Object Inclusion

Instruction: Generate an image of a crowded beach. Please include a single snowman in the image.

Good Example: FLUX 1.1 PRO



Evaluation: ✔

Bad Example: IMAGEN 3



Evaluation: ✘
Analysis: It's not a snowman but a "sand-man", affected by its context of a beach.

Figure 5: Examples for the task: Object Inclusion

Object Exclusion

Instruction: Generate an image of a birthday party. Do not include cakes in the image.

Good Example: IDEOGRAM 2



Evaluation: ✔

Bad Example: DALLE 3



Evaluation: ✘
Analysis: It fails to exclude cakes in the context of a birthday party.

Figure 6: Examples for the task: Object Exclusion

Object Count

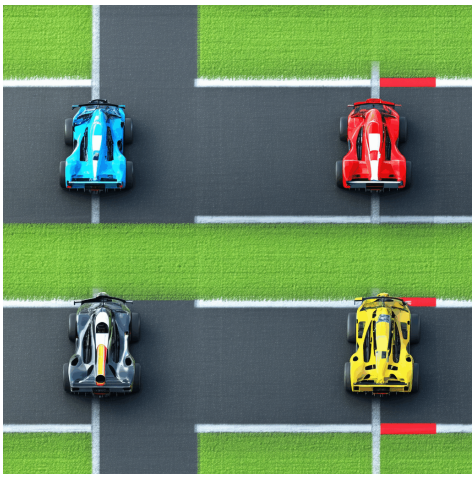
Instruction: Generate an image of a race track with 3 clearly visible race cars, evenly spaced and not overlapping. The race cars should be of regular size and common shapes.

Good Example: RECRAFT V3



Evaluation: ✔

Bad Example: SD 3.5



Evaluation: ✘
Analysis: Neither the number of tracks nor the number fo cars is correct.

Figure 7: Examples for the task: Object Count

Object Reasoning

Instruction: Generate an image of a single fruit named after a flightless bird native to New Zealand. Do not add additional elements or details in the background.

Good Example: IMAGEN 3



Evaluation: ✔

Bad Example: SD 3.5



Evaluation: ✘
Analysis: Wrong object.

Figure 8: Examples for the task: Object Reasoning

Object Attribution

Instruction: Generate an image of a single red giraffe with green stripes.

Good Example: IMAGEN 3



Evaluation: ✔

Bad Example: DALLE 3



Evaluation: ✘
Analysis: The giraffe has green legs instead of stripes.

Figure 9: Examples for the task: Object Attribution

Comparison Relation

Instruction: Generate an image of only one nail that is longer than only one snake.

Good Example: IDEOGRAM 2



Evaluation: ✔

Bad Example: GEMINI IMAGE



Evaluation: ✘
Analysis: The snake is longer than the nail.

Figure 10: Examples for the task: Comparison Relation

Universal Relation

Instruction: Generate an image of a toy box where all toys are animals except one car.

Good Example: GPT IMAGE



Evaluation: ✔

Bad Example: RECRAFT V3



Evaluation: ✘
Analysis: Have two cars instead of one.

Bad Example: FLUX 1.1 PRO



Evaluation: ✘
Analysis: Have building blocks which are not allowed in the instruction.

Figure 11: Examples for the task: Universal Relation

Absolute Spatial Relation

Instruction: Generate an image of a countryside porch, with a single rocking chair at the bottom left quarter of the image and a single lantern at the up right quarter of the image.

Good Example: GEMINI IMAGE



Evaluation: ✔

Bad Example: DALLE 3



Evaluation: ✘
Analysis: The number of lanterns is two. Generating wrong number of objects is a prevelant failure mode for all image generation tasks other than obejct count task.

Bad Example: SD 3.5



Evaluation: ✘
Analysis: The latten is at the left-up quarter and the chair at bottom-right.

Figure 12: Examples for the task: Absolute Spatial Relation

Relative Spatial Relation

Instruction: Generate an image about a quiet library. Please include a single wooden chair and and a single lamp. Make sure that the wooden chair is to the right of the lamp.

Good Example: DALLE 3



Evaluation: ✔

Bad Example: SD 3.5



Evaluation: ✘
Analysis: The wooden chair is to the left of the lamb.

Figure 13: Examples for the task: Relative Spatial Relation

Region Fill

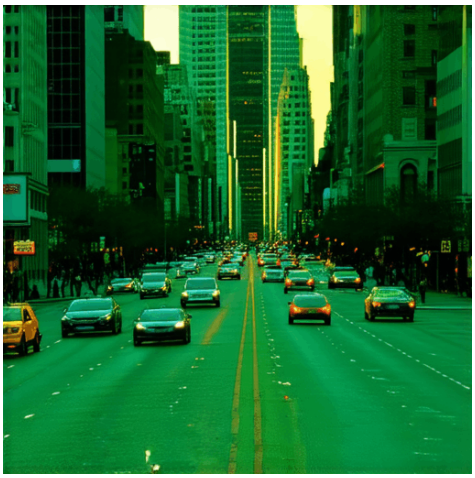
Instruction: Generate an image about a bustling city street at dusk. The left half should be pure green.

Good Example: GPT IMAGE



Evaluation: 0.892 (similarity score)

Bad Example: SD 3.5



Evaluation: 0.327 (similarity score)
Analysis: The whole image is green.

Figure 14: Examples for the task: Region Fill

Border Fill

Instruction: Generate a serene mountain landscape at sunset. The entire image should be surrounded by a simple and flat, solid and green border and approximately 10% of the image width on all sides.

Good Example: GPT IMAGE



Evaluation: 0.859 (similarity score)

Bad Example: GEMINI IMAGE



Evaluation: 0.0 (similarity score)
Analysis: The border is too wide

Bad Example: IDEOGRAM 2



Evaluation: 0.0 (similarity score)
Analysis: There is no border.

