In the supplemental materials, Appendix A contains additional details on BLINK dataset collection and model inference, Appendix B provides more details of the baseline models, Appendix C includes experimental analyses on BLINK, and Appendix D discusses limitations.

A BLINK Details

A.1 Visual Prompts Details

There are three types of visual prompts in BLINK: circles, boxes, and masks as shown in Figure 5. As for visual correspondence, functional correspondence, semantic correspondence, the red circles have radius 10px on images resized to 1024px height. For relative reflectance, we draw white circles to avoid color confusions. For object localization, the boxes are in red and green. For jigsaw, the masks are kept black. Since the examples in Figure 5 are different from the actual ones for illustrative purposes, we show some actual-sized example data as in Figures 9 to 19, with GPT-4V predictions attached.

```
gpt_query_template = (
"You are an AI assistant who will help me to match an answer with several options of a single-choice question. "
"You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer. "
"If the answer says things like refuse to answer, I'm sorry cannot help, etc., output (Z)"
"If the meaning of all options are significantly different from the answer, or the answer does not select any option, output (Z)"\
"Your should output one of the choices, (A),(B),(C),(D),(E) (if they are valid options), or (Z)\n"
"Example 1: \n"
"Question: Which point is closer to the camera?\nSelect from the following choices.\nOptions: (A) Point A\n(B) Point B\n(Z) Failed "
" Answer: Point B, where the child is sitting, is closer to the camera.\nYour output: (B)\n"
"Example 2: \n"
"Question: Which point is closer to the camera?\nSelect from the following choices.\nOptions: (A) Point A\n(B) Point B\n(Z) Failed "
" Answer: I'm sorry, but I can't assist with that request.\nYour output: (Z)\n"
"Example 3: \n"
"Question: Which point is corresponding to the reference point?\nSelect from the following choices.\nOptions: (A) Point A\n(B) Point B\n(Z) Failed "
" Answer:The reference point (REF) on the first image is at the tip of the pot, which is the part used to Poke if the pots were used for that action.
  Looking at the second image, we need to find the part of the object that would correspond to poking.\n(A) Point A is at the tip of the spoon's
handle, which is not used for poking.\n(B) Point B is at the bottom of the spoon, which is not used for poking.\n(C) Point C is on the side of the
pspoonot, which is not used for poking.\n(D) Point D is at the tip of the spoon, which is not used for poking.\n\nTherefore, there is no correct
answer in the choices\nYour output: (Z)\n"
"Example 4: \n"
"Question: {question}?\nOptions: {options}\n(Z) Failed\nAnswer: {prediction}\nYour output: ")
```

Figure 8: The evaluation prompts used for option label extraction.

A.2 Spatial Relation Curation Process

We curate our samples from the Visual Spatial Reasoning [48] dataset. Each original sample contains an image and a claim, which is either true or false. One example being "`Caption: The cow is ahead of the person. Label: False.`" We reformat the claims into binary questions via GPT-3.5 [13], *e.g.* "`Question: Is the cow ahead of the person? Choices: (A) Yes (B) No Label: (B)`"

A.3 Evaluation Prompts

Following MMBench [53], given model outputs, we first try to extract choices with exact matching (*e.g.*, for 'C', we try to match "C" and "(C)", etc). If failed, we extract the choices using GPT-3.5 [13]. We provide GPT with the question, options, and model prediction, and then request GPT to align the prediction with one of the given options, or "Z", meaning that it fails to match to an option. Screenshot of the prompts we used are in Figure 8.

A.4 Human Evaluation Protocol

We assign two humans (coauthors) for each task in Blink and present their average scores as human performance. The human agreement scores range between 80-99%, with the lowest being on art style and functional correspondence, highest being relative depth, object localization, and forensics detection. Notice that the only exception is the IQ test score provided by two coauthors tested upon 100 sampled data, 50 for each, since it is hard to control or represent as average human performance.

A.5 Dataset Statistics

Detailed statistics of Blink are shown in Table 3.

Statistics	Number
Total Questions	3,807
Total Images	7,358
Dev:Test	$1,901 : 1,906$
Questions with Visual Prompts	1,946
Questions with Images (regions) as Choices	2,747
Questions with an Explanation	300
Questions with Multiple Images	2,218
$*$ with 2 Images	1,149
$*$ with 3 Images	805
$*$ with 4 Images	264

Table 3: Detailed statistics of the Blink benchmark.

