From Easy to Hard: The MIR Benchmark for Progressive Interleaved Multi-Image Reasoning

Supplementary Material

1.1. Dataset Generation

The MIR benchmark dataset was collected with meticulous attention to detail, ensuring that each image and its associated context contribute meaningfully to improving multimodal reasoning capabilities. Our data sources span a wide range of origins, including public datasets, synthetic generation shooting, web scraping, and scientific publications.

Public Datasets: We sample the images required for the Image Puzzle task from the ALLaVA-4V. The Speed Comparison and Motion Detection categories used frames extracted from videos in the YouTube-BoundingBoxes dataset, ensuring motion dynamics in sequential images.

Synthetic Generation and Shooting: We developed custom pipelines using random parameter generation and rendering tools to generate 3D views. For weight comparison, we used Python scripts to generate a combination of objects placed on the balance for comparison. We manually captured single-object images of 60 objects with consistent backgrounds for Spatial Relationships.

Web Scraping: Images for Forced Perspective were sourced through targeted searches for ”forced perspective” photography. The cooking process involved extracting key step frames from available cooking tutorial videos. Satirical illustrations for ”YES, BUT” were carefully selected from works shared by Anton Gudim on his Instagram page.

Scientific Preprint: For Chart Analysis, the charts were extracted from the papers of arXiv, which includes four types: astrophysics, artificial intelligence, quantitative economics, and statistics, with a small sample of the dataset MultiChartQA.

To ensure the integrity, usability, and ethical compliance of the benchmark MIR dataset, a rigorous filtering process was implemented. All images were subjected to a relevance check, where they were manually reviewed to confirm alignment with the specific objectives of the task, and any irrelevant or ambiguous content was excluded. Ethical considerations were prioritized, as any potentially harmful, offensive, or sensitive material was removed during preprocessing. Additionally, every effort was made to secure proper authorization for all included content, including seeking explicit consent from artists and verifying licensing agreements for web-sourced images.

The detailed generation process for each category will be introduced in the following.

Image Puzzle The process begins by dividing the complete image into a 4x4 grid and removing the 4 patches from the central 2x2 area, separating the image into the removed section and the remaining portion. Using DINOv2, the similarity between these four patches is calculated and samples with an average similarity greater than 0.65 are eliminated. The captions are then generated for the remaining 2x2 subimages, with two distractor patches added: one from another part of the original image and one from a different sample. These are randomly shuffled and passed through GPT-4o to identify the correct 4 patches based on their captions. Samples where the selection is accurate are retained, and the reasoning behind the correct choices is recorded as text2region reasoning. Next, the identified patches and the original image with the central region removed are inputted into GPT-4o to match the patches to their correct positions in the 2x2 area. Samples with at least one correctly matched position are kept and incorrect positions are reevaluated until only samples with all four correctly matched positions remain. The reasoning process is compiled into

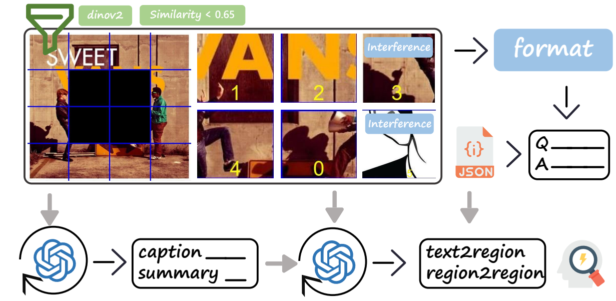
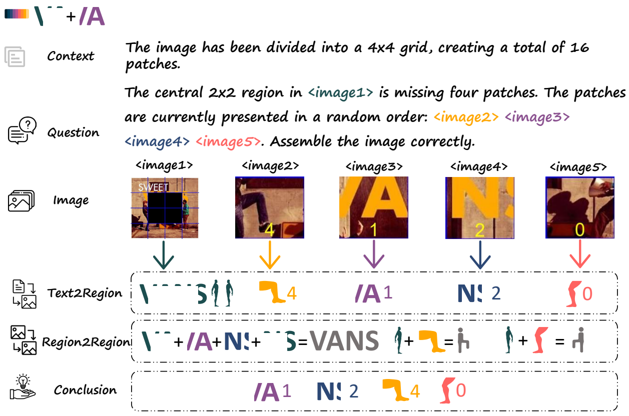


Figure A1. The generation process of the task Image Puzzle.



