

A Appendix

The subsequent sections offer essential insights into reproducibility, encompassing detailed explanations and examples regarding model evaluation, analysis, and data collection.

For access to the complete Visual Riddles dataset, visit <https://huggingface.co/datasets/visual-riddles/visual-riddles>. Comprehensive information about the dataset fields and loading instructions is provided therein.

A.1 Reproducibility and Resources

We executed all evaluation code on a single Colab notebook, accessible on our website <https://visual-riddles.github.io/>. For models using APIs like Gemini and GPT-4, we employed CPU resources. However, for models like Llava and InstructBLIP, which required GPU resources, we utilized a single A100 GPU. We utilized a paid plan for models with APIs, enabling us to generate 400 predictions for each task within a timeframe of less than 30 minutes, equivalent to the runtime of Llava and InstructBLIP models used locally with GPUs. Table 6 lists the models specified in the APIs, and how they were used.

Auxiliary Data: Attributions For each visual riddle containing an attribution field, we extracted the textual content from the webpage URL using Selenium WebDriver (<https://www.selenium.dev/>). Consequently, prompts containing attributions may be lengthy.

Table 6: APIs model specification

API	Model Name	Specified	Type of Model Used	Tasks Used
GPT4	gpt-4-vision-preview		VLM	Open-Ended, Multiple-Choice (Distractors + Evaluation), Auto-Eval Judge
Gemini-Pro-1.5	gemini-1.5-pro-latest		VLM	Open-Ended, Multiple-Choice, Auto-Eval Judge, Auto-Rater
Gemini-Pro-1.5	gemini-1.5-pro-latest		LLM	Caption → LLM on Open-Ended
Gemini-Pro-Vision	gemini-pro-vision		VLM	Open-Ended, Multiple-Choice, Auto-Eval Judge

A.2 Image Generation Guidelines

The Visual Riddles benchmark was created by seven designers including four women and three men, most of whom are authors of this paper, all from the same country, experienced in generating images using text-to-image models. The annotators were instructed to create images that integrate information with world knowledge and common sense to answer a given textual question. The answer must be grounded in the data presented in the image, making it impossible to respond accurately without comprehending the image and identifying the embedded clues. To generate high-quality images, our designers are given access to advanced text-to-image models, including Midjourney¹, Ideogram², Canva³, DALL-E 3 [38], and Stable-Diffusion [39]. Each image must be sufficiently challenging such that at least one of the evaluation models (e.g., Gemini-Pro-1.5 [7], GPT4-turbo-preview [11] and LLaVA-1.6-34B [40]) fails to correctly answer the question based on the provided image. In addition, the designers provided not only the correct answer to the question but also a hint to the image that should guide where to look in the image when trying to solve the riddle (Example in Fig. 1)

¹<https://www.midjourney.com/home>

²<https://www.ideogram.ai>

³<https://www.canva.com>

Call for AI image designers!

We are collecting human designers to assist with the generation of a general world knowledge and common-sense challenge data set.

The generation task involves the creation of unique instances each consisting of an image, a question, and an answer. These instances should test the limits of AI's understanding of general world knowledge by presenting scenarios that require multi-step reasoning based on visual cues. The main goal is to craft questions that an AI model cannot easily answer just by analyzing the image and the text, thereby pushing the boundaries of current AI capabilities.

The images should be in a .jpg format 1024x1024 size, the textual file should contain the question, the answer, the relevant image file name for each instance, and the name of the models that you were used to generate each of your images. This file should be in a .csv format:

Question	Answer	Image file name	Image caption	Generative Model name	Hint	Attribution
----------	--------	-----------------	---------------	-----------------------	------	-------------

Examples for generative models:

<https://ideogram.ai/top/1>, <https://screenshot.googleplex.com/333oLFUd2FuhfPZ>,
https://stablediffusionweb.com/#google_vignette, <https://www.canva.com/ai-image-generator/>,
<https://www.midjourney.com/home>, <https://gemini.google.com/app>

Image Creation

Realism and AI Generation: Each image should be crafted to look as realistic as possible, utilizing advanced AI-based image generation tools such as Dall-E, Midjourney, or Stable Diffusion etc. The chosen imagery should directly relate to the question and answer, serving as a visual foundation for the challenge.