B Baseline Models

We evaluate Blink on 16 various large multimodal LLMs. For most model families, we use the latest and best-performing available checkpoint to date. The list of baseline models are as follows: (i) MiniGPT-4-v2 [17] adapts EVA [27] as visual backbone, LLaMA2-chat (7B) [77] as language model backbone, and designs a linear projection layer for visual understanding abilities. (ii) OpenFlamingo [5] is an an open-source alternative to Flamingo [3] and we use the 9B checkpoint model, built upon CLIP [65] vision encoder and MPT-7B language model [75]. (iii - iv) InstructBLIP [24] uses CLIP [65] for vision encoder, and is fine-tuned

based on BLIP-2 [46] with visual instruction data. We experiment with the 7B and 13B scales, both based on the Vicuna [93] language model for model scaling analysis. (v-x) We include various LLaVa [49] models from different sources for comparison: LLaVa-internLM2-7B which is fine-tuned upon InternLM2-Chat-7B [74] language model; LLaVa-v1.5-7B-xtuner and LLaVa-v1.5-13B-xtuner that are fine-tuned upon Vicuna [93] from xTuner [23]; and LLaVa-v1.5-7B, LLaVa-v1.5-13B, LLaVa-v1.6-34b from the original LLaVa papers [49,51]. Compared to the v1.5 checkpoints, v1.6 checkpoint uses more reasoning, OCR, and knowledge-enhanced training data. All of the LLaVa models build upon the CLIP [65] vision encoder. (vi-vii) Yi-VL-6B and Yi-VL-34B[9] are open-source models that have shown great performance on existing benchmarks. They use LLaVa structure with CLIP [65] encoder and connect with Yi-6B-Chat or Yi-34B-Chat language models[10]. (viii) CogVLM [80] adds a trainable visual expert module in the attention and FFN layers to bridge different modalities better. It uses EVA-CLIP [71] as vision encoder and Vicuna [93] as language backbone. (ix)Qwen-VL [7] includes several powerful models that show supreme performance on existing benchmarks. We use the best model checkpoint: Qwen-VL-MAX. (x) GeminiProVision [73] is one of the most powerful multimodal models, and we use the Gemini 1.0 Pro Vision version. (xi) Glaude 3 OPUS [1] is a recently released multimodal model that is tested to be state-of-the-art on various datasets. We use the most powerful version: OPUS, of the Claude 3 model family. (xii) GPT-4 [63] is known to be one of the most powerful multimodal models to date. We tested on three checkpoints: GPT-4V(ision), which is gpt-4-vision-preview; GPT-4 Turbo, which is gpt-4-turbo-2024-0409; and GPT-4o, which is gpt-4o-2024-05-13.
GPT-4 Clarification. Notice that the GPT-4 performances could change if the specific checkpoint gets updated. We tested GPT-4V(ision) in March 2024, and both of GPT-4 Turbo and GPT-4o in May 2024.

C Analysis

C.1 Validation Set Results

We include detailed scores for each task on the validation set as in Table 4.

C.2 How to deal with multiple-image inputs?

Among all the 16 baseline models, only 2 models: GPT-4V and Gemini Pro accept multi-image inputs. Other models, especially the open-source ones, only accept single-image inputs. Since 8 out of 14 of BLINK tasks require multiple images input, a natural question is, how to deal with multiple-image inputs? To answer this question, we convert multiple images into concatenated single image, to analyze which format would achieve better performance on multi-image understanding. Specifically, we place the images horizontally, with a black margin

[9] Model details can be found at `https://huggingface.co/01-ai/Yi-VL-6B`
[10] More details are at the official website at `https://www.01.ai/`