The image has been divided into a 4x4 grid, creating a total of 16 patches.

The central 2x2 region in <image1> is missing four patches. The patches are currently presented in a random order: <image2> <image3> <image4> <image5>. Assemble the image correctly.

Figure A2. An example of reasoning annotation in the task Image Puzzle.

region2region reasoning. Finally, a question template is created manually, and processed through various models to generate thought processes for summary templates, and DeepSeek is used to create Options and Answer options for each sample, completing the entire workflow.

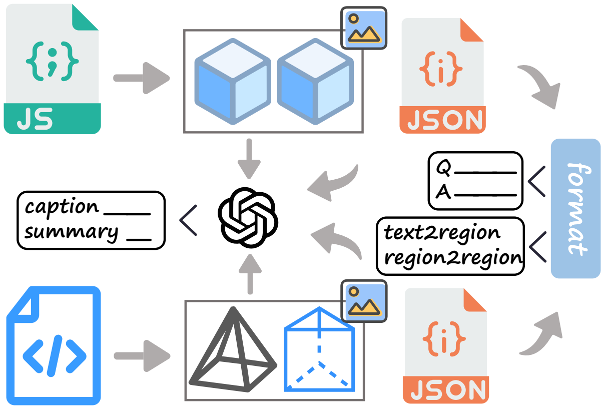
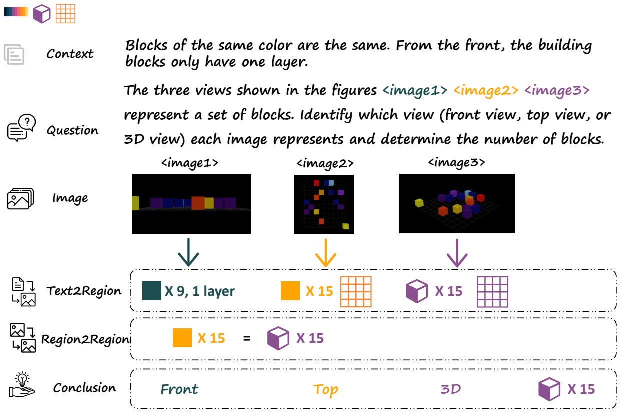


Figure A3. The generation process of the task 3D View.



Blocks of the same color are the same. From the front, the building blocks only have one layer.

The three views shown in the figures <image1> <image2> <image3> represent a set of blocks. Identify which view (front view, top view, or 3D view) each image represents and determine the number of blocks.

Figure A4. An example of reasoning annotation in the task 3D View

3D View The 3D View task programmatically generates three-view and 3D perspective images and their corresponding annotation data for two types of questions: building blocks and prisms/pyramids. For the building blocks type, Three.js is used to render front, top, and 3D perspective views as PNG images. For the prisms and pyramids type, SVG technology dynamically generates main, auxiliary, and top views while displaying geometric features and color gradients. Each question produces a JSON file that includes the question stem, image references, options, correct answers, and multiple sets of text descriptions such as summaries and explanations. The process handles questions in batches of 50 to optimize performance and uses JSZip to package all images and data into a ZIP file for easy download. React hooks are employed to encapsulate core functionalities and clean up resources to avoid memory leaks, ensuring the efficient generation of high-quality 3D view questions and their annotation information.

Depth Relationship The task involves generating depth information for objects through multi-image inference using both the original image and its corresponding RGB depth map. This process integrates large language models to produce structured image annotations. The primary objects and their spatial relationships are identified from the original image, generating descriptive captions. Subsequently, the depth map is utilized to analyze the distance relationships between different objects, thereby completing the inference of depth information. The combination of the original image and depth map undergoes two stages: text-to-region mapping for object localization and region-to-region analysis for depth estimation. The resulting data is then processed using a language model, such as Qwen/Qwen2-VL-72B-Instruct, to generate structured captions and logical inferences while adhering to strict formatting requirements. Finally, all generated images, depth maps, and annotations are consolidated into a JSON file for storage and future analysis.