Accuracy Verification: After generating an image, ensure that it accurately represents the intended scenario or concept (the concept must be a real-world one, don't use ideas that do not happen in reality). The image should visually encapsulate all the elements necessary to lead to the question and, by extension, the answer.

Crafting Questions

Challenge Through Indirect Connection: Questions should be designed to challenge AI models by requiring two-step logical reasoning (two-hop) based on the information depicted in the image. For instance, an image showing leafless trees and lush greenery suggests a winter scene, which implies colder temperatures, leading to questions about appropriate attire for the conditions.

Specificity and Relevance: Each question must directly relate to the visual and conceptual content of the image, prompting the responder to make logical connections between what is seen and the broader implications or facts of the world.

additional note : verify that each question cannot be answered without the relevant image.

Formulating Answers

The answers can be short and free-style, please follow these:

Detail-Oriented Responses: Answers should be comprehensive, detailing all aspects of the image relevant to the question. This includes describing the visible elements that lead to the understanding necessary to address the question accurately.

Completeness and Advisory: Beyond just stating facts, answers should also provide a conclusive statement or recommendation based on the question posed. The response should leave no ambiguity about the appropriate course of action or the logical conclusion to be drawn from the image-question context.

PAY ATTENTION: the answer should refer to both the visual clue in the image, and the commonsense/world knowledge that is required for the solution.

Hint: a textual hint that directs the solver (human or model) for the visual clue. cannot be the clue itself.

Attribution: for questions that their answer requires a "reliable source" like answers that talk about rules/countries/biological conditions/cultural rules/etc, there is a need to provide a link for a webservice that confirms your answer (does not have to be only wikipedia).

Figure 7: Guidelines for human designers to create a visual riddle — Part 1. This section includes instructions on the visual riddle creation process including requirements for captions, generative models, hints, and attribution.

and, for riddles requiring world knowledge, include attributions to relevant sources (see the full instructions in Fig. 7, Fig. 8 and the bottom example in Fig. 1). After creation, each image-question pair undergoes a peer review by at least three other designers to ensure the hint's clarity and the riddle's solvability.

A.3 Categories and Difficulty Level: Breakdown and Results

To examine how different types of world knowledge and difficulty levels impact model performance, we analyze their correlations to identify particularly challenging combinations for current models. The distribution of the different categories and difficulty level (with 3 - most difficult and 0 - least

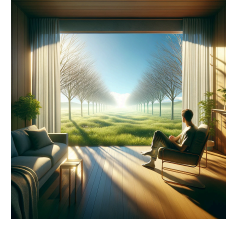
Examples:

1. **Q: He wants to go outside, what should he wear?**

A: "In his window, there are leafless trees and high greenery, features characteristic of the winter season. Therefore, despite the sunny appearance, the temperature could be low. It is advisable to wear warm clothing when going outside."

H: Look at the plants outside.

At: <https://forestryandland.gov.scot/blog/trees-in-autumn>



2. **Q: What is the language she would probably speak in this place?**

A: Salmon and cream cheese bagel is a popular food in NY, USA. so she probably will speak english there.

H: What does she eat?.

At: <https://www.alphafoodie.com/the-best-lox-bagel/>



3. **Q: Why does she look like this?**

A: The girl in the image holds a balloon, and her hair is standing up. When a balloon is rubbed against hair, electrons are transferred from the hair to the balloon, creating a static charge that causes the static hair phenomenon. So the reason for the appearance of standing hair is static electricity.

H: what is she holding?

At: https://en.wikipedia.org/wiki/Static_electricity



4. **Q: How to get out?**

A: There is a bat in this cave, therefore a way out of the cave is the way that the bat came in through. So, to get out, one must follow the entrance-path of the bat.

H: What can you see in the cave?



These examples illustrate the intended structure and depth of each instance within the dataset. Designers are encouraged to use this as a template for creativity, ensuring that each element of the instance—the image, question, and answer—works harmoniously to challenge and enhance AI understanding of complex, real-world knowledge.

Figure 8: Guidelines for human designers to create a visual riddle — Part 2. This section includes examples, and, similarly to the presented examples, designers were asked to provide additional details such as the image caption, the generative model used, a hint, and attribution for their riddles.

difficult) are illustrated in Tables 7 and 8. Fig. 9 presents a heatmap illustrating model failures and human failures across various categories and difficulty levels within the Visual Riddles challenge. Across 16 categories and 4 difficulty levels, our analysis identifies model weaknesses in categories like “Object Counting”, “Temporal Principles”, and “Physical Principles”. Notably, human performance on “physical principles” appears to be even more susceptible to errors compared to the models on difficulty index level of 3. Naturally, models as well as humans also seem to struggle more with the higher difficulty instances (levels 2 and 3 in Fig. 9), highlighting the most demanding elements of the benchmark for current vision-and-language models. Yet this highlight the gaps between humans and models performances.