Figure 15: Examples for the task: Border Fill

Single Text Rendering

Instruction: Generate an image of a sign and the only text on it is “Love The World Forever”. The text should be written in a clear, standard, easily readable typeface without any artistic distortions.

Good Example: RECRAFT V3



Evaluation: ✔

Bad Example: DALLE 3



Evaluation: ✘
Analysis: The text rendered (“love the world forevver”) is wrong.

Bad Example: IDEOGRAM 2



Evaluation: ✘
Analysis: Artistic distortion makes it hard to recognize.

Figure 16: Examples for the task: Single Text Rendering

Double Text Rendering

Instruction: Generate an image of exactly one sticky note on the left side of a white desk and exactly one coffee mug on the right. The sticky note should have "Stay bold, chase dreams" written on it while the coffee mug should have "Coffee fuels infinite morning possibilities" printed on it. The text should be written in a clear, standard, easily readable typeface without any artistic distortions. The objects should be clearly visible and avoid overlapping.

Good Example: GPT IMAGE



Evaluation: ✔

Bad Example: RECRAFT V3



Evaluation: ✘
Analysis: The text is not rendered on the coffee mug.

Bad Example: IMAGEN 3



Evaluation: ✘
Analysis: The text rendered on the coffee mug is wrong.

Bad Example: RECRAFT V3



Evaluation: ✘
Analysis: The coffee mug is to the left of the sticky note.

Figure 17: Examples for the task: Double Text Rendering

Multi-Lingual Text Rendering

Instruction: Generate an image of a mug and the only text on it is "美好". The text should be written in a clear, standard, easily readable typeface without any artistic distortions.

Good Example: GPT IMAGE



Evaluation: ✔

Bad Example: RECRAFT V3



Evaluation: ✘

Analysis: Distorted Chinese-like characters. In fact, all other models fails in generating Chinese except GPT IMAGE and GEMINI IMAGE.

Figure 18: Examples for the task: Multi-Lingual Text Rendering

Semantic Consistency

Instruction: Create four images of traditional dice faces, each displaying a single face with an increasing number of pips, starting from two up to the highest number.

Good Example: GPT-4O + GPT IMAGE
Response: Here are the images of traditional dice faces, each displaying a single face with an increasing number of pips:

1. Two pips:



2. Three pips:



3. Four pips:

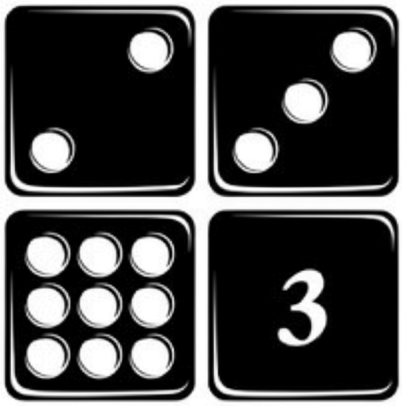


2. Five pips:



Evaluation: ✔

Bad Example: GEMINI IMAGE
Response: I will generate four images showing individual faces of a standard six-sided die. The first image will show the '2' face, the second will show the '3' face, the third will show the '4' face, and the final image will show the '5' face. Each face will be depicted with the traditional arrangement of white pips on a black background.



Evaluation: ✘
Analysis: Generate only one image; pip numbers are inconsistent with textual response; the last face is a number rather than pips.

Figure 19: Examples for the task: Semantic Consistency

Multi-Angle Consistency

Instruction: Using the provided image as the reference angle, create four additional images of the object from these specific angles: 30 degrees to the right, 60 degrees to the right, 30 degrees to the left, and 60 degrees to the left.



Unfortunately all the models we evaluate perform poorly on this task. Here we will demonstrate our gold reference images.

Reference Answer:

1. 30 degrees to the right:



2. 60 degrees to the right:



3. 30 degrees to the left:



4. 60 degrees to the left:



Bad Example: GEMINI IMAGE

Response: Generating an image of the floral patterned seat cushion rotated 30 degrees to the right, showing more of its right side and slightly obscuring the left.



Generating an image of the floral patterned seat cushion rotated 60 degrees to the right, revealing much of its right side and significantly obscuring the left.



Generating an image of the floral patterned seat cushion rotated 30 degrees to the left, showing more of its left side and slightly obscuring the right.



Generating an image of the floral patterned seat cushion rotated 60 degrees to the left, revealing much of its left side and significantly obscuring the right.



Evaluation: [0.184, 0.401, 0.321, 0.320] (image similarity)

Analysis: Fail to rotate the object correctly at the specific angles.

Figure 20: Examples for the task: Multi-Angle Consistency

Multi-View Consistency

Instruction: Using the provided image as a reference for the frontal view, create four additional images from the following angles, in order: 15 degrees to the right, 60 degrees to the left, 30 degrees to the right, and 45 degrees to the left.



Unfortunately all the models we evaluate perform poorly on this task. Here we will demonstrate our gold reference images.

Reference Answer:

1. 15 degrees to the right:



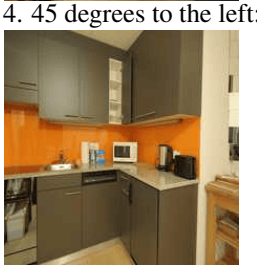
2. 60 degrees to the left:



3. 30 degrees to the right:



4. 45 degrees to the left:



Bad Example: GEMINI 2.5 + GPT IMAGE

Response: Okay, here are the four images of the kitchen from the specified angles, using your provided image as the frontal view reference.

First, here is the kitchen viewed from 15 degrees to the right:



Next, the kitchen viewed from 60 degrees to the left:



Then, the kitchen viewed from 30 degrees to the right:



And finally, the kitchen viewed from 45 degrees to the left:



Evaluation: [0.302, 0.304, 0.299, 0.351] (image similarity)

Analysis: Fail to rotate the scene correctly at the specific angles

Figure 21: Examples for the task: Multi-View Consistency

Composition Consistency

Instruction: Create three images that sequentially display the addition of an orange, a banana, and a kiwi into an initially empty fruit bowl. Ensure that previously added fruits remain in the bowl.