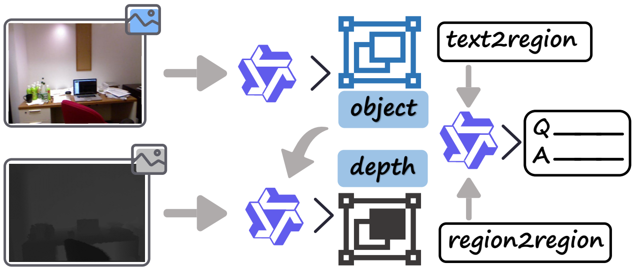
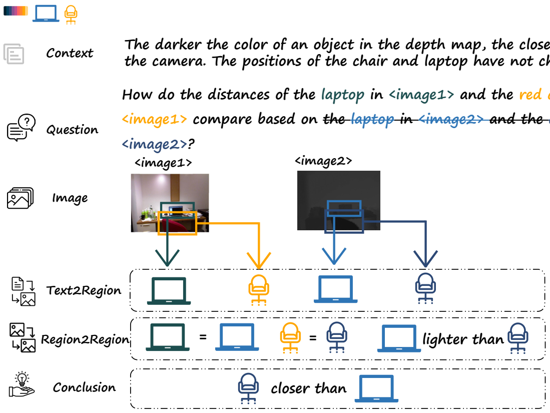


Figure A5. The generation process of the task Depth Relationship.



The darker the color of an object in the depth map, the closer it is to the camera. The positions of the chair and laptop have not changed.

How do the distances of the laptop in <image1> and the red chair in <image1> compare based on ~~the laptop in <image2> and the chair in~~ <image2>?

Figure A6. An example of reasoning annotation in the task Depth Relationship.

Forced Perspective This task begins by finding real

”forced perspective” images online, which typically show objects appearing much larger or smaller than they actually are due to clever camera angles and positioning. These images are then paired with reference objects to form a series of image sets designed to test whether the model can accurately interpret forced perspective effects, such as a giant beach ball appearing larger than a person. These images are then processed using a language model to generate structured captions and logical inferences while adhering to strict formatting requirements.

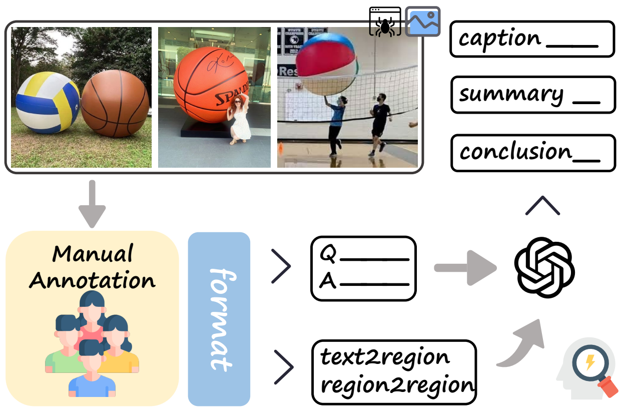
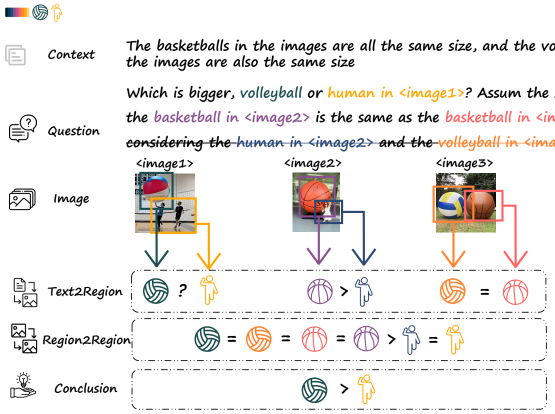


Figure A7. The generation process of the task Forced Perspective.



The basketballs in the images are all the same size, and the volleyballs in the images are also the same size

Which is bigger, volleyball or human in <image1>? Assum the size of the basketball in <image2> is the same as the basketball in <image3>. ~~considering the human in <image2> and the volleyball in <image3>.~~

Figure A8. An example of reasoning annotation in the task Forced Perspective.