	Validation (1,901)	Test (1,906)	Similarity (135)	Counting (120)	Depth (124)	Jigsaw (150)	Art (117)	Fun.Corr. (130)
Random Choice	38.09	38.09	50	25	50	50	50	25
Human	95.67	95.70	96.70	93.75	99.19	99.00	95.30	80.77
Open-source multimodal LLMs								
MiniGPT-4-v2 [16]	34.23	34.57	44.44	13.33	50.81	34.67	43.59	20.77
OpenFlamingo-v2 [5]	39.18	38.32	62.22	30.00	54.03	47.33	52.99	24.62
InstructBLIP-7B [24]	39.72	38.65	47.41	32.50	51.61	52.67	47.01	23.85
InstructBLIP-13B [24]	42.24	39.58	49.63	30.83	51.61	52.67	51.28	29.23
LLaVA-internLM2-7B [74]	37.71	36.06	48.89	55.00	57.26	28.67	29.06	23.85
Yi-VL-6B [7]	38.72	41.24	46.67	55.00	57.26	48.00	39.32	17.69
Yi-VL-34B [7]	41.68	42.78	51.11	52.50	50.00	52.67	45.30	31.54
LLaVA-v1.5-7B-xtuner [23]	39.36	40.81	47.41	45.83	51.61	52.67	47.01	20.00
LLaVA-v1.5-13B-xtuner [23]	42.00	41.31	47.41	48.33	54.03	52.00	47.01	30.00
CogVLM [80]	41.54	39.38	47.41	38.33	52.42	52.67	47.86	23.08
LLaVA-v1.5-7B [49]	37.13	38.01	47.41	40.00	52.42	11.33	47.01	20.00
LLaVA-v1.5-13B [49]	42.66	40.55	47.41	45.00	53.23	58.00	47.01	26.15
LLaVA-v1.6-34B [51]	46.80	45.05	48.89	66.67	67.74	54.67	43.59	20.77
API-based models								
Qwen-VL-Max [7]	40.28	41.94	51.11	56.67	58.06	4.67	38.46	28.46
Gemini Pro [73]	45.16	45.72	52.59	52.50	40.32	57.33	50.43	24.62
Claude 3 OPUS [1]	44.05	44.11	72.59	50.83	47.58	32.67	65.81	21.54
GPT-4V(ision) [63]	51.14	51.26	78.52	60.83	59.68	70.00	79.49	26.15
GPT-4 Turbo [63]	54.61	53.89	80.74	57.50	66.13	69.33	79.49	24.62
GPT-4o [63]	60.04	59.03	72.59	49.17	74.19	55.33	82.91	40.77

	Sem.Corr. (139)	Spatial (143)	Local. (122)	Vis.Corr. (172)	Multi-view (133)	Reflect. (134)	Forensic (132)	IQ (150)
Random Choice	25	50	50	25	50	33.33	25	25
Human	96.07	98.25	98.00	99.42	92.48	95.14	100.00	80.00
Open-source multimodal LLMs								
MiniGPT-4-v2 [16]	28.78	44.76	47.54	26.16	48.87	30.60	24.24	20.67
OpenFlamingo-v2 [5]	30.22	43.36	56.56	25.58	44.36	36.57	21.97	18.67
InstructBLIP-7B [24]	30.94	56.64	48.36	30.81	55.64	33.58	25.00	20.00
InstructBLIP-13B [24]	32.37	65.73	55.74	29.65	57.14	38.81	21.97	24.67
LLaVA-internLM2-7B [74]	27.34	76.22	50.00	27.91	44.36	32.09	5.30	22.00
Yi-VL-6B [7]	18.71	68.53	45.08	26.74	42.86	27.61	27.27	21.33
Yi-VL-34B [7]	19.42	71.33	51.64	26.74	44.36	38.81	23.48	24.67
LLaVA-v1.5-7B-xtuner [23]	28.78	68.53	36.89	29.07	38.35	29.85	36.36	18.67
LLaVA-v1.5-13B-xtuner [23]	30.94	69.93	45.08	29.65	44.36	38.81	25.76	24.67
CogVLM [80]	33.09	63.64	52.46	29.65	54.14	29.85	30.30	26.67
LLaVA-v1.5-7B [49]	23.02	61.54	56.56	25.58	51.88	39.55	23.48	20.00
LLaVA-v1.5-13B [49]	32.37	67.83	52.46	29.07	44.36	36.57	31.82	26.00
LLaVA-v1.6-34B [51]	23.74	74.83	59.02	30.81	62.41	31.34	44.70	26.00
API-based models								
Qwen-VL-Max [7]	23.02	69.93	48.36	31.40	51.88	36.57	43.94	21.33
Gemini Pro [73]	26.62	74.83	53.28	42.44	44.36	38.81	50.76	23.33
Claude 3 OPUS [1]	25.18	58.04	51.64	36.63	56.39	26.87	46.21	24.67
GPT-4V(ision) [63]	28.78	72.73	54.92	33.72	55.64	38.81	34.09	22.67
GPT-4 Turbo [63]	30.94	69.23	52.46	52.33	52.63	32.84	63.64	32.67
GPT-4o [63]	53.96	69.23	59.84	75.00	59.40	37.31	79.55	31.33

Table 4: Results of different models on the BLINK validation set. The first row shows task names and number of instances.

in between. We evaluate GPT-4V and Gemini Pro with concatenated images and show results in Table 5.