Good Example: GPT-4O + GPT IMAGE
Response: Here are the images showing the sequential addition of fruits into the bowl:
1. Empty Fruit Bowl with an Orange Added:



2. Fruit Bowl with an Orange and a Banana Added:



3. Fruit Bowl with an Orange, a Banana, and a Kiwi Added:



These images will illustrate the gradual addition of each fruit into the bowl.
Evaluation: ✔

Bad Example: GEMINI IMAGE
Response: Image 1: An empty, clear glass fruit bowl sits on a light wooden surface, ready to be filled.



Image 2: A vibrant orange has been placed in the clear glass fruit bowl, resting slightly off-center. The light wooden surface is visible beneath and around the bowl.



Image 3: A bright yellow banana has been added to the bowl, curving gently alongside the orange. Both fruits are now visible inside the clear glass bowl on the light wooden surface.

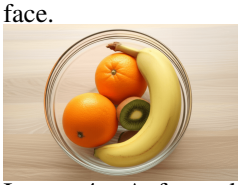


Image 4: A fuzzy brown kiwi has been placed in the bowl with the orange and banana. The three fruits now sit together in the clear glass bowl on the light wooden surface.



Evaluation: ✘
Analysis: Fail to follow the required order of adding objects and generate incorrect number of images.

Figure 22: Examples for the task: Composition Consistency

Decomposition Consistency

Instruction: Create three images that sequentially show the result after removing the kiwi, then the banana, and finally the orange from the original image. Do not replace the removed objects.



Good Example: GEMINI 2.5+GPT IMAGE
Response: Okay, here are the images showing the sequential removal of the fruits.
First, here is the image with the kiwi removed:



Next, here is the image with the banana also removed:

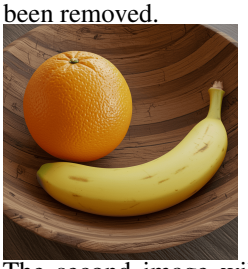


Finally, here is the image with the orange also removed, leaving an empty bowl:

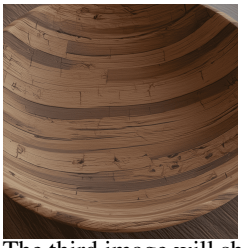


Evaluation: ✔

Bad Example: GEMINI IMAGE
Response: I will generate three images. The first image will show the wooden bowl with the orange and banana, after the kiwi has been removed.



The second image will show the wooden bowl with only the orange, after the banana has been removed from the previous image.



The third image will show the empty wooden bowl, after the orange has been removed from the previous image.

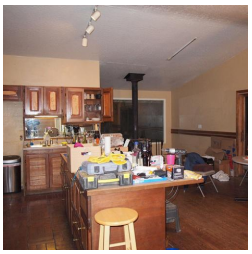


Evaluation: ✘
Analysis: Fail to follow the sequence of removing objects in the second image.

Figure 23: Examples for the task: Decomposition Consistency

Interleaved Image Editing - Object Adding

Instruction: Create an image displaying the result after placing the object <image_0> from <image_1>.

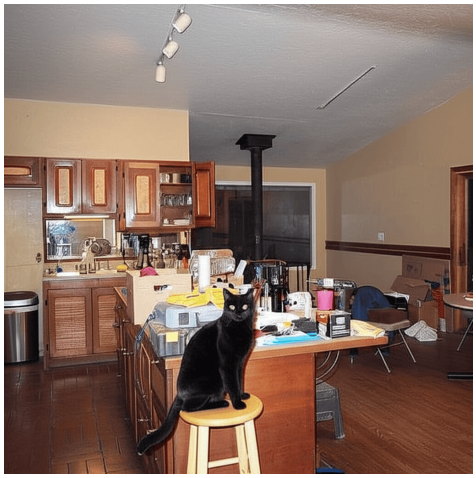


<image_0>



<image_1>

Good Example: GEMINI IMAGE



Evaluation: 0.770 (image similarity)

Bad Example: GEMINI 2.5 + GPT IMAGE

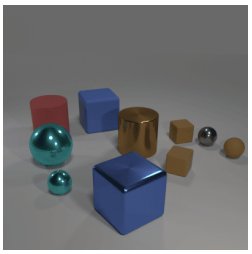


Evaluation: 0.342 (image similarity)
Analysis: The position of the cat is wrong and unreasonable. This is a common failure pattern among all the models.

Figure 24: Examples for the task: Interleaved Image Editing - Object Adding

Interleaved Image Editing - Color Modifying

Instruction: Generate an image that shows the result after changing the color of the largest ball in <image_0> to <image_1>.