To provide a clearer understanding of the performance across difficulty levels, we included aggregated results for open-ended VQA accuracy of both humans and models, categorized by different categories and difficulty index levels (see Figure 10). The analysis shows that certain categories, such as Temporal Principles, Object Counting, and Biological Principles, pose greater challenges for models, while Geographic Knowledge and Physical Principles are more difficult for humans. Interestingly,



Figure 9: A heatmap depicting how models (Top) and human annotators (Middle) fail to solve visual riddles in the Visual Riddles dataset, visualizing by the color intensity of each cell. The x-axis represents different categories of riddles, while the y-axis shows the difficulty index levels. Red cells highlight particularly challenging areas, with vertical red bands indicating categories that consistently pose difficulties and horizontal red bands confirming the appropriateness of the difficulty index levels assigned. This visualization underscores the diverse range of commonsense knowledge needed to effectively tackle the visual-riddles. Below, there is an example of images related to some of our categories.

there is no significant correlation between the difficulty categories for humans and models, with a correlation coefficient of -0.12. For difficulty levels, while human performance consistently surpasses models, both show a decline as difficulty increases, with a strong correlation of 0.94 (see top and middle parts of Fig. 9). To further illustrate, examples of images corresponding to different difficulty levels are provided in Fig. 10 (bottom).