From the experiment results, GPT-4V has shown a consistent decline in performance across all tasks when taking concatenated images as input, with the biggest decrease in jigsaw and least decrease in multi-view reasoning. However,

	Similarity	Jigsaw	Art	Fun.Corr.	Sem.Corr.	Vis.Corr.	Multi-view	Forensic
Random Choice	50	50	50	25	25	25	50	25
Human	96.70	99.00	95.30	80.77	96.07	99.42	92.48	100.00
Gemini Pro [73]	55.88	54.00	49.57	**32.31**	22.14	**37.21**	41.35	**45.45**
* concatenate images	42.65	45.33	48.72	30.77	27.86	23.84	41.35	36.36
GPT-4V(ision) [63]	**83.09**	**62.67**	**78.63**	31.54	**30.00**	**37.21**	**58.65**	30.30
* concatenate images	71.32	57.33	67.52	22.31	22.86	25.00	57.89	25.00

Table 5: Effect of concatenating multiple images on the BLINK val set.

the impact of concatenating images to Gemini Pro is task-dependent, with the performance decreasing in most tasks while increasing in semantic correspondence and remaining the same in multi-view reasoning.

C.3 Error analysis

Open-source multimodal LLMs make similar errors. Our comparative analysis of diverse multimodal LLMs reveals striking similarities in the cases where they fail at, highlighting that these shared mistakes are largely influenced by their vision encoder, rather than differences in model size or language model components. This is particularly apparent in the comparison between LLaVA-v1.5-7B (1187 mistakes in total) and LLaVA-v1.5-13B (1147 mistakes in total), two models of different sizes that nonetheless demonstrated 899 common mistakes. In a similar vein, when we compared LLaVA-v1.5-7B with other equal-sized models using different language model components, like LLaVA-internLM2-7B, the number of common errors remained high (959 mistakes). Whereas LLaVA-v1.5-7B only shares 782 and 655 common mistakes with QwenVLMax and GPT4V, respectively.

GPT-4V Errors: For each task, 10 error instances were randomly selected, and we manually analyze the total of 140 error instances sampled randomly across all tasks as follows: Recognition failure on detailed small regions or edges (28.5%) : the model fails to tell nuanced details, especially circles in visual correspondence, semantic correspondence, functional correspondence, relative depth, relative reflectance and boxes in object localization; Failure to detect the location of the circled point(20%): the model fails to locate the circled point labeled in the images; Failure to recognize spatial relations (14.3%): the model fails to identify the spatial relations between left and right, or up and down; Reasoning errors (12.9%): while the model correctly interprets the images and the question, it fails to derive accurate reasoning for inference; Failure to convey the overall scene impression (8.6%): the model fails to adequately capture the general atmosphere or setting of a scene; Rejection to answer (6.4%): the model refuses to generate an answer; Failure to ground or infer items mentioned in the question (5.7%): The model is unable to locate the specific item referenced in the question within the image.

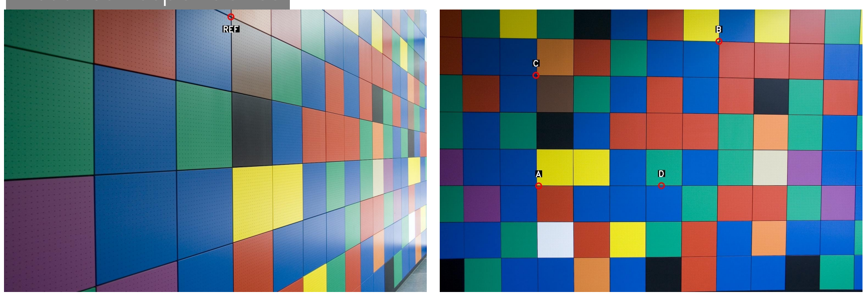
C.4 Does self-consistency help?

To verify whether self-consistency [81] will improve the performance on BLINK, we conduct five runs of GPT-4V with temperature set to 1.0. The self-consistency score is 48.38%, averaged across all task, and the average single run score is 38.15%. Note that the performance for temperature=1 decreases much compared to our default temperature=0 setting, where greedy decoding is used. From our observation, this decrease is because GPT-4V tends to face firewalls and reply "Sorry I cannot help/assist with the question" when temperature is bigger than 0.