Cooking Process To organize the cooking steps provided by a dataset, one can utilize the Qwen API to generate initial question-and-answer pairs. These questions should then be carefully selected and refined to ensure clarity and simplicity, avoiding overly detailed descriptions. identify the relevant cooking steps for each question along with the necessary reasoning analysis. Based on these steps, create text marked with image tags and capture corresponding video frames from the cooking process for association. Additionally, employ models to generate varied incorrect answer options to offer feedback. During the creation of Reasoning Steps, summarize and align questions directly with the dataset’s provided images to mark related visuals. The accuracy of the answer is confirmed through a large model

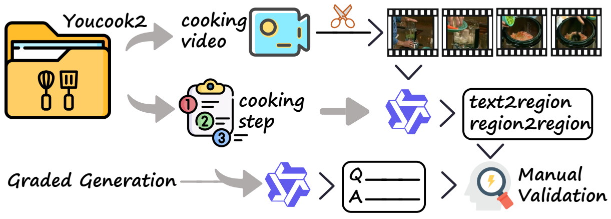
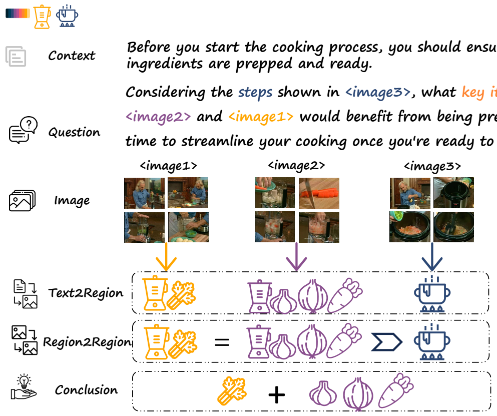


Figure A9. The generation process of the task Cooking Process.



Before you start the cooking process, you should ensure certain ingredients are prepped and ready.

Considering the steps shown in <image3>, what key items mentioned in <image2> and <image1> would benefit from being prepared ahead of time to streamline your cooking once you're ready to begin?

Figure A10. An example of reasoning annotation in the task Cooking Process.

Speed Comparison Relative velocities between a target object (such as a vehicle) and the photographer are inferred through changes in their positional relationships across a series of images. Specific objects within video frames are identified and annotated, with calculations made for positional offsets to exclude instances of excessively rapid camera movements or scene transitions. Video frames are downloaded, and any sets of frames not meeting the criteria are manually removed, and categorized into motion speed categories. LLMs recognize primary objects and pose questions about their relative positions, conducting multiple rounds of inference to refine answers. Different batches of answers are consolidated, and final answers are determined based on the sequence of single-image responses.

Motion Detection The task aims to determine whether the subject or the background/environment is moving by analyzing changes in distance and position between the target object and the photographer. Specific objects, such as cars or buses, are identified and annotated within video frames, with calculations made for changes in the central positions

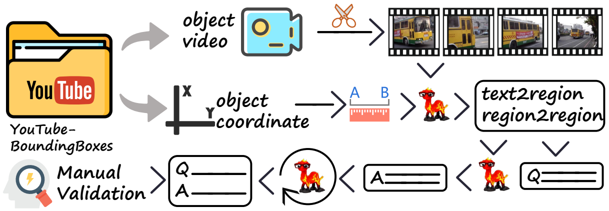
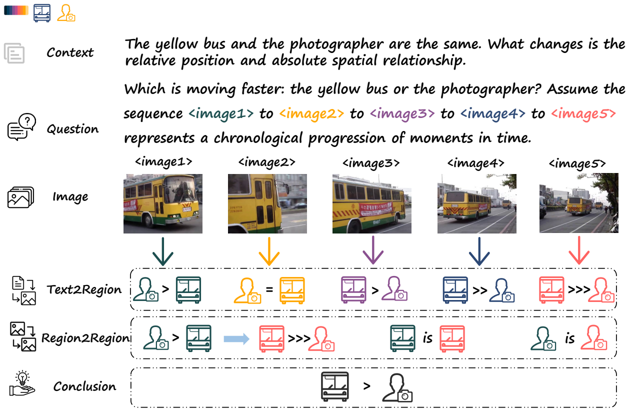


Figure A11. The generation process of the task Speed Comparison.