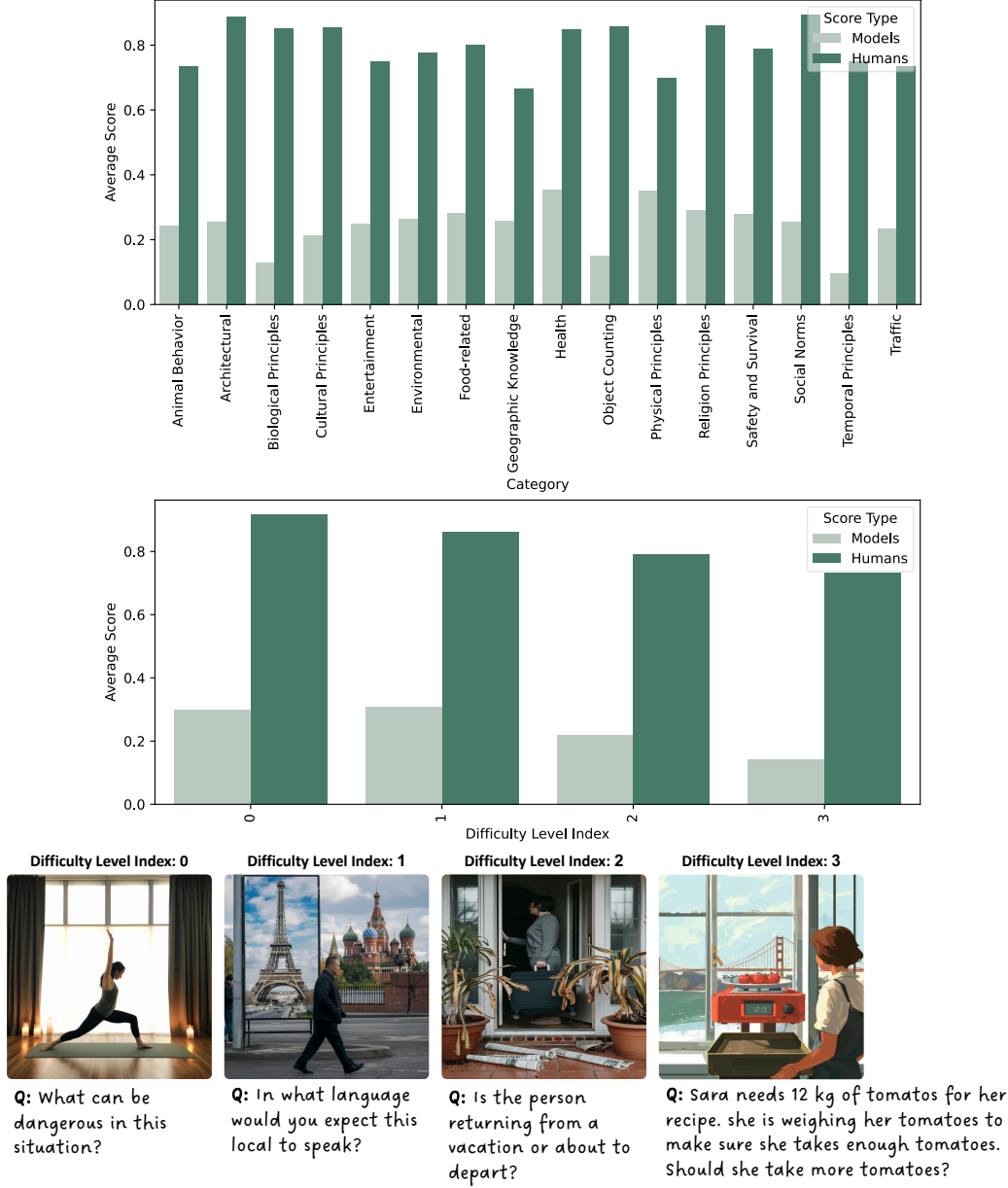


Figure 10: Aggregated results for model and human performance across categories (top) and difficulty levels (middle), along with visual riddle examples illustrating varying difficulty levels (bottom). The top section shows the average open-ended VQA accuracy for models and humans by category, with a correlation of -0.12 between their performances. The middle section presents the average open-ended VQA accuracy for models and humans by difficulty level, showing a correlation of 0.94.

A.4 Prompts for Different Models

There are several tasks with different prompts:

Multiple-Choice VQA: A specific prompt is used for generating distractors when there are insufficient incorrect answers using GPT-4. Additionally, for multiple-choice VQA, we evaluate the models' performance across three settings: (1) given only the image, the question, and five possible answers (with only one correct answer), (2) given the image, the question, the possible answers, and a hint,

Table 7: Distribution of Categories Across Images

Category	% Images
Cultural Principles	19.50
Social Norms	15.75
Safety and Survival	8.25
Food-related	7.50
Object Counting	7.00
Health	6.00
Traffic	5.00
Animal Behavior	5.00
Geographic Knowledge	4.50
Entertainment	4.00
Architectural	3.75
Temporal Principles	3.00
Environmental	3.00
Religion Principles	3.00
Physical Principles	2.50
Biological Principles	2.25

Table 8: Distribution of Categories Across Images

Difficulty Level	% Images
0	11.10
1	33.25
2	38.25
3	17.50

and (3) given the image, the question, the possible answers, and an attribution. The prompts structure is outlined in Table 9.

VQA Automatic Evaluation: To find the best judge (auto-rater), we evaluate the models in two scenarios - *reference-free* and *reference-based*. The prompts structure is outlined in Table 10.

Auto-Rater for Open-Ended VQA and Ablation Study with Modified Images: Using the best Auto-Rater we evaluate all models automatic ratings on the Open-Ended VQA task. We also use the same prompt in our analysis we perform ablation study to explore whether models base their answers solely on text or consider visual clues. The prompts structure is outlined in Table 10.

A.5 Open-Ended VQA: Annotators Solve Visual Riddles Guidelines

In order to evaluate how well humans are capable in solving the visual riddles questions, a human response were collected using Amazon Mechanical Turk platform www.mturk.com. To gather human responses for the benchmark’s riddles, we used the Amazon Mechanical Turk platform, paying annotators \$18 per hour. We contacted workers with a proven track record in similar tasks and invited them to a qualification round that began with a review of task guidelines and included solving five riddles of varying difficulty. Following an assessment of their responses and providing personalized feedback, only those demonstrating a strong understanding of the task qualified. Of 14 candidates, 10 proceeded to the actual annotation, where each of the 400 riddles was solved by three annotators. In the guidelines, annotators were presented with five examples. For each example, they were initially shown the visual riddle and subsequently given the solution and the process for solving the question. We recommended that annotators first attempt to solve the riddles on their own before reviewing our provided answers.

Two examples of the guidelines are in Fig. 11 and Fig. 12. In the first, answering the question required only common-sense and counting capabilities of objects in the image while in the second example, world knowledge about cultural principles is needed therefore, the workers were expected to search the different items, understand the clues given in the image and only then answer the question.

