

RISEBench: Benchmarking Reasoning-Informed viSual Editing

**Xiangyu Zhao*, Peiyuan Zhang*, Kexian Tang*, Xiaorong Zhu*,
Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia,
Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang**†, Haodong Duan†****



上海人工智能实验室
Shanghai Artificial Intelligence Laboratory



**PRINCETON
UNIVERSITY**

Rethinking Image Editing Task

Change Motion



Change the dog's action to running through the snow.

Change Material



Change the sofa material to cotton and linen.



Draw the left view.

Limitations of Traditional Evaluation

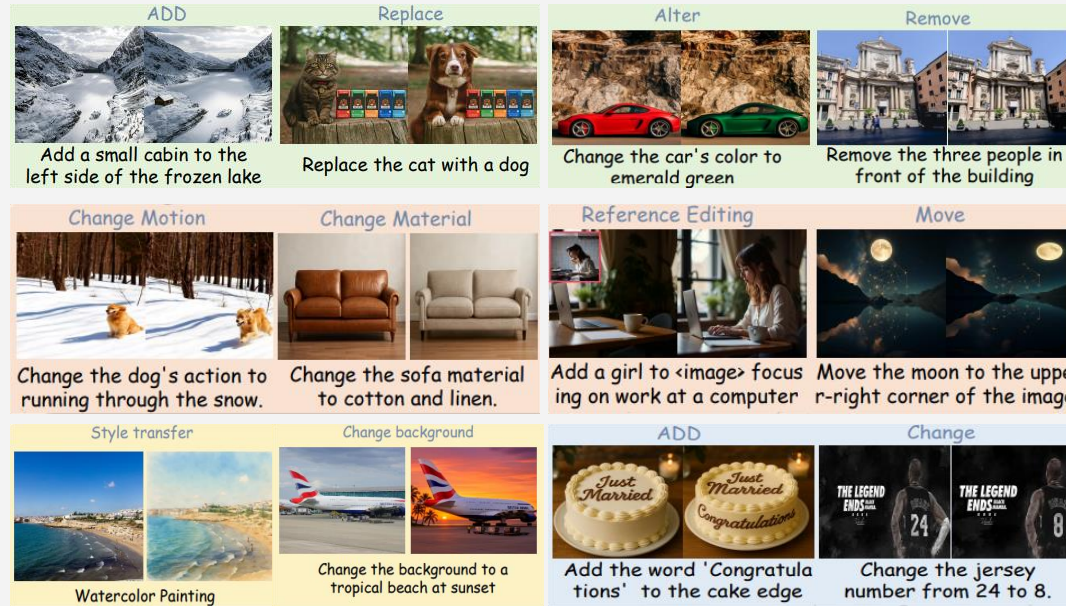
"Replace the tie with a superhero cape"



"Replace the word with 'pure'"



commands are short
and overly explicit



tasks are low-complexity and lack diversity



Make the bears wearing tiny hats
of different colors



Change the action of the horses to
galloping

Simplistic instructions

Limited task types

Quality inconsistency

A standardized benchmark for evaluating Reasoning-Informed viSual Editing remains absent.

Limitations of Traditional Evaluation

❑ Simplistic instructions

- Current datasets rely on **shallow, template-based** editing commands.
- Such instructions offer **limited semantic depth** and poorly reflect real user expression.
- More **complex, reasoning-informed** editing instructions are therefore needed.



Limitations of Traditional Evaluation

□ Limited task types

- Most existing datasets focus on **basic tasks** like add, replace, alter, remove, and style transfer.....
- Such operations remain **low in semantic complexity** and offer **limited demands on reasoning**.
- This simplicity prevents robust assessment of a model's deeper editing abilities.



Replace the bubble with **Fire**



Make this image a **watercolor painting**



Turn the horse into a **colourful unicorn**



Let there be a **shark** in the water



Turn the environment into a **snowy landscape**



Add a **golden retriever** to the foreground



Remove his **hat**

Limitations of Traditional Evaluation

□ Quality inconsistency

- A portion of existing datasets suffers from **low-quality** annotations generated through automated construction pipelines.
- These pipelines often introduce **inaccuracies or poorly executed edits**.
- Such noise ultimately **undermines the reliability** of the data for training and evaluation.



