
- "How do {{F1}} and {{F2}} complement each other while highlighting different aspects of
{{concept}}"
- "What key contrast in {{attribute/concept/function/event}} can be identified between
{{F1}} and {{F2}}?"

2) Spatial-contextual
- "How does the position or organization of {{entity}} in {{F1/F2}} correspond to the
structural context shown in {{F2/F1}}?"
- "How does the position of {{entity}} change between {{F1}} and {{F2}}?"
- "How does {{F1/F2}} spatially elaborate {{structure/arrangement}} implied in {{F2/
F1}}?"
- "How does {{entity}}'s placement in {{environment}} differ between {{F1}} and {{F2}}?"
- "How does {{F1/F2}} provide spatial or contextual detail that complements {{content}}
in {{F2/F1}}?"

3) Sequential
- "What is the sequence between {{F1}} and {{F2}} with respect to {{attribute/concept/
function}}?"
- "If these figures are stages of {{concept}}, what stage is shown in {{F1}} and what
stage in {{F2}}?"
- "What changed from {{F1/F2}} to {{F2/F1}} that indicates progression of
{{state_change}}?"
- "If {{F1/F2}} precedes {{F2/F1}}, what likely intermediate step occurs between them?"
- "Write a short timeline linking {{F1}} and {{F2}} for the {{concept}} using terms from
the figures."
- "How do {{F1}} and {{F2}} progress with respect to {{attribute/concept/function}}?"

4) Cause-effect
- "What in {{F1/F2}} could plausibly cause {{case/situation/outcome}} shown in {{F2/
F1}}?"
- "Which detail in {{F1/F2}} most likely produced {{effect/outcome}} in {{F2/F1}}?"
- "List multiple possible causes visible in {{F1/F2}} that could explain {{outcome}} in
{{F2/F1}}."
- "If {{F1/F2}} is not the cause of {{event/outcome}} in {{F2/F1}}, what alternative
cause could explain {{F2/F1}}?"
- "How are {{F1}} and {{F2}} related in cause and effect regarding {{concept}}?"

5) Structural-functional
- "Which parts or components shown in {{F1/F2}} are associated with {{functions}}
described in {{F2/F1}}?"
- "How does the structure in {{F1/F2}} support what is described in {{F2/F1}}?"
```

Figure 32: The template for relational questions

Relational Question Templates (Continued)

```
- "How do the components shown in {{F1/F2}} relate to how {{system}} operates as
explained in {{F2/F1}}?"
- "What structural features highlighted in {{F1/F2}} correspond to {{functional}}
explanation provided in {{F2/F1}}?"
- "What aspect of the structure in {{F1/F2}} appears necessary for {{function/role}}
described in {{F2/F1}}?"

6) Hierarchical
- "How does the concept shown in {{F1/F2}} relate hierarchically to {{constituents/
subunits}} described in {{F2/F1}}?"
- "How does {{F1/F2}} break down or decompose {{concept}} introduced in {{F2/F1}} into
parts or levels?"
- "How do {{F1}} and {{F2}} illustrate a hierarchy of {{concepts/components}}?"
- "How can {{concepts/information/components}} in {{F1}} and {{F2}} be aligned into a
single hierarchy from broader concepts to finer-grained components?"
- "How does {{F1/F2}} provide a top-level view that is refined or expanded by {{concepts/
information/components}} in {{F2/F1}}?"

7) Analogical
- "What shared principle connects {{situation}} in {{F1}} and {{F2}}?"
- "What plays the same role in {{F1/F2}} as {{role_object}} does in {{F2/F1}}?"
- "How do {{F1}} and {{F2}} provide parallel explanations of {{concepts}} at different
levels?"
- "How is the relationship shown in {{F1/F2}} analogous to the relationship shown in
{{F2/F1}}?"
- "How does {{organization}} shown in {{F1/F2}} mirror or correspond to the organization
shown in {{F2/F1}}?"

8) Conditional
- "If {{condition_change}} happened in {{F1/F2}}, how would it likely affect {{outcome}}
in {{F2/F1}}?"
- "If {{key_factor}} were absent in {{F1/F2}}, would {{result}} still occur? Use {{F2/
F1}} as justification."
- "Suppose {{scenario}}. Which figure ({{F1}} or {{F2}}) better matches the expected
result, and why?"
- "Under what condition would {{situation}} in {{F1/F2}} transition into one like {{F2/
F1}}?"
- "If {{assumption}} is true, what conclusion follows when comparing {{F1}} and {{F2}}?"

9) Other
- How does {{F1/F2}} relate to {{F2/F1}}?
- What is the relationship between {{F1}} and {{F2}}?
- How does the information in {{F1/F2}} connect to what is shown in {{F2/F1}}?
```

Figure 33: The template for visual questions (continued)