Table 9: Prompts for different models for Multiple choice VQA

Task	Prompt
Creating Distractor	“Here is a question regarding the image, and a ground-truth answer. \n Question:< <i>question</i> > \n Ground-Truth Answer < <i>ground_truth_answer</i> > \n\n Please generate < <i>num_of_incorrect_distractors</i> > wrong answers (that are kind-of similar to the ground-truth answer, and in a similar length) to the question based on the image, in the format of: \n\n YOUR FIRST ANSWER@@@YOUR SECOND ANSWER@@@...”
Clean	“This is a multiple-choice question concerning the image. Out of the options labeled (1)-(5), only one is correct. Please provide your answer as a single digit that corresponds to the correct option. For instance, if the correct answer is (3), you should respond with 3. \n\n Question: < <i>question</i> > \n\n Candidate answers: \n (1)< <i>candidate1</i> > \n (2)< <i>candidate2</i> > ...”
+ Hint	“This is a multiple-choice question concerning the image. Out of the options labeled (1)-(5), only one is correct. Please provide your answer as a single digit that corresponds to the correct option. For instance, if the correct answer is (3), you should respond with 3. \n\n Question: < <i>question</i> > \n\n Hint: < <i>Hint</i> > \n\n Candidate answers: \n (1)< <i>candidate1</i> >\n (2)< <i>candidate2</i> > ...”
+ Attribution	“This is a multiple-choice question concerning the image. Out of the options labeled (1)-(5), only one is correct. Please provide your answer as a single digit that corresponds to the correct option. For instance, if the correct answer is (3), you should respond with 3. Additionally, an attribution, which is a textual content from a webpage providing the basis for the correct answer, is also included below. Use this information to select the correct answer. \n\n Question: < <i>question</i> > \n\n Attribution: \n\''\''\''\''< <i>attribution</i> >\n\''\''\''\''\n\n Candidate answers: \n (1)< <i>candidate1</i> >\n (2)< <i>candidate2</i> > ...”

Table 10: Prompts for different models for Automatic Evaluation

Task	Prompt
Reference-Free, Ablation Study with Modified Images	“Answer with only Yes OR No. Given the image and the question, is the candidate answer correct? \n Question: < <i>question</i> > \n Candidate Answer: < <i>candidate_answer</i> > \n”
Reference-Based, Auto-Rater for Open-Ended VQA	“Answer with only Yes OR No. Given the image, the question and the ground-truth answer, is the candidate answer correct? \n Question: < <i>question</i> > \n Ground-Truth Answer:< <i>ground_truth_answer</i> > \n Candidate Answer:< <i>candidate_answer</i> > \n”

In Fig. 13 there is an example of the UI page for annotation of solving a visual riddle.

A.6 Open-Ended VQA: Humans and Models Answers Annotation Evaluation Guidelines

In order to evaluate how well models and humans answered the visual riddle, we used MTurk platform, paying 18\$ per hour. We contacted workers and invited them to a qualification round that began with a review of task guidelines and included solving five riddles of varying difficulty. Following an assessment of their responses and providing personalized feedback, only those with a strong understanding of the task moved on to the actual annotation phase. Of 9 candidates, 6 proceeded to the actual annotation phase. To ensure greater cultural diversity in the benchmark, all AMT annotators were selected from countries different from those of the visual riddles creators.

1. Visual Riddle

Question: How many bracelets are there in the picture?



2. Answer

Answer:

there are two bracelets on the table. in addition, the person in the back is wearing three more bracelets. therefore, there are in total 5 bracelets in the picture.

3. Explanation

To correctly answer many such riddles, you would have to observe the entire image, and not only what is in the front. Many of the hints would appear in the background or any other location that is not in plain sight, and it'd be your job to pick up on all those hidden hints.

Figure 11: Example of the guidelines for MTurk annotators for solving a visual riddle as an open ended question. (1) Question: Solving visual riddles composed of images and open-ended questions. (2) Answer: the answer to the question. (3) Explanation & Notes: explanation of why this is the correct answer given the question and the image.

During annotation, In the LVLM and the Caption → LLM cases, each annotator was provided with the image, the question, and the correct answer, along with the candidate answers from six different models and three humans (the human answers were obtained as described in §5.2). In the same way (image, question and ground truth answer), we annotated the candidate answers for human (oracle) and Gemini-Pro-1.5 captions from the best model as described in §5.2.