Make the bears wearing tiny hats of different colors



Remove his beard



Replace the school bus with a truck



Change the action of the horses to galloping



Adjust the background to a garden



Replace the text 'TRAIN' with 'PLANE'



RISEBench: Towards General Editing Intelligence

What kind of questions do we have?

Temporal Reasoning



Draw what they will look like after being kept in a daily environment for a year.



Draw what it will look like 5 seconds later.



This image is observed from the Northern Hemisphere. Draw what it will look like 7 days later.



Draw what it looked like three hours later.



Draw what it looked like ten minutes ago at a temperature of 25 degrees Celsius.



Draw what it will look like after 30 seconds in summer.

Causal Reasoning



Draw what it will look like after being bitten by people.



Draw the consequence of capillary action.



Draw what the scene will look like when the speaker plays strong, powerful music.



Draw what it will look after it has been shaken vigorously and then opened.



Draw what it will look like after the knot is untied.



Draw what the iron nail looks like after being left in the beaker of copper sulfate.

Spatial Reasoning



Draw the top view.



Draw the objects arranging from left to right in order: red, orange, yellow, green, blue, and purple.



Draw the image after the black car has moved forward towards the camera.



Generate an image assembling the components into a complete clock displaying 9:45.



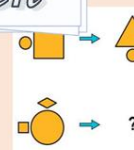
Draw the scene that the pink sofa and the oval coffee table is neatly arranged.



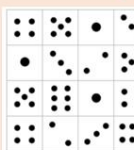
Draw the scene that shows the view from sitting on the red chair facing toward the bookshelf on the left side of the wall.

RISE Bench

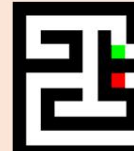
Logical Reasoning



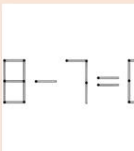
Complete the shape represented by the question mark.



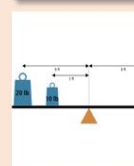
Cut out six adjacent squares that form a valid unfolded dice net.



This is a maze. Draw the path from the red dot to the green dot in blue.



Move two matchsticks on the left side of the equation to make the equation true.



Draw the final state of the lever.



Draw the solution, ensuring that each row and each column contains the numbers 1, 2, 3 and 4.

- Four Knowledgeable Categories
- 100% Human Expert Manually Annotated
- Challenging and Zero-shot Instruction
- Precise Textual or Image GT Answer
- Various Image Sources

How to design a good metric?



How to design a good metric?

Instruction Reasoning

1

Output Image

Reference Description



T-shirt with faint but clearly visible brown stains.

Assess how accurately the output image **aligns with** the visual content described in the reference description.....

Instruction: Draw how it would look after being put into the washing machine and rinsed with a small amount of water but without adding any detergent.

Prompt



Step-by-Step Evaluation

1. Instruction Analysis: -----
2. Reference Description: -----
3. Output Image Analysis: -----
4. Comparison to Reference Description:-----
5. Scoring Justification: -----

Judge & Final Score

2

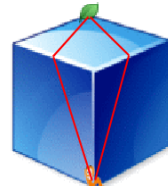
2

Output Image

Reference Image



&



Assess how accurately the output image **aligns with** the visual content

Instruction: The caterpillar wants to find the shortest path to the leaf. Please mark the two shortest paths in the diagram.

Prompt



Step-by-Step Evaluation

1. Problem Analysis: -----
2. Reference Image Analysis: -----
3. Output Image Comparison: -----
4. Scoring Justification: -----

Judge & Final Score

1

Metric 1

Instruction Reasoning

- For samples that are easy to describe (images with main object and simple constructure), we give **Reference Description** to describe the image in text.
- For samples that can not be described in several sentences(like spatial view transform or logical questions), we annotate the **GroundTruth image**.

How to design a good metric?

Metric 2

Appearance Consistency

The model should preserve the irrelevant visual attributes of the original image.

Appearance Consistency

Input & Output Evaluate how **consistent Image B remains with Image A** in all aspects except those explicitly changed by the instruction. You must **ignore the instructed changes** and only assess **unintended differences**.