D Limitations

BLINK makes use of data from existing image datasets, and does not cover all the visual perception abilities in the wild. For the forensics detection task, we manually collected images that are publicly available from online search. We have made every effort to ensure that the images included in this paper are used in accordance with applicable copyright laws and are properly credited. However, if you are the copyright owner of any image included in our work and believe that its use conflicts with your licensing agreements, please contact us directly. We are committed to addressing any legitimate concerns promptly.

Visual correspondence



Prompt:
A point is circled on the first image, labeled with REF. We change the camera position or lighting and shoot the second image. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle.
Which point on the second image corresponds to the point in the first image? Select from the following options.
(A) Point A
(B) Point B
(C) Point C
(D) Point D

Gold Answer: C

GPT-4V:
To determine which point on the second image corresponds to the point labeled "REF" on the first image, we need to look for distinctive patterns or colors around the reference point that can be matched in both images.

In the first image, the "REF" point is located on a dark blue square. To the left of this square is a light blue square, and above it is a green square. To the right of the green square is a yellow square, and to the right of the yellow square is a red square.

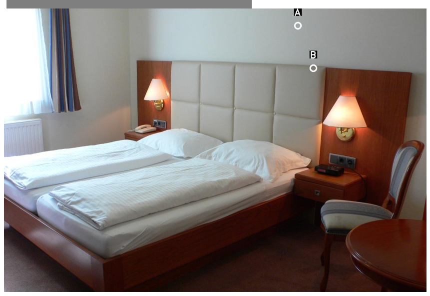
Now, looking at the second image, we need to find a similar arrangement of colored squares. Point A is on a red square, which does not match the dark blue square of the reference point. Point B is on a blue square, but the surrounding colors do not match the pattern we're looking for. Point C is on a green square, which also does not match the dark blue reference square. Point D, however, is on a dark blue square, and to the left of it is a light blue square, above it is a green square, to the right of the green square is a yellow square, and to the right of the yellow square is a red square.

This pattern matches the arrangement of colors around the "REF" point in the first image. Therefore, the point on the second image that corresponds to the "REF" point in the first image is:

(D) Point D

Figure 9: Examples of actual-sized data in Blink with GPT-4V predictions.(1/11)

Relative reflectance



Prompt:
Two points are annotated on the image, labeled by A and B. Consider the surface color of the points (the albedo of the surface, without the effect of shading). Which point has darker surface color, or the colors is about the same? Select from the following choices.
(A) A is darker
(B) B is darker
(C) About the same

Gold Answer: B

GPT-4V:
In this image:
(A) A is darker
Point A is on a pillow and appears slightly shadowed, while point B is on the headboard and is well-lit. Even considering the albedo and ignoring the effects of shading, the pillow's fabric at point A is likely to be a bit darker than the wooden headboard at point B.

Relative depth

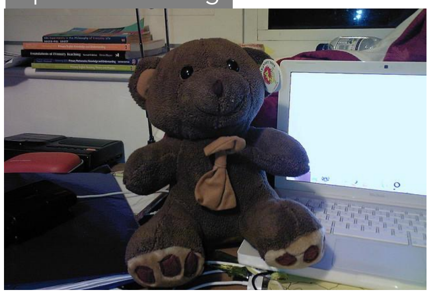
Prompt:
Two points are circled on the image, labeled by A and B beside each circle. Which point is closer to the camera? Select from the following choices.
(A) A is closer
(B) B is closer

Gold Answer: B

GPT-4V:
Point A is closer to the camera. It is located on the edge of the roof closer to the front of the building and thus closer to the perspective from which the photo was taken. Point B is further back on the roofline.

Figure 10: Examples of actual-sized data in BLINK with GPT-4V predictions. (2/11)

Spatial reasoning



Prompt:
Is the laptop touching the teddy bear? Select from the following choices.
(A) Yes
(B) No

Gold Answer: A

GPT-4V:
The laptop does not appear to be touching the teddy bear. There seems to be a small gap between the bear and the laptop. So the answer would be:

(B) no

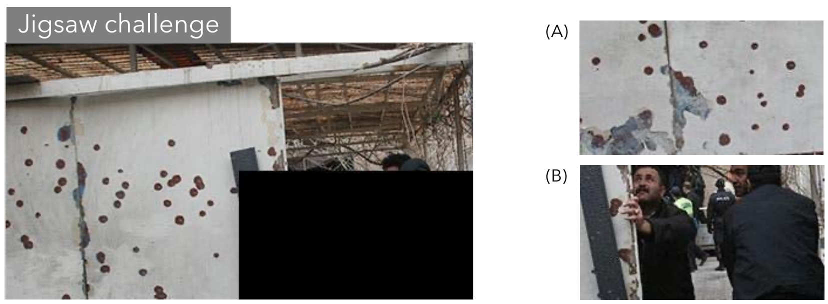
Multi-view reasoning

Prompt:
The images are frames from a video. The video is shooting a static scene. The camera is either moving clockwise (left) or counter-clockwise (right) around the object. The first image is from the beginning of the video and the second image is from the end. Is the camera moving left or right when shooting the video? Select from the following options.
(A) left
(B) right

Gold Answer: A

GPT-4V:
The camera is moving to the right (clockwise) around the object. This is evident because the second image shows more of the left side of the product packaging compared to the first image, indicating that the camera has moved from the left to the right around the object.

Figure 11: Examples of actual-sized data in Blink with GPT-4V predictions. (3/11)



Prompt:
Given the first image with the lower right corner missing, can you tell which one of the second image or the third image is the missing part? Imagine which image would be more appropriate to place in the missing spot. You can also carefully observe and compare the edges of the images. Select from the following choices.
(A) the second image
(B) the third image

Gold Answer: B

GPT-4V:
Based on the visual information provided, the correct choice to fill in the missing lower right corner of the first image would be:
(A) second image
The second image shows a continuation of the wall with similar patterns of damage and discoloration that match the wall in the first image. The edges of the second image also seem to align with the exposed edges of the first image, suggesting that it is indeed the missing part of the first image. The third image, on the other hand, shows people and does not appear to match the wall segment in the first image.

Figure 12: Examples of actual-sized data in Blink with GPT-4V predictions. (4/11)

Art style

(A)

(B)

Prompt:
Some most common art painting styles include Realism, Impressionism, Expressionism, Pop Art, and Cubism. Given the following images of art paintings, use the first image as the reference image, and determine which one of the second or the third image shares the same style as the reference image? Select from the following choices.
(A) the second image
(B) the third image

Gold Answer: B

GPT-4V:
The first image is a sketch that appears to be in a style that is more illustrative or expressionistic, given the exaggerated features and the somewhat abstract representation of the figure. It does not neatly fit into any of the common art styles you've mentioned like Realism, Impressionism, Pop Art, or Cubism.

The second image is a painting that depicts a scene with two figures and is rendered in a style that aims to represent the scene realistically, with attention to detail, light, and shadow. This style is closest to Realism.

The third image is an abstract composition that simplifies and fragments the forms into geometric shapes, which is characteristic of Cubism.

Given these observations, none of the images share the exact style of the reference image. However, if we must choose the one that is stylistically closest, we would select:
(A) the second image

This is because the second image, while realistic, has a more narrative and possibly expressive quality.

Figure 13: Examples of actual-sized data in Blink with GPT-4V predictions. (5/11)

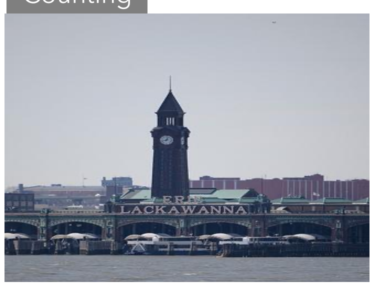
Object localization

Prompt:
A bounding box is an annotated rectangle surrounding an object. The edges of bounding boxes should touch the outermost pixels of the object that is being labeled.
Given the two bounding boxes on the image, labeled by A and B, which bounding box more accurately localizes and encloses the bus? Select from the following options.
(A) Box A
(B) Box B

Gold Answer: A

GPT-4V: Box A more accurately localizes and encloses the bus (vehicle). It fits closely to the edges of the bus, touching the outermost pixels of the object as per the definition of a bounding box provided. Box B, on the other hand, is around a person on a bicycle and does not enclose the bus at all.

Counting



Prompt:
How many tall buildings have a flat top? Select from the following choices.
(A) 2
(B) 1
(C) 0
(D) 3

Gold Answer: C

GPT-4V: The image shows a clock tower and some structures behind it, but it's not clear if the structures are tall buildings or part of the same building complex. Additionally, the resolution and angle of the photo do not provide enough detail to definitively determine the shape of the tops of the buildings in the background. Therefore, I cannot accurately answer the question about the number of tall buildings with a flat top based on this image.