The yellow bus and the photographer are the same. What changes is the relative position and absolute spatial relationship.

Which is moving faster: the yellow bus or the photographer? Assume the sequence <image1> to <image2> to <image3> to <image4> to <image5> represents a chronological progression of moments in time.

Figure A12. An example of reasoning annotation in the task Speed Comparison.

of these objects to filter out invalid data caused by rapid camera movements. Queries are formulated based on two sub-questions: changes in the absolute position of the object and changes in the object's position relative to the photographer. Answers to these queries contribute to generating multi-image question-answer pairs. By integrating responses from these questions, the task identifies which entity is in motion.

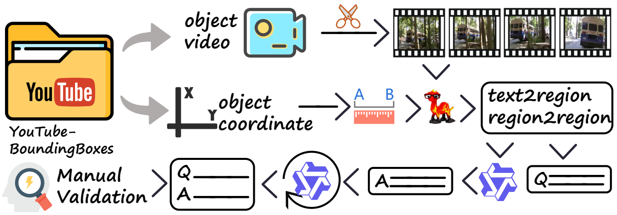
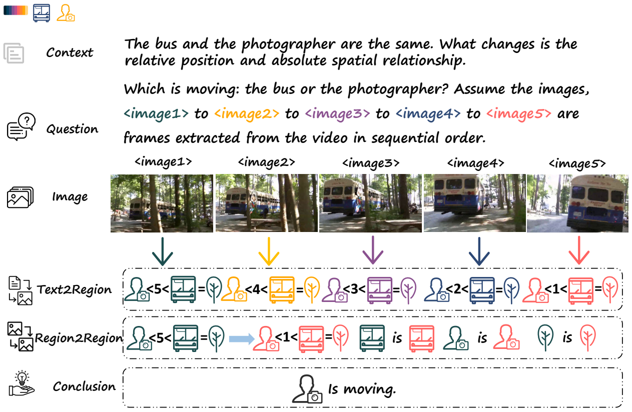


Figure A13. The generation process of the task Motion Detection.

Location Relationship The task involves analyzing the relationships between identical objects across multiple images to infer the spatial arrangement of the leftmost and rightmost objects in a panoramic view. Sixty individual object images are captured with consistent backgrounds, and each object is manually annotated with its ID and name. These single-object images are then stitched together to create multi-object position relationship images, with Gaussian blur applied at the stitching points to smooth transitions. Four types of options are designed: ”A is to the left of B,” ”A is to the right of B,” ”Insufficient intermediate positional information to determine,” and ”Object A or B does not exist.” The number of images and their order for stitching are randomly selected to generate question-answer pairs. Specific questions are formulated based on the sequence of the stitched images.



The bus and the photographer are the same. What changes is the relative position and absolute spatial relationship.

Which is moving: the bus or the photographer? Assume the images, <image1> to <image2> to <image3> to <image4> to <image5> are frames extracted from the video in sequential order.

Figure A14. An example of reasoning annotation in the task Motion Detection.

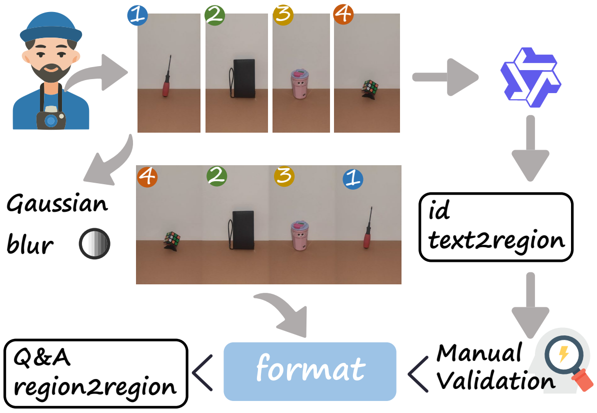


Figure A15. The generation process of the task Location Relationship.

”YES, BUT” Satire Select a series of satirical illustrations with the theme ”YES, BUT” from Anton Gudim's Instagram page. These illustrations showcase the contradictions in human behavior through starkly contrasting scenarios. Utilize the GPT-4o model to generate detailed descriptions for each illustration. These descriptions should cover not only the content depicted but also include an analysis of the satirical elements and their underlying socio-cultural significance. Based on the aforementioned descriptions, compile corresponding questions and answer annotations.