Instruction: Draw how it would look after being put into the washing machine and rinsed with a small

Prompt: ?



Step-by-Step Evaluation

1. Instruction Analysis: -----
2. Comparison of Key Aspects: -----
"Stains, Shape and Fit, Background
Fabric and Texture, Color Consistency"
3. Evaluation of Consistency: -----

Judge & Final Score

4

Visual Plausibility

Output Image



Assess the **overall visual realism** and **generation fidelity** of the image. Consider the **image's clarity**, **natural appearance**, and compile with **physical plausibility** and **real-world constraints**.

Prompt: ?



Step-by-Step Evaluation

The image depicts a t-shirt with a colorful, splattered design. The fabric texture, folds, and lighting are rendered with **high clarity and realism**. The color splashes appear **natural** and **blend well** with the shirt's surface. There are **no noticeable distortions** or **unrealistic elements**.

Judge & Final Score

5

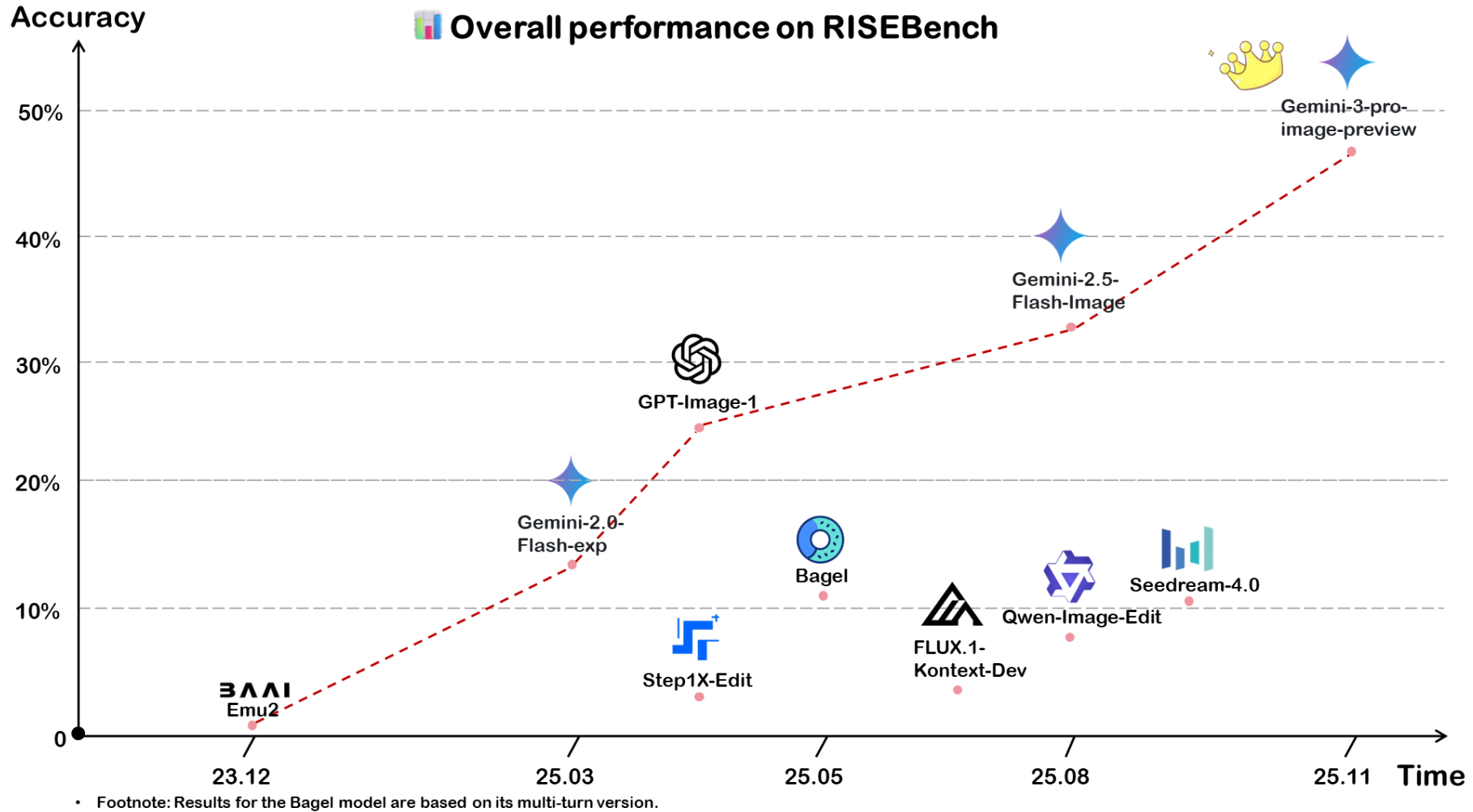
Metric 3

Visual Plausibility

The output should be coherent, realistic, and physically or logically plausible within context.