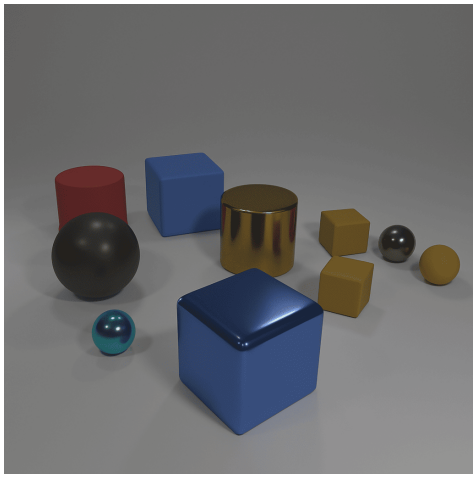


<image_0>



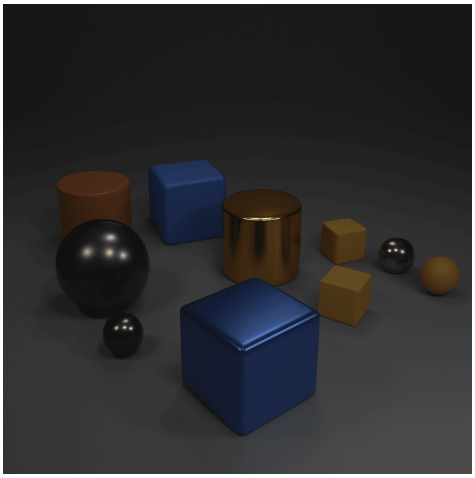
<image_1>

Good Example: GPT-4O + GPT IMAGE



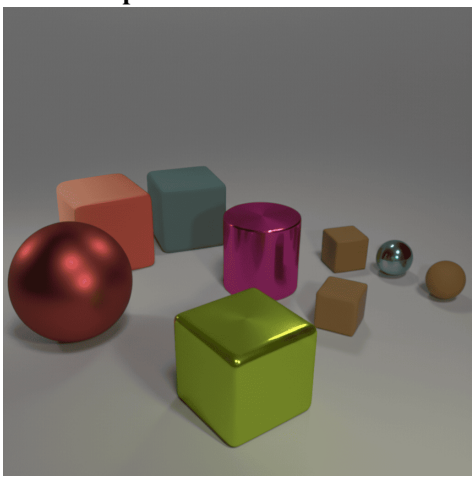
Evaluation: 0.722 (image similarity)

Bad Example: GEMINI 2.5 + GPT IMAGE



Evaluation: 0.483 (image similarity)
Analysis: The colors of background and some other objects are changed.

Bad Example: GEMINI IMAGE

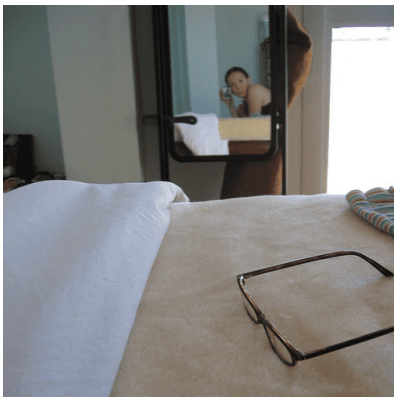


Evaluation: 0.608 (image similarity)
Analysis: The colors and shapes of many objects are changed.

Figure 25: Examples for the task: Interleaved Image Editing - Color Modifying

Image Editing - Text Editing

Instruction: Create an image displaying the result after inserting the word "clean" onto the white linen closest to the viewer, while leaving the rest of the image untouched.

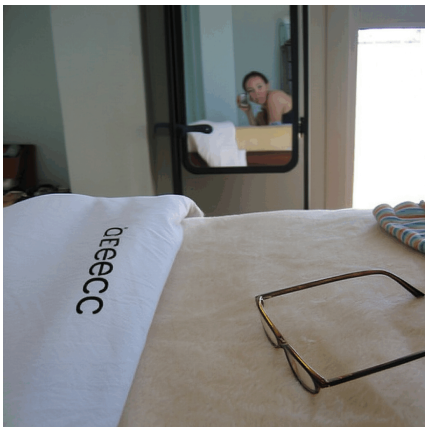


Good Example: GEMINI 2.5 + GPT IMAGE



Evaluation: ✔, 0.786 (image similarity)

Bad Example: GEMINI IMAGE



Evaluation: ✘
Analysis: The text rendered is wrong.

Bad Example: GEMINI 2.5 + IMAGEN 3



Evaluation: ✘
Analysis: The original image is changed. This is inevitable if using a text-to-image model incapable of image editing.

Figure 26: Examples for the task: Image Editing - Text Editing

Image Editing - Object Adding

Instruction: Create an image displaying the result after placing a blue sticker on the meter, with all other elements in the image remaining unchanged.



Good Example: GEMINI IMAGE



Evaluation: ✔, 0.911 (image similarity)

Bad Example: GPT-4O + GPT IMAGE



Evaluation: ✘
Analysis: No obviously visible blue sticker is added.

Figure 27: Examples for the task: Image Editing - Object Adding

Image Editing - Object Removing

Instruction: Create an image that displays the result after removing the man's wig, while leaving all other elements of the image unchanged.



Good Example: GEMINI 2.5 + GPT IMAGE



Evaluation: ✔, 0.630(image similarity)

Bad Example: GEMINI IMAGE



Evaluation: ✘
Analysis: The wrong object (head) is removed.

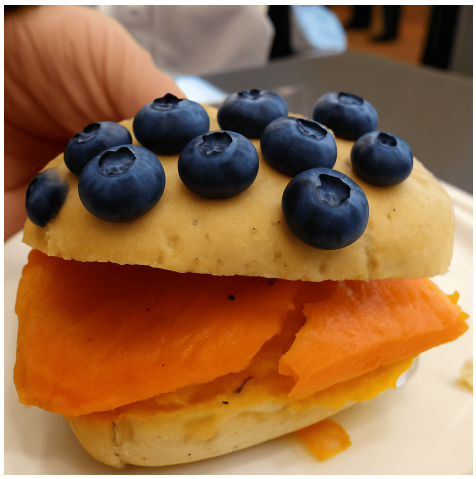
Figure 28: Examples for the task: Image Editing - Object Removing

Image Editing - Object Modifying

Instruction: Generate an image displaying the result after replacing the nuts with blueberries, while keeping all other elements of the image unchanged.



Good Example: GEMINI 2.5 + GPT IMAGE



Evaluation: ✔, 0.853 (image similarity)

Bad Example: GPT-4O + GPT IMAGE



Evaluation: ✘

Analysis: Change elements (salmon to oranges) that should remain the unchanged.

Figure 29: Examples for the task: Image Editing - Object Modifying

Self Count

Instruction: Create an image of a kitchen counter displaying multiple apples—specifically, more than two but fewer than seven—that are clearly visible, not stacked, and not overlapping. The apples should be of standard size and typical colors. After generating the image, return ONLY the number of apples N in the following XML format: <count>N</count>.