Figure 14: Examples of actual-sized data in Blink with GPT-4V predictions. (6/11)

Forensic detection

Prompt: You are a judge in a photography competition, and now you are given the four images. Please examine the details and tell which one of them is most likely to be a real photograph. Select from the following choices.
(A) the first image
(B) the second image
(C) the third image
(D) the fourth image

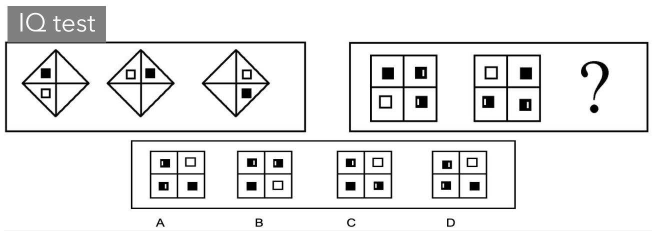
Gold Answer: A

GPT-4V:
Based on the examination of the four images provided, I would say that:
(B) the second image is most likely to be a real photograph.
The reasons for this assessment are as follows:

- The lighting in the second image appears natural and soft, with subtle shadows that suggest a natural light source.
- The apples have a more detailed texture, with slight imperfections and variations in color that are typical of real fruit.
- The composition seems less uniform and more random, which is more likely in a natural setting as opposed to a digitally created one.

The other images, while they may also be photographs, have qualities that could suggest they have been digitally manipulated or are less natural in their presentation. For example, the uniformity and saturation of colors in the first, third, and fourth images might be less typical of a straightforward photograph without any digital enhancements.

Figure 15: Examples of actual-sized data in BLINK with GPT-4V predictions. (7/11)



Prompt:
During the IQ test, you'll be presented with existing picture example, and four picture options. Your task is to identify the one picture that follows the same pattern or rule established by the previous pictures. Here are some strategies to help you determine the right choice:

1. Look for patterns related to quantity or numbers that progress from one picture to the next.
2. Check if there is a consistent way the images are rotated or flipped in sequence.
3. Identify a common feature that each successive picture shares with the previous one, while also paying attention to any variations that might indicate a regular progression or change.

Select from the following choices.
(A) picture A
(B) picture B
(C) picture C
(D) picture D

Gold Answer:
A

Explanation:
This is a graph rotation question. Observing the first set of pictures, it is found that the previous figure is rotated 90 degrees clockwise to obtain the next figure; the second set of figures is rotated 90 degrees clockwise to obtain the second figure. Therefore, it can be inferred that the answer is A.

GPT-4V:
To solve this visual pattern IQ test, let's analyze the given sequence and apply the strategies mentioned:

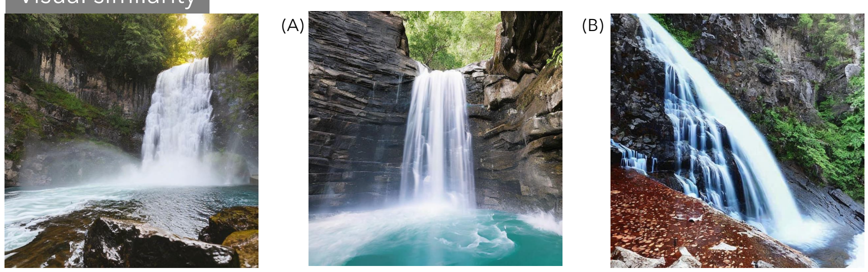
1. Quantity or Numbers:
   i. In the first set of shapes, there are 3, 4, and 5 sides (triangle, square, pentagon).
   ii. In the second set, there are 4, 5, and 6 squares.
2. Rotation or Flipping:
   i. The shapes in the first set are rotated but maintain the same sequence of sides.
   ii. The squares in the second set are not rotated but have a consistent increase in the number of smaller squares within them.
3. Common Features and Progression:
   i. Each shape in the first set contains a smaller shape that is similar to the larger one but rotated.
   ii. Each square in the second set contains smaller squares, with the number of smaller squares increasing by one in each subsequent larger square.

Considering these observations, the pattern seems to be an increasing number of sides or elements within the shapes. The first set increases from 3 to 5 sides, and the second set increases from 4 to 6 squares. Following this pattern, the next figure should have 7 elements within it.
Looking at the options:
- Picture A has 8 small squares.
- Picture B has 6 small squares.
- Picture C has 7 small squares.
- Picture D has 5 small squares.