Results: How Far Are We Going to Editing Intelligence

Performance Trend



Accuracy measures the proportion of complete samples (achieving perfect score (5) in all three dimensions)

Comparison

Spatial Reasoning



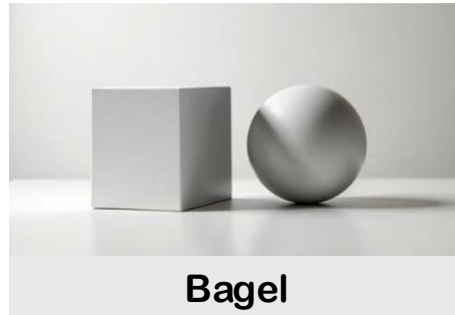
Draw the left view.

Reference Description

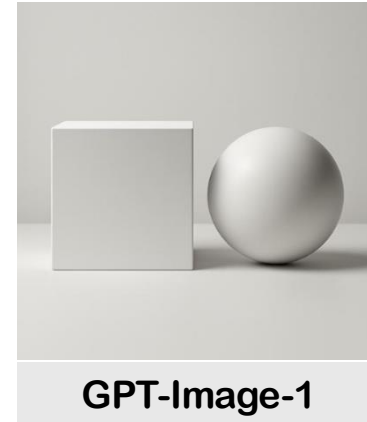
An image which only contains the front view of a white cube (a white square).



Banana pro



Bagel



GPT-Image-1



Seedream4.0

Comparison

Logical Reasoning



Draw a clear red line path from a red-and-white spotted mushroom house to a round mud pit.

Reference Image:



Banana pro



Bagel



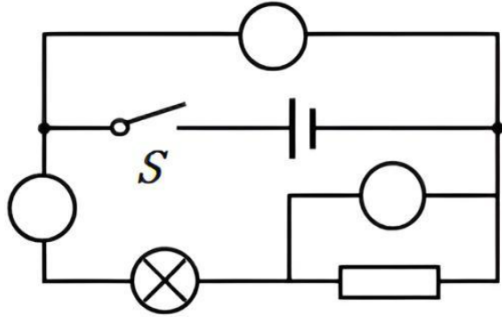
GPT-Image-1



Seedream4.0

Comparison

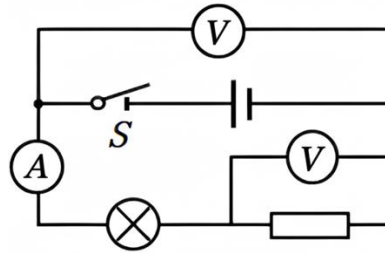
Logical Reasoning



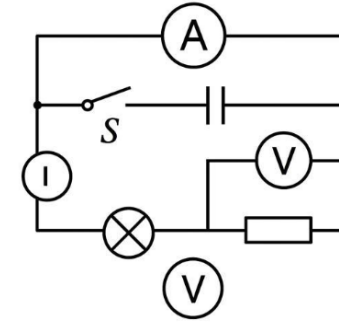
Complete a picture via filling in the Ammeter "A" and Voltmeter "V" at the circle in the circuit to make it the correct circuit.

Reference Description

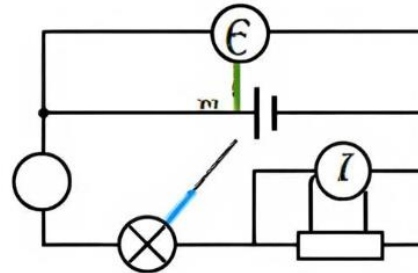
The circle at the very top is filled with "V", the circle at the far left is filled with "A", and the circle at the bottom right is filled with "V".