Good Example: GEMINI 2.5 + GPT IMAGE
Response:



<count>5</count>
Evaluation: ✔

Bad Example: GEMINI 2.5 + IMAGEN 3
Response:



<count>5</count>
Evaluation: ✘
Analysis: The number of apples is four.

Figure 30: Examples for the task: Self Count

Self Color Recognition

Instruction: Create an image of a kitchen that includes a chair, a microwave, and a vase, ensuring that each object appears exactly once. Assign a unique color to each object, choosing from the following: blue, yellow, and red. Once the image is complete, return only a JSON object showing the color assigned to each item, formatted as: {"chair": "[color]", "microwave": "[color]", "vase": "[color]"}.

Good Example: GPT-4O + GPT IMAGE
Response:



{"chair": "blue", "microwave": "yellow", "vase": "red"}
Evaluation: ✔

Bad Example: GEMINI IMAGE
Response:



Evaluation: ✘
Analysis: No text response. Forgetting to output text is a common mistake for interleaved image-text generation.

Figure 31: Examples for the task: Self Color Recognition

Self Size Recognition

Instruction: Create an image of a living room that includes a carpet, a TV, and a poster, each with a different area and appearing only once in the scene. After generating the image, compare the areas of these three objects and provide ONLY a ranked list from smallest to largest area, such as [“poster”, “TV”, “carpet”].

Good Example: GPT-4O + GPT IMAGE
Response:



[“poster”, “TV”, “carpet”]
Evaluation: ✔

Bad Example: GEMINI 2.5 + IMAGEN 3
Response:



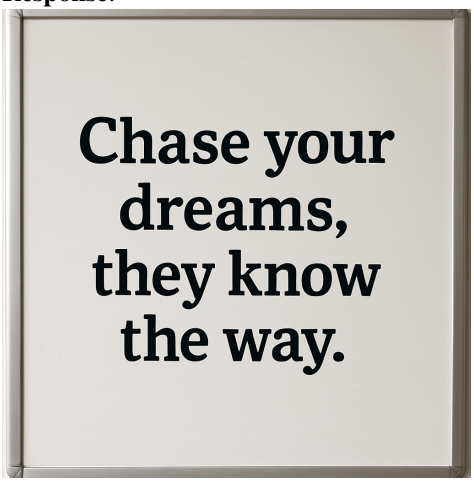
[“poster”, “TV”, “carpet”]
Evaluation: ✘
Analysis: The poster is larger than then the TV

Figure 32: Examples for the task: Self Size Recognition

Self OCR

Instruction: Create an image of a whiteboard displaying a motivational quote. The quote must be between 5 and 10 words long, written in a clear, standard, and easily readable typeface without any artistic alterations. The quote should encourage people to follow their dreams. Once the image is generated, return ONLY the text from the whiteboard enclosed in XML format: <text>board text</text>.

Good Example: GEMINI 2.5+GPT IMAGE
Response:



<text>chase your dreams, they know the way.</text>
Evaluation: ✔

Bad Example: GEMINI IMAGE
Response:



<text>follow yousved theie to say, dreams their their cleay drevcrd experition.</text>
Evaluation: ✘
Analysis: The generated image has distorted text and does not align with model's response.

Figure 33: Examples for the task: Self OCR

Self Relative Spatial Relation Recognition

Instruction: Create an image showing a bathroom counter that includes a toothbrush, a bar of soap, and a tube of toothpaste. Each item should appear exactly once in the image. Once the image is complete, respond to these two questions: 1. The soap is to the ___ of the toothpaste tube: A. left B. right C. neither obviously left nor right. 2. The toothbrush is positioned ___ than the soap: A. higher B. lower C. neither obviously higher nor lower. Provide ONLY a list of letter answers (e.g., ["A", "B"]).

Good Example: GEMINI 2.5 + GPT IMAGE
Response:



["A", "A"]
Evaluation: ✔

Bad Example: GEMINI IMAGE
Response:



["A", "C"]
Evaluation: ✘
Analysis: The toothbrush is positioned lower than the soap.

Figure 34: Examples for the task: Self Relative Spatial Relation Recognition

Self Absolute Spatial Relation Recognition

Instruction: Create an image of a kitchen countertop featuring a blender, a bowl, and a cutting board, each positioned in a different quadrant of the image. Ensure that each object appears only once. Once the image is complete, answer the following two questions: 1. In which quadrant is the blender located? A. bottom left B. bottom right C. top left D. top right. 2. In which quadrant is the cutting board located? A. bottom left B. bottom right C. top left D. top right. Respond with ONLY a list of the corresponding letter choices (e.g., [“A”, “D”])..

Good Example: GEMINI 2.5 + IMAGEN 3
Response:



[“C”, “B”]
Evaluation: ✔

Bad Example: GEMINI 2.5 + GPT IMAGE
Response:



[“C”, “A”]
Evaluation: ✘
Analysis: The cutboard is at the bottom center of the image.

Bad Example: GPT-4O + GPT IMAGE
Response:

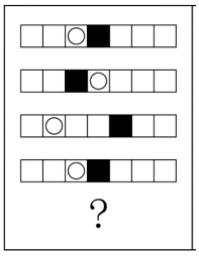


[“C”, “B”]
Evaluation: ✘
Analysis: There are multiple bowls. The wrong number of objects is a common mistake for all tasks.

Figure 35: Examples for the task: Self Absolute Spatial Relation Recognition

Interleaved Math Reasoning

Instruction: Carefully examine the sequence and determine the most suitable pattern to replace the question mark. Analyze the sequence step by step and clearly explain your reasoning. Present your final answer enclosed in double angle brackets as follows: «a detailed text description of the pattern ONLY». Then, create an image representing the same pattern to fill in the question mark.



Unfortunately all the models we evaluate perform poorly on this task. Here we will demonstrate our gold reference images.