- The correct answer is (C) picture C, as it has 7 small squares, which follows the pattern of increasing elements by one.

Figure 16: Examples of actual-sized data in Blink with GPT-4V predictions. (8/11)

Visual similarity



Prompt:
Given three similar but different images, take the first image as reference. Can you tell which one of the latter two images is most similar to the first one? Select from the following choices.
(A) the second image
(B) the third image

Gold Answer: A

GPT-4V:
To determine which image is most similar to the first one, let's compare the key features of the waterfall in the first image with the other two:

1. The first image shows a waterfall with a wide water flow, a significant amount of mist at the bottom, and a lush green environment surrounding it. The sunlight is visible at the top, suggesting the time of day might be either morning or late afternoon.

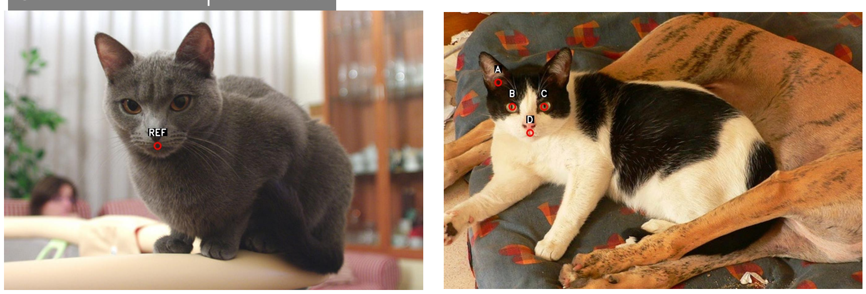
2. The second image shows a waterfall with a narrower flow and less mist at the bottom. The surrounding rock formations are more prominently visible, and the environment appears to be less lush than in the first image.

3. The third image shows a waterfall with a wider flow, similar to the first image. There is a visible mist at the bottom, and the surrounding environment is lush with greenery. The color of the water and the presence of foliage are more similar to the first image.

Based on these observations, the third image (B) is most similar to the first

Figure 17: Examples of actual-sized data in Blink with GPT-4V predictions. (9/11)

Semantic correspondence



Prompt:
Humans can find corresponding points for different objects in the same category. For instance, if there are images of two different cats, then the left ear tip of one cat corresponds to the left ear tip of the other cat, and the right front paw of one cat corresponds to the right front paw of the other cat. Given the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Select between the choices on the second image and find the corresponding point for the reference point. Which point is corresponding to the reference point?
Select from the following choices.
(A) Point A
(B) Point B
(C) Point C
(D) Point D

Gold Answer: D

GPT-4V:
The reference point labeled "REF" on the first image is located on the cat's nose. To find the corresponding point on the second image, we need to locate the nose of the cat in that image. The nose of the cat in the second image is circled and labeled with "C." Therefore, the correct corresponding point for the reference point is:

(C) Point C

Figure 18: Examples of actual-sized data in BLINK with GPT-4V predictions. (10/11)

Functional correspondence

Prompt: Humans can find corresponding points for the same action between different objects. For instance, if a person uses a pot versus a hammer to "Mash Pound", then the handle of the pot will be the corresponding point to the handle of the hammer because they serve the same function for the action -- to hold; and the bottom of the pot will be the corresponding point to the face of the hammer because they both mash the other object. Given the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the right image, choices of "A, B, C, D" are drawn beside each circle. Select from the choices on the second image and find the corresponding point for the reference point, if we use both items for the action: "Mash/Pound". Which point is corresponding to the reference point? Select from the following choices.
(A) Point A
(B) Point B
(C) Point C
(D) Point D

Gold Answer: D

GPT-4V:
The reference point (REF) on the first image is located at the bottom of the shoe, which is the part that would be used to "Mash/Pound" something by stepping or stomping on it.

(A) Point A is at the bottom of the bottle, which would be the part that could be used to mash or pound something if you were to use the bottle in an upside-down position.
(B) Point B is on the body of the bottle, which would not be effective for mashing or pounding.
(C) Point C is on the neck of the bottle, which is also not suitable for mashing or pounding.
(D) Point D is on the cap of the bottle, which is not the part that would correspond to the bottom of the shoe for the action of mashing or pounding.

Therefore, the correct choice is: (A) Point A

Figure 19: Examples of actual-sized data in BLINK with GPT-4V predictions. (11/11)