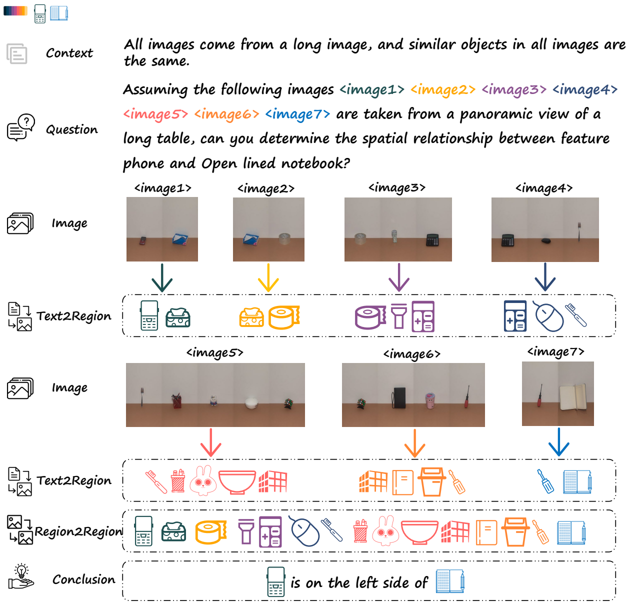

Context
All images come from a long image, and similar objects in all images are the same.
Question
Assuming the following images <image1> <image2> <image3> <image4> <image5> <image6> <image7> are taken from a panoramic view of a long table, can you determine the spatial relationship between feature phone and Open lined notebook?
Image
<image1>
<image2>
<image3>
<image4>
Text2Region
<image5>
<image6>
<image7>
Region2Region
Conclusion
is on the left side of

Figure A16. An example of reasoning annotation in the task Location Relationship.

All generated descriptions, questions, and answers undergo manual review and revision to ensure accuracy. Finally, all annotated data are filtered once more to remove any content that does not meet project standards, ensuring the quality and compliance of the final dataset.

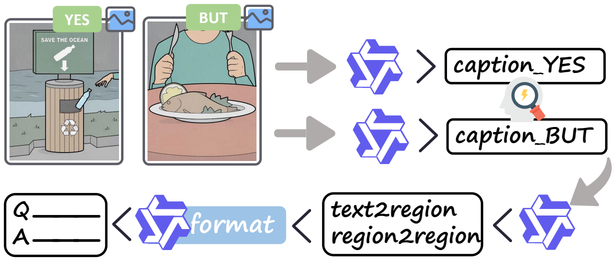


Figure A17. The generation process of the task "YES, BUT" Satire.

Scale Blur Invoke the image generation model's API and combine keywords of different objects to batch-generate a series of related images. These images should include a common reference object for size comparison purposes. Utilize this generated image set to formulate questions regarding the relative sizes of the items, and conduct logical reasoning based on the reference object to derive the correct answers. For instance, use an ice cream cone as a reference to compare the relative sizes of a chandelier and a coffee cup. Document these questions along with their corresponding answers and detailed reasoning processes. After undergoing manual review and compilation, this will culminate in a finalized dataset aimed at enhancing the spatial reasoning capabilities of large models concerning size comparisons. The aim is to enable the model to recognize counterintuitive objects in images based on inference rather than prior knowledge. Throughout this process, selecting commonly found and AI-familiar objects is crucial for boosting the accuracy of the dataset.