Reference Answer:



Evaluation: ✔

Bad Example: GEMINI 2.5 + GPT IMAGE

Response:

To determine the pattern that replaces the question mark, let's analyze the positions of the white circle and the black square in each row of the sequence. We'll number the squares from 1 (leftmost) to 6 (rightmost).

Step 1: Analyze the position of the white circle.

Row 1: The white circle is in the 3rd square.

Row 2: The white circle is in the 4th square. (Movement: 3 → 4, which is +1)

Row 3: The white circle is in the 2nd square. (Movement: 4 → 2, which is -2)

...

A white circle in the 1st square.

A black square in the 6th square.

The squares in between (2nd, 3rd, 4th, and 5th) will be empty.

«A row of six squares with a white circle in the first square from the left and a black square in the sixth square from the left. The second, third, fourth, and fifth squares are empty. »



Evaluation: ✘

Analysis: All the moving patterns are correct, but the last step of calculating the position of the black square goes wrong.

Figure 36: Examples for the task: Interleaved Math Reasoning

Interleaved Code Reasoning

Instruction: [SVG Code] What does this SVG code represent? Analyze the elements step by step, then create a rendered image showing how it would appear in a browser. (reference image is not given)



reference image

Good Example: GEMINI 2.5+GPT IMAGE
Response:
This SVG code describes a cartoon image, likely an emoji or a simple illustration. Let's break down its components:
1. `<path fill="#AF7E57" d="M6 18.562..."/>`: This is the main shape of the face, filled with a tan color (`#AF7E57`). It's a large, somewhat pear-shaped element forming the head.
2. `<path fill="#DD2E44" d="M30.961 10c..."/>`: This path, filled with red (`#DD2E44`), forms the main part of a hat sitting on top of the head.
...
11. `<path fill="#DD2E44" d="M18 30.651c..."/>`: A red shape at the bottom of the face, forming a smiling mouth.
12. `<path fill="#D1D3D4" d="M10 25c0.../>` and `<path fill="#D1D3D4" d="M26.003 25c0..."/>`: Two light gray shapes at the very bottom, possibly representing parts of a beard, chin, or stylized feet if it were a full figure. Given the context, they are likely part of the lower face or beard detail.
Collectively, these elements form an image of a **Santa Claus face emoji**. Here is a rendered image of how it would appear:



Evaluation: ✔, 0.781 (similarity score)

Bad Example: GPT-4O + GPT IMAGE
Response:
The SVG code represents a stylized cartoon face, possibly a bear or a similar animal, with various decorative elements. Let's break down the elements step by step:
1. **Main Face Shape**: The main face is created using a path with a fill color of `#AF7E57`, which is a brownish color. This path forms the central part of the face.
2. **Ears**: Two ellipses with a fill color of `#D1D3D4` are used to create the ears. These are positioned at the top left and right of the face.
...
8. **Mouth**: A path with a fill color of `#915A34` is used to create the mouth, adding detail to the expression.
9. **Background Elements**: The SVG includes additional paths and ellipses to add depth and detail to the overall design.
Now, let's create a rendered image showing how this SVG would appear in a browser:



Evaluation: ✘
Analysis: Wrong final image.

Figure 37: Examples for the task: Interleaved Code Reasoning

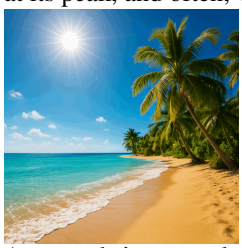
Image-Text Order

Instruction: Use visual storytelling to depict the four seasons. For each season, begin by creating an image of a landscape that represents that time of year, followed by a written description in natural language highlighting its key features. Provide only the requested content—do not include explanations, titles, or extra commentary.

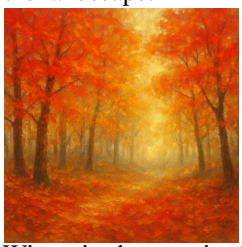
Good Example: GEMINI 2.5+GPT IMAGE
Response: This spring scene showcases fresh green grass, trees bursting with new leaves and pink blossoms, and a variety of colorful flowers under a bright, clear sky.



Summer is represented by a bright, sunny day with deep blue skies, lush green foliage at its peak, and often, warm, inviting waters.



Autumn brings a palette of warm colors as leaves turn red, orange, and yellow; the air becomes crisp, and a sense of calm pervades the landscape.



Winter is characterized by a blanket of snow covering the ground and trees, often with frozen bodies of water and a quiet, still atmosphere.



Evaluation: ✔

Bad Example: GEMINI IMAGE
Response: A vibrant green field stretches under a clear blue sky, bordered by a line of lush, dark green trees in the distance. The grass shows subtle stripes from recent mowing, indicating care and growth.
A stark white landscape is dominated by a snow-covered path leading into the distance, flanked by wooden posts connected by dark railings. The sky is a pale, overcast white, blending seamlessly with the snow-covered ground.
Patches of snow dust a field of dry, golden-brown grass, contrasting with the remaining green leaves on the distant trees under a bright blue sky with wispy clouds. A large tree branch with vibrant orange leaves extends into the frame from the upper left.
A winding path covered in fresh snow cuts through a forest where trees are heavily laden with snow, creating a monochrome scene in shades of white and dark gray. A wooden fence runs along the side of the path, also covered in snow.



Evaluation: ✘
Analysis: Generate only one image.

Figure 38: Examples for the task: Image-Text Order

Sound Begin-End

Instruction: Create an audio that begins with the sound of chirping birds and concludes with the sound of a mouse click.

Good Example: MAKE-AN-AUDIO 2
Audio Description: an audio with the beginning 3 seconds of chirping birds and the last 1 sec of mouse click.
Evaluation: ✔

Bad Example: STABLE AUDIO
Audio Description: an audio with chirping birds all the time, and a mouth click only in the 2nd second, not the end.
Evaluation: ✘

Figure 39: Examples for the task: Sound Begin-End

Sound Positional Inclusion

Instruction: Create an audio of a city street, ensuring a police car siren is included in the first half.