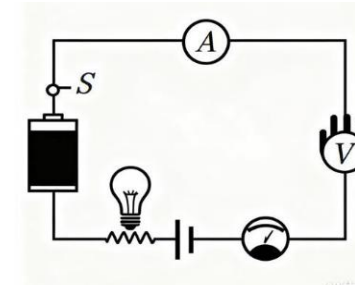
Banana pro



GPT-Image-1



Bagel



Seedream4.0

Comparison

Spatial Reasoning



Draw an image showing the view from someone sitting on the green chair, looking at the table screen.

Reference Description

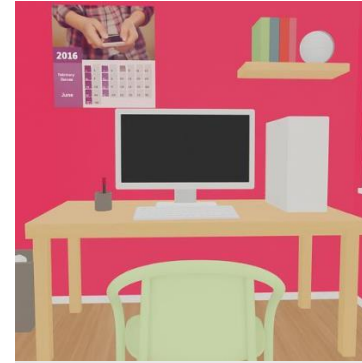
The picture shows a brown chair, a table and a computer screen, while the screen on the table is partially obscured by the back of a brown chair in the foreground.



Banana pro



Bagel



GPT-Image-1



Seedream4.0

Future Work: What could we do?

Future Work

Scaling

- The current **scale (360) is limited** for large-scale statistical analysis.

Evaluation Pipeline

- Current MLLM-based evaluators are not always accurate; they sometimes miss subtle artifacts or misinterpret complex logic.
- A more **robust and human-aligned evaluator** is strictly necessary.

Diversity

- Introduce more cognitively demanding tasks, specifically those requiring **multi-hop reasoning or creative imagination**.

Temporal Reasoning



Draw what they will look like after being kept in a daily environment for a year.



Draw what it will look like 5 seconds later.



This image is observed from the Northern Hemisphere. Draw what it will look like 7 days later.



Draw what it looked like three hours later.



Draw what it looked like ten minutes ago at a temperature of 25 degrees Celsius.



Draw what it will look like after 30 seconds in summer.

Causal Reasoning



Draw what it will look like after being bitten by people.



Draw the consequence of capillary action.



Draw what the scene will look like when the speaker plays strong, powerful music.



Draw what it will look after it has been shaken vigorously and then opened.



Draw what it will look like after the knot is untied.

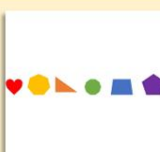


Draw what the iron nail looks like after being left in the beaker of copper sulfate.

Spatial Reasoning



Draw the top view.



Draw the objects arranging from left to right in order: red, orange, yellow, green, blue, and purple.



Draw the image after the black car has moved forward towards the camera.



Generate an image assembling the components into a complete clock displaying 9:45.

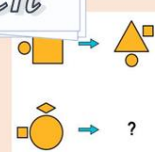


Draw the scene that the pink sofa and the oval coffee table is neatly arranged.

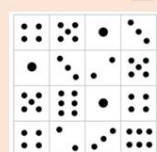


Draw the scene that shows the view from sitting on the red chair facing toward the bookshelf on the left side of the wall.

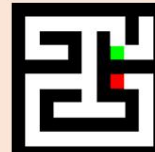
Logical Reasoning



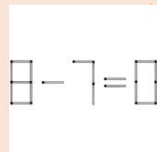
Complete the shape represented by the question mark.



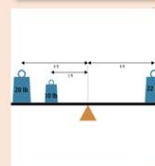
Cut out six adjacent squares that form a valid unfolded dice net.



This is a maze. Draw the path from the red dot to the green dot in blue.



Move two matchsticks on the left side of the equation to make the equation true.



Draw the final state of the lever.

		4	2
4	1		
	4	3	1
3		1	

Draw the solution, ensuring that each row and each column contains the numbers 1, 2, 3 and 4.

Thanks



z2855064151@sjtu.edu.cn

<https://github.com/PhoenixZ810/RISEBench>