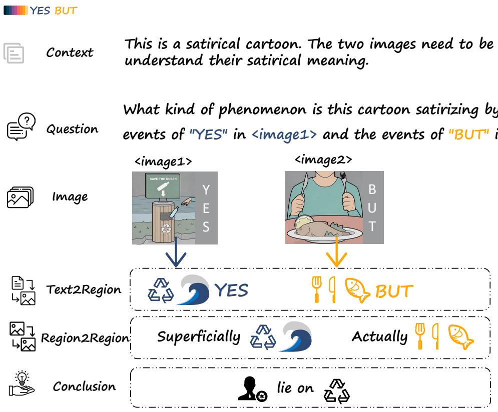

YES BUT
Context
This is a satirical cartoon. The two images need to be linked together to understand their satirical meaning.
Question
What kind of phenomenon is this cartoon satirizing by comparing the events of "YES" in <image1> and the events of "BUT" in <image2>?
Image
<image1>
<image2>
Y E S
B U T
Text2Region
YES
BUT
Region2Region
Superficially
Actually
Conclusion
lie on

Figure A18. An example of reasoning annotation in the task "YES, BUT" Satire.

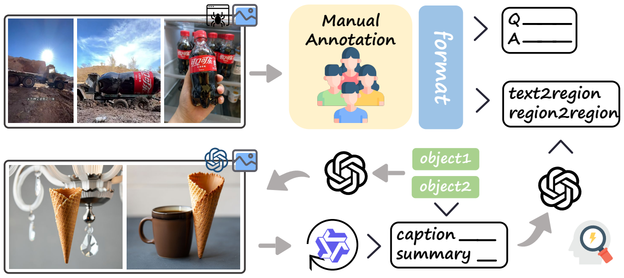


Figure A19. The generation process of the task Scale Blur.

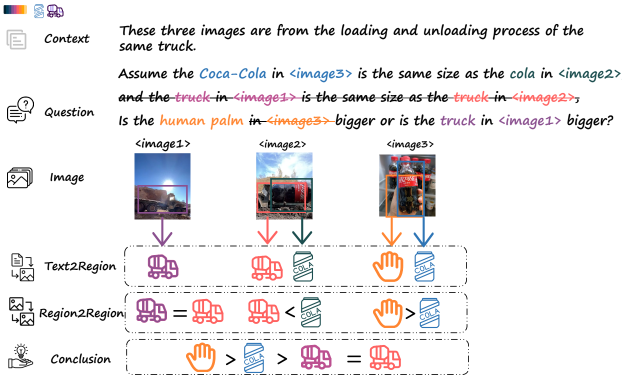

Context
These three images are from the loading and unloading process of the same truck.
Question
Assume the Coca-Cola in <image3> is the same size as the cola in <image2>,
~~and the truck in <image1> is the same size as the truck in <image2>~~,
Is the human palm ~~in <image3>~~ bigger or is the truck in <image1> bigger?
Image
<image1>
<image2>
<image3>
Text2Region
Region2Region
=
<
>
Conclusion
>
>
=

Figure A20. An example of reasoning annotation in the task Scale Blur.

Weight Comparison This task is primarily achieved through programming. It starts by setting different types and quantities of entities on both sides of a balance scale and then combining these entities with images of the scale. By designing multiple inequality groups to generate counterintuitive answers—where logically the Entities with relatively light intuitive feelings turn out to be heavier—it produces several sets of weight comparison images aimed at testing the model's reasoning ability rather than relying on prior knowledge. These images are then processed using a language model to generate structured captions and logical inferences while adhering to strict formatting requirements.

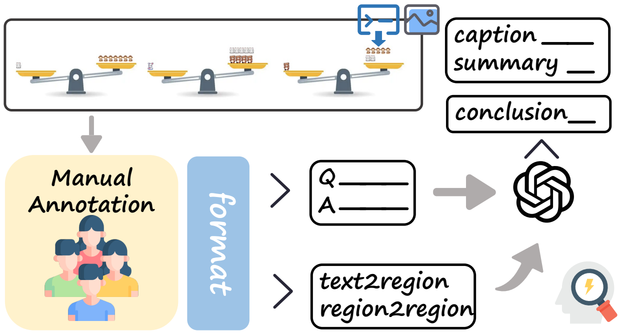
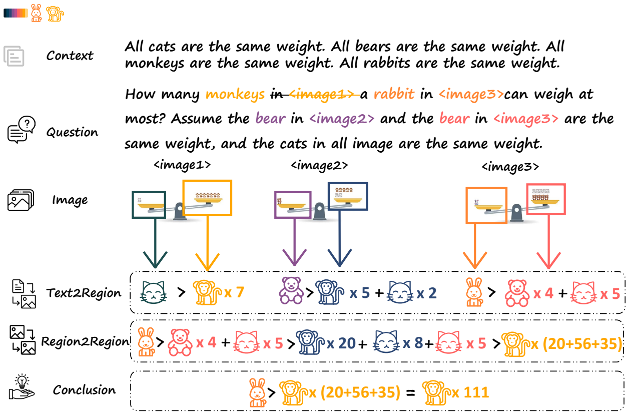


Figure A21. The generation process of the task Weight Comparison.