In the guidelines, annotators were presented with five examples. For each example, they were initially shown the image, the question, the gold-answer, and the candidate answers. Afterwards, they could see which answers were correct and the reasoning behind the correctness or incorrectness of each candidate answer. Marking a candidate answer as correct required that it not only accurately answered the question based on the given answer but also contained no hallucinations.

1. Visual Riddle

Question: what holiday is probably taking place today?



2. Answer

Answer:

the Russian flag in the background suggests this is a Russian holiday. Additionally, there is a traditional Shuba salad on the table, which is usually prepared for new years or christmas. therefore, it is most likely the holiday celebrated is either new years or christmas

3. Explanation

Another type of question requiring cultural knowledge would revolve around holidays and special festivals from around the world. As in the previous examples, we search for potential hints within the image, and if necessary look up relevant details on Google. → [see search example in the next slide](#)

4. Attribution

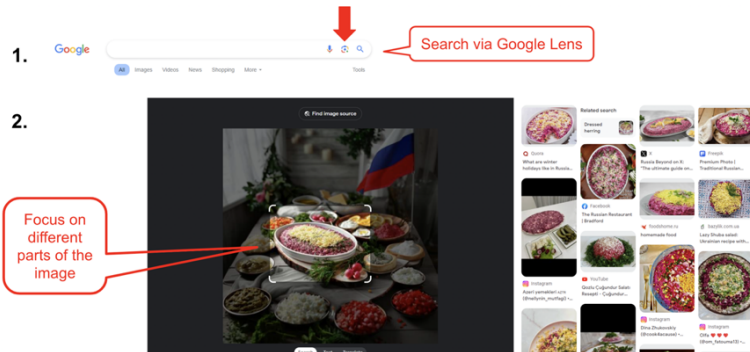
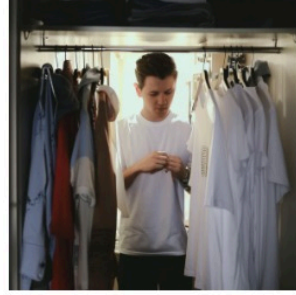


Figure 12: Example of the guidelines for MTurk annotators for solving a visual riddle as an open ended question. (1) Question: Solving visual riddles composed of images and open-ended questions. (2) Answer: the answer to the question. (3) Explanation & Notes: explanation of why this is the correct answer given the question and the image. (4) Attribution: for this image a world knowledge is required therefore - a search in google is helpful in getting the data.

An example of the guidelines as well as a UI page is in Fig 14. Using these annotations as mentioned in §5.3 we compose the Multiple-Choice VQA prompts by sampling three answers annotated as incorrect. If there are fewer than three, we generate additional incorrect answers. We then include the ground-truth answer provided by the riddle designer, along with the "cannot determine" option.

Instructions: Given an image and a question, answer the question based on the image.

Important: Make sure you look for hints and details that are hidden within the image.



Question: Does Dani have any black clothes?

Answer the question based on the image

Submit

Figure 13: Example of MTurk UI page for answering the question given an image and a question

1. Evaluation

Question: Is there a way out?

Answer: In a pitch-black cave with no apparent entrance or exit, a bat hangs from the ceiling. Given the presence of the bat, it implies there must be a passage through which it entered, indicating the existence of a way out of the cave.

The image you've provided appears to be a stylized or artistic representation of a cave or a dark, enclosed space. It's not a real photograph, so it's not possible to determine if there's a way out based on the image alone. In a real-world scenario, if you were in a cave or a similar environment, you would need to explore the cave to find an exit or a way back to the entrance. It's important to have proper equipment and knowledge when exploring caves to ensure safety.

☐ There is a way out, but it is not easy. The way out is to fly up and out of the cave. However, the bat cannot fly because its wings are injured. The bat will need to find another way out of the cave.

☐ Yes there is a way out. There is a bat hanging inside this cave. There has to be an opening that the bat flew in through, and the same opening would also be a way to get out.

☐ The image depicts a dark cave with a bat hanging from the ceiling, suggesting that there is a way out of the cave.

☐ The cave is very dark, so that suggests that there is no way out, but a bat has managed to find its way inside, and would need to find its way out. The cave has no opening or an exit, which allows the bat to fly out and find its way to the bright light.