Good Example: AUDIOGEN
Audio Description: an audio of a city street with the first 3 seconds including a police car siren. **Evaluation:** ✔

Bad Example: STABLE AUDIO
Audio Description: an audio of police car siren mixed with normal cars passing by street all the time.
Evaluation: ✘

Figure 40: Examples for the task: Sound Positional Inclusion

Sound Reasoning

Instruction: Produce the sound of a black bird commonly linked to death or misfortune.

Good Example: MAKE-AN-AUDIO 2
Audio Description: an audio of very typical crow cry.
Evaluation: ✔

Bad Example: STABLE AUDIO
Audio Description: bright and melodious chirp of a certain kind of bird that's obviously not crow.
Evaluation: ✘

Figure 41: Examples for the task: Sound Reasoning

Sound Silence

Instruction: Create an audio that begins with a loud car horn, followed by a long silence, and concludes with a distant siren.

Good Example: MakeAnAudio2
Audio Description: an audio with a loud car horn in the beginning 3 seconds, and then comes 4 seconds of silence, with the last 3 seconds of a distant siren.
Evaluation: ✔

Bad Example: AUDIOLDM 2
Audio Description: an audio without any silent time, and the two sounds are mixed together.
Evaluation: ✘

Figure 42: Examples for the task: Sound Silence

Music Instrument Inclusion

Instruction: Create a seamless saxophone improvisation.

Good Example: TANGO MUSIC
Audio Description: a casual piece of saxophone improvisation.
Evaluation: ✔

Bad Example: YUE
Audio Description: an audio starting with 3 seconds of laughter and then 5 seconds of improvised jazz music including piano, drums and saxophone.
Evaluation: ✘

Figure 43: Examples for the task: Music Instrument Inclusion

Music Instrument Exclusion

Instruction: Create an audio of a city street, ensuring a police car siren is included in the first half.

Good Example: MUSICGEN
Audio Description: an audio of smooth jazz music featuring bass but without drums.
Evaluation: ✔

Bad Example: STABLE AUDIO
Audio Description: an audio of jazz music with rhythms played by drums.
Evaluation: ✘

Figure 44: Examples for the task: Music Instrument Exclusion

Music Intensity

Instruction: Compose a cinematic orchestral piece that gradually fades out at the end.

Good Example: STABLE AUDIO
Audio Description: an audio of orchestral piece featuring a cinematic build with rich instrumentation and gradually fading out, creating a smooth ending.
Evaluation: ✔

Bad Example: TANGO MUSIC
Audio Description: an audio quite the opposite, with a tranquil start and getting more intense.
Evaluation: ✘

Figure 45: Examples for the task: Music Intensity

Music Tempo

Instruction: Create a laid-back lo-fi hip-hop beat at 100 BPM.

Good Example: MusicGen
Audio Description: an audio of a hip-hop beat at approximately 102 BPM.
Evaluation: ✔

Bad Example: AUDIOLDM2
Audio Description: an audio of a hip-hop beat at approximately 64 BPM.
Evaluation: ✘

Figure 46: Examples for the task: Music Tempo

Speaker Voice Attribution

Instruction: Generate an audio of a man speaking rapidly in a low-pitched voice, saying, "The detective carefully examined the crime scene, noting every detail that could lead him to the truth, knowing that even the smallest clue might be the key to solving the mystery."

Good Example: GEMINI 2.5 + VOXINSTRUCT
Speech Transcript: (low-pitched male voice talking rapidly) The detective carefully examined the crime scene, noting every detail that could lead him to the truth, knowing that even the smallest clue might be the key to solving the mystery.
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOICELDM
Speech Transcript: (high-pitched male voice talking rapidly) The detective carefully examined the crime scene, noting every detail that could lead him to the truth, knowing that even the smallest clue might be the key to solving the mystery.
Evaluation: ✘

Figure 47: Examples for the task: Speaker Voice Attribution

Multi-Lingual Speech

Instruction: Generate an audio of a man slowly speaking: "窗外的雨滴敲打着玻璃，滴滴答答的声音仿佛一首温柔的旋律，让她的思绪飘回了那个久远而温暖的夏天。"

Good Example: GEMINI 2.5 + VOXINSTRUCT
Speech Transcript: (an audio of a man gently and slowly speaking Chinese) 窗外的雨滴敲打着玻璃，滴滴答答的声音仿佛一首温柔的旋律，让她的思绪飘回了那个久远而温暖的夏天。
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOXINSTRUCT
Speech Transcript: (an audio of a man speaking Chinese hastily) 窗外的雨滴敲打着玻璃，仿佛一首温柔的旋律，让她的思绪飘回了那个久远而温暖的夏天。
Evaluation: ✘

Figure 48: Examples for the task: Multi-Lingual Speech

Speaker Voice Replication

Instruction: Create an audio of reading the sentence, "The aroma of fresh coffee and warm pastries filled the air as she stepped into the café, instantly feeling a sense of comfort and familiarity in the cozy atmosphere," using the same voice as the reference speaker. (reference speech: a gentle male sound)

Good Example: GEMINI 2.5 + VOICELDM
Speech Transcript: (a warm, middle-pitched gentle male sound) The aroma of fresh coffee and warm pastries filled the air as she stepped into the café, instantly feeling a sense of comfort and familiarity in the cozy atmosphere.
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOXINSTRUCT
Speech Transcript: (a hoarse male sound) The aroma of fresh coffee and warm pastries filled the air as she stepped into the café.
Evaluation: ✘

Figure 49: Examples for the task: Speaker Voice Replication

Transcript Generation

Instruction: Create a speech on the internet revolution, making sure the word "digital" is used at least twice.