All cats are the same weight. All bears are the same weight. All monkeys are the same weight. All rabbits are the same weight.

How many monkeys ~~in <image1>~~ a rabbit in <image3>can weigh at most? Assume the bear in <image2> and the bear in <image3> are the same weight, and the cats in all image are the same weight.

Figure A22. An example of reasoning annotation in the task Weight Comparison.

Chart Analysis The dataset images were sourced from Gallup Inc. and the Bureau of Economic Analysis (BEA) in the United States, alongside chart data extracted from arXiv papers covering "Astrophysics," "Artificial Intelligence," "Quantitative Finance," and "Statistics" in order to ensure that each set of images has interrelated charts. Additionally, a small sample was taken from the Multi-Chart-QA dataset. We rigorously screened these images to exclude any that did not meet our criteria or were unlawful. Using the Qwen model, we generated detailed descriptions for individual images. Based on these descriptions, we synthesized corresponding annotations of questions, answers, and reasoning steps, which were subsequently reviewed and edited by human annotators. An additional screening process was conducted to eliminate any annotations that did not meet our stringent quality standards, ensuring the final dataset's quality and compliance.

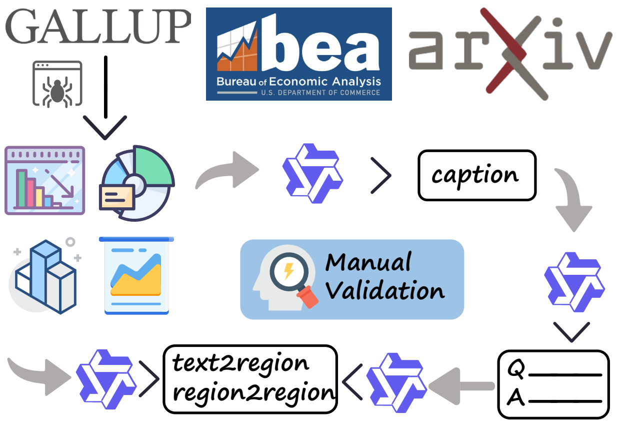
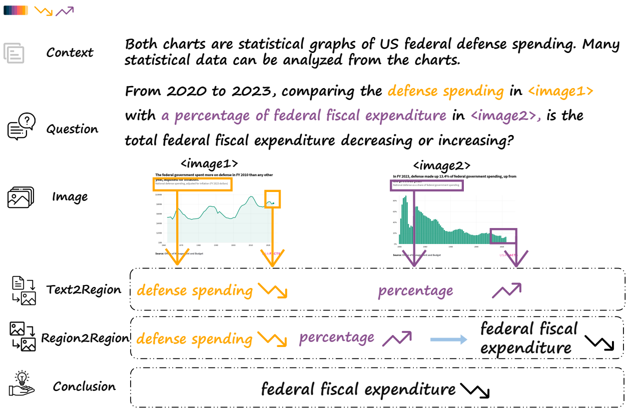


Figure A23. The generation process of the task Chart Analysis.


Context
Question
Image
<image1>
The federal government spent more on defense in FY 2010 than any other year, adjusted for inflation.
<image2>
In FY 2023, defense made up 13.4% of federal government spending, up from the previous year.
Text2Region
defense spending
percentage
Region2Region
defense spending
percentage
federal fiscal expenditure
Conclusion
federal fiscal expenditure

Both charts are statistical graphs of US federal defense spending. Many statistical data can be analyzed from the charts.

From 2020 to 2023, comparing the defense spending in <image1> with a percentage of federal fiscal expenditure in <image2>, is the total federal fiscal expenditure decreasing or increasing?

Figure A24. An example of reasoning annotation in the task Chart Analysis.