☐ Yes, there is a way out. The image shows a cave with a bat flying through it, and the bat is heading towards a bright light. This suggests that the cave has an opening or an exit, which allows the bat to fly out and find its way to the bright light.

☐ The image depicts a bat hanging upside down in a dark cave environment with rock textures that might be challenging to traverse. Whether there is a "way out" of this cave could depend on various factors not visible in the image, such as the layout of the cave, the presence of any exits or entrances other than the one the bat is hanging near, and so on. Without additional information or a broader view of the cave, it's not possible to determine if there is a way out from this scene alone.

☐ Yes, because though there's no apparent entrance or exit, there's a bat hanging from the ceiling. Bats have yet to master teleportation, indicating there must be an entry, and therefore, an exit from the cave.

☐ It appears there could be a way out of the cave. The image shows a dark cave with rough rock walls and a small opening at the end of a tunnel, which could lead outside. However, with the bat hanging from the top of the cave, it may be wise to proceed with caution.

2. Evaluation

Question: Is there a way out?

Answer: In a pitch-black cave with no apparent entrance or exit, a bat hangs from the ceiling. Given the presence of the bat, it implies there must be a passage through which it entered, indicating the existence of a way out of the cave.

The image you've provided appears to be a stylized or artistic representation of a cave or a dark, enclosed space. It's not a real photograph, so it's not possible to determine if there's a way out based on the image alone. In a real-world scenario, if you were in a cave or a similar environment, you would need to explore the cave to find an exit or a way back to the entrance. It's important to have proper equipment and knowledge when exploring caves to ensure safety.

☐ There is a way out, but it is not easy. The way out is to fly up and out of the cave. However, the bat cannot fly because its wings are injured. The bat will need to find another way out of the cave.

☒ There is a way out. There is a bat hanging inside this cave. There has to be an opening that the bat flew in through, and the same opening would also be a way to get out.

☒ The image depicts a dark cave with a bat hanging from the ceiling, suggesting that there is a way out of the cave.

The cave is very dark, so that suggests that there is no way out, but a bat has managed to find its way inside, and would need to find its way out. The cave has no opening or an exit, which allows the bat to fly out and find its way to the bright light.

☐ Outside again in order to exit, so we can be certain that there is at least one exit to this cave. The way out may be very small, and only accessible to a small flying animal, but it will be there, somewhere.

☐ Yes, there is a way out. The image shows a cave with a bat flying through it, and the bat is heading towards a bright light. This suggests that the cave has an opening or an exit, which allows the bat to fly out and find its way to the bright light.

☐ The image depicts a bat hanging upside down in a dark cave environment with rock textures that might be challenging to traverse. Whether there is a "way out" of this cave could depend on various factors not visible in the image, such as the layout of the cave, the presence of any exits or entrances other than the one the bat is hanging near, and so on. Without additional information or a broader view of the cave, it's not possible to determine if there is a way out from this scene alone.

☒ Yes, because though there's no apparent entrance or exit, there's a bat hanging from the ceiling. Bats have yet to master teleportation, indicating there must be an entry, and therefore, an exit from the cave.

☐ It appears there could be a way out of the cave. The image shows a dark cave with rough rock walls and a small opening at the end of a tunnel, which could lead outside. However, with the bat hanging from the top of the cave, it may be wise to proceed with caution.

3. Evaluation explanation

(1) ✗ The image you've provided appears to be a stylized or artistic representation of a cave or a dark, enclosed space. It's not a real photograph, so it's not possible to determine if there's a way out based on the image alone. In a real-world scenario, if you were in a cave or a similar environment, you would need to explore the cave to find an exit or a way back to the entrance. It's important to have proper equipment and knowledge when exploring caves to ensure safety.

The answer ignores the presence of the bat, which implies there must be a passage through which it entered, indicating the existence of a way out of the cave. Therefore, the answer is incorrect.

(1) ✗ There is a way out, but it is not easy. The way out is to fly up and out of the cave. However, the bat cannot fly because its wings are injured. The bat will need to find another way out of the cave.

The answer identifies that there is a way out. However, it indicates that the bat's wings are injured which is incorrect. Therefore, the answer is incorrect.

(2) ✓ Yes there is a way out. There is a bat hanging inside this cave. There has to be an opening that the bat flew in through, and the same opening would also be a way to get out.