Good Example: GEMINI 2.5 + VOXINSTRUCT
Speech Transcript: As you just heard, the internet revolution has profoundly impacted our lives, marking a significant shift into a new digital era. This ongoing digital transformation continues to unfold, bringing new opportunities and challenges.
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOICELDM
Speech Transcript: Friends, colleagues, we stand at the precipice of an era by connectivity. The internet revolution has reshaped our world, ushering in a new digital age with transformation connecting billions and democratizing information, and its story is still being written.
Evaluation: ✘
Analysis: keyword "digital" appears only once.

Figure 50: Examples for the task: Transcript Generation

Transcript Editing

Instruction: Revise the provided speech by replacing the third sentence with: "Join hands, break barriers, and transform the world." (Provided speech transcript: "We must act now. Every action counts. We can make a difference together if we remain united. Progress comes from shared responsibility.")

Good Example: GEMINI 2.5 + VOICELDM
Speech Transcript: We must act now. Every action counts. We can make a difference together if we remain united. Join hands, break barriers, and transform the world.
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOXINSTRUCT
Speech Transcript: We must act now. Every action counts. Join hands, break barriers, and transform the world.
Evaluation: ✘
Analysis: The third sentence is missing.

Figure 51: Examples for the task: Transcript Editing

Speech Conversation

Instruction: Create five consecutive speeches forming a dialogue between James and Victoria about architectural innovations for sustainable urban living, with each speech delivered by a single speaker. Use the following speaking order: James, Victoria, Victoria, James, James.
Constraints: 1. Victoria's second speech must begin with the phrase "Contrary to your point."
2. James's final speech must contain exactly two uses of the word "design".
Ensure each speaker maintains a consistent voice throughout the conversation.

Good Example: GEMINI 2.5 + VOICELDM
Conversation Transcript:
James: I've been fascinated by the potential of kinetic pavements in high-traffic urban areas to generate clean energy.
Victoria: That's an interesting concept, James! I also think smart water management systems like rainwater harvesting and greywater recycling are vital.
Victoria: Contrary to your point about focusing on generation, I believe water conservation innovations offer more immediate and widespread benefits for sustainability.
James: Both are crucial, Victoria. Perhaps integrating biomimetic architecture that learns from nature's efficiency could address energy and water simultaneously.
James: The overall urban design must be rethought. We need a cohesive design approach that incorporates all these innovations seamlessly for truly sustainable cities.
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOXINSTRUCT
Conversation Transcript:
James: I believe modular construction is key to sustainable urban development, significantly reducing waste and build times.
Victoria: That's a valid point, James. Integrating green roofs and vertical gardens can also drastically improve air quality and biodiversity.
Victoria: Contrary to your point about modular construction being the key, I think retrofitting existing buildings with smart technologies offers a more immediate impact on energy consumption.
James: Retrofitting is important, Victoria, but new constructions set the standard. We need innovations in material science too, like self-healing concrete.
James: Ultimately, a holistic approach to urban design is crucial.
Evaluation: ✘
Analysis: James's final speech only contain one use of the word "design"

Figure 52: Examples for the task: Speech Conversation

Audio-Text Order

Instruction: Develop a 2-step guide for delivering a persuasive presentation. For each step, begin with detailed written instructions in natural language, followed by a speech from a communication expert demonstrating the technique. For the second step, also include an additional speech at the end that illustrates how the technique is applied in a formal business context. Generate only what is requested—no explanations, titles, or extra commentary.

Good Example: GEMINI 2.5 + VOXINSTRUCT
Response: Step 1: Thoroughly research your audience to understand their needs, motivations, existing knowledge, and potential objections. Adapt your language, examples, and level of detail to resonate specifically with them. Frame your core message in a way that directly addresses their interests and demonstrates clear value from their perspective, ensuring your arguments are supported by credible evidence and logical reasoning.
(speech) *To truly persuade, you must first connect—understand who you're speaking to, what their concerns are, their aspirations. Tailor your message to their world, not yours. Make it relevant, make it resonate.*
Step 2: Structure your presentation for maximum impact by creating a clear, logical, and compelling narrative. Begin with a strong opening that grabs attention and establishes relevance. Develop your core arguments with supporting evidence, data, and illustrative examples, addressing potential counterarguments proactively. Conclude with a memorable summary and a clear call to action, guiding your audience on the next steps.
(speech)*Your delivery is as crucial as your content. Project confidence through your voice and body language. And when faced with questions or doubts, listen carefully, acknowledge their points, and respond thoughtfully. This builds trust.*
(speech) *While I understand concerns about the initial investment, consider the long-term ROI. Our projections show a 20% return within two years, significantly boosting our profit margins and market competitiveness. This strategic move positions us for success.*
Evaluation: ✔

Bad Example: GEMINI 2.5 + VOICELDM
Response:
Step 1: Thoroughly understand your audience and craft a compelling narrative. This involves researching their demographics, interests, pre-existing beliefs, and potential pain points. Use this understanding to build a story that resonates, typically following a structure that introduces a problem, presents your proposed solution, highlights the benefits, and concludes with a clear call to action. Employ vivid language, relatable examples, and appropriate emotional appeals to make your message memorable and impactful.
Step 2: Master your delivery and prepare to handle objections gracefully. Effective delivery encompasses strong body language—maintaining eye contact, adopting an open posture, and using purposeful gestures—as well as vocal variety, modulating your tone, pace, and volume to maintain engagement.
(speech) *To truly persuade, you must first connect—understand who you're speaking to, what their concerns are, their aspirations. Tailor your message to their world, not yours. Make it relevant, make it resonate.*
(speech) *A persuasive presentation flows like a good story. Hook your audience early, build your case with compelling evidence, and then guide them to action.*
(speech) *Good morning. Our analysis indicates a 15% market share increase is achievable by Q4 if we implement the proposed strategy. This directly addresses the growth targets set last quarter and positions us ahead of key competitors. We seek your approval to proceed.*

Evaluation: ✘
Analysis: Speeches and texts are not interleaved in the expected order.

Figure 53: Examples for the task: Audio-Text Order