The answer identifies the presence of the bat that is hanging from the ceiling. Which suggests that there is a way out. Therefore, the answer is correct.

Figure 14: Example of guidelines for MTurk annotators evaluating different answers to the visual riddle. (1) Annotator UI page displaying the visual riddle, the correct answer, and the candidate answers. (2) Solution: Identification of which candidate answers correctly solve the visual riddle. (3) Explanation & Notes: Detailed reasoning for the correctness or incorrectness of each candidate answer (only explanations for three candidates are shown due to space constraints).

A.7 Multiple Choice VQA Analysis: Model Hesitation

As mentioned in §5.3, a notable occurrence was the frequent selection of the "cannot determine" option by some models. Further analysis excluding these instances revealed improvements in comparison to the results with this option (see Table 2), with GPT-4 achieving 52% accuracy instead of 45%, and Gemini-Pro-1.5 reaching 48% accuracy instead of 38%, respectively. These results suggest that **some models hesitate to answer certain questions, opting for the "cannot determine" option when lacking sufficient information to decide**. Additionally, as shown in Table 11, providing models with auxiliary information via hints and attributions reduces their likelihood of selecting the "cannot

determine" option. For example, Gemini-Pro-1.5 achieves an accuracy rate of 74% with a hint, while GPT-4 achieves an accuracy rate of 84% with an attribution.

Table 11: Percentage of "Cannot Determine" Answers

	% Cannot Determine	+ Hint % Cannot Determine	+ Attribution % Cannot Determine
Gemini Pro 1.5	20	12	9
Gemini-Pro-Vision	3	2	
GPT4	12	3	3
LLaVA-1.6-34B	8	10	
LLaVA-1.5-7B	0	0	
Claude 3.5 Sonnet	4	1	
GPT4o	17	4	
Qwen-VL-Max	3	1	
Molmo-7B	1	1	

Table 12: Accuracies Excluding "Cannot Determine" Answer

	% Accuracy w/o Cannot Determine	+ Hint % Accuracy w/o Cannot Determine	+ Attribution % Accuracy w/o Cannot Determine
Gemini Pro 1.5	48	74	79
Gemini-Pro-Vision	42	64	
GPT4	52	71	84
LLaVA-1.6-34B	26	34	
LLaVA-1.5-7B	17	29	
Claude 3.5 Sonnet	48	45	
GPT4o	67	87	
Qwen-VL-Max	37	53	
Molmo-7B	35	43	

A.8 The Visual Riddles Prompt Set: A Challenge for Text-to-Image Models

In this section, we present the complete breakdown of the models' performance in generating images that fit the visual-riddles prompts. Table 13 presents the full results, indicating that models struggle to generate images that include delicate hints hidden within them, with the best-generating model, SDXL-Turbo [48], creating only 15% of images that match the given prompt.

Table 13: Model success rates

	% Generation Success ↑
SD-1.4	7
SD-2.1	12
SDXL-LCM	12
SDXL	14
SDXL-Turbo	15

B Discussion any potential negative societal impacts Visual Riddles

While Visual Riddles offers a unique platform for enhancing visual reasoning and commonsense understanding, it also presents potential challenges that merit careful consideration. One concern is the inadvertent reinforcement of biases. Despite rigorous efforts to design inclusive and culturally neutral visual riddles, the possibility remains that some content might unintentionally reflect or amplify societal stereotypes. Moreover, the complexity of certain riddles could disadvantage users with specific cognitive or sensory impairments, thereby limiting their participation and representation in the research facilitated by this benchmark.

Another limitation is the reliance on automated evaluation methods. Such methods may not fully capture the depth of human reasoning and interpretation, potentially affecting the robustness and transparency of the evaluations. This might inadvertently prioritize certain types of reasoning over others, skewing the development and assessment of AI systems.

To address these issues, we are committed to a continuous review process involving diverse stakeholders to help identify and mitigate any biases or exclusions. This process will include refining the riddles and improving the evaluation methods to ensure they are as inclusive and representative as possible. Additionally, ongoing adjustments will be made to the evaluation protocols to enhance their ability to assess nuanced and complex responses, thereby ensuring that the benchmark remains a fair and effective tool for advancing AI research in an ethically responsible manner.

C Designer Consent

We acknowledge and extend our gratitude to all designers who contributed to the benchmark. The credit for each visual riddle is included as part of our dataset. All designers have consented to contribute their creations to this research.