NEURAL INFORMATION PROCESSING SYSTEMS 2025

The City of SAN DIEGO

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

SHANGHAI JIAO TONG UNIVERSITY 1896

TSINGHUA UNIVERSITY -1911-

PRINCETON UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS 2025

The City of SAN DIEGO

Rethinking Image Editing Task

Draw the left view.

Limitations of Traditional Evaluation

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.

NEURAL INFORMATION PROCESSING SYSTEMS 2025

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.

Zhao, Haozhe, et al. NIPS 2024

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.

Wei, Cong, et al. ICLR 2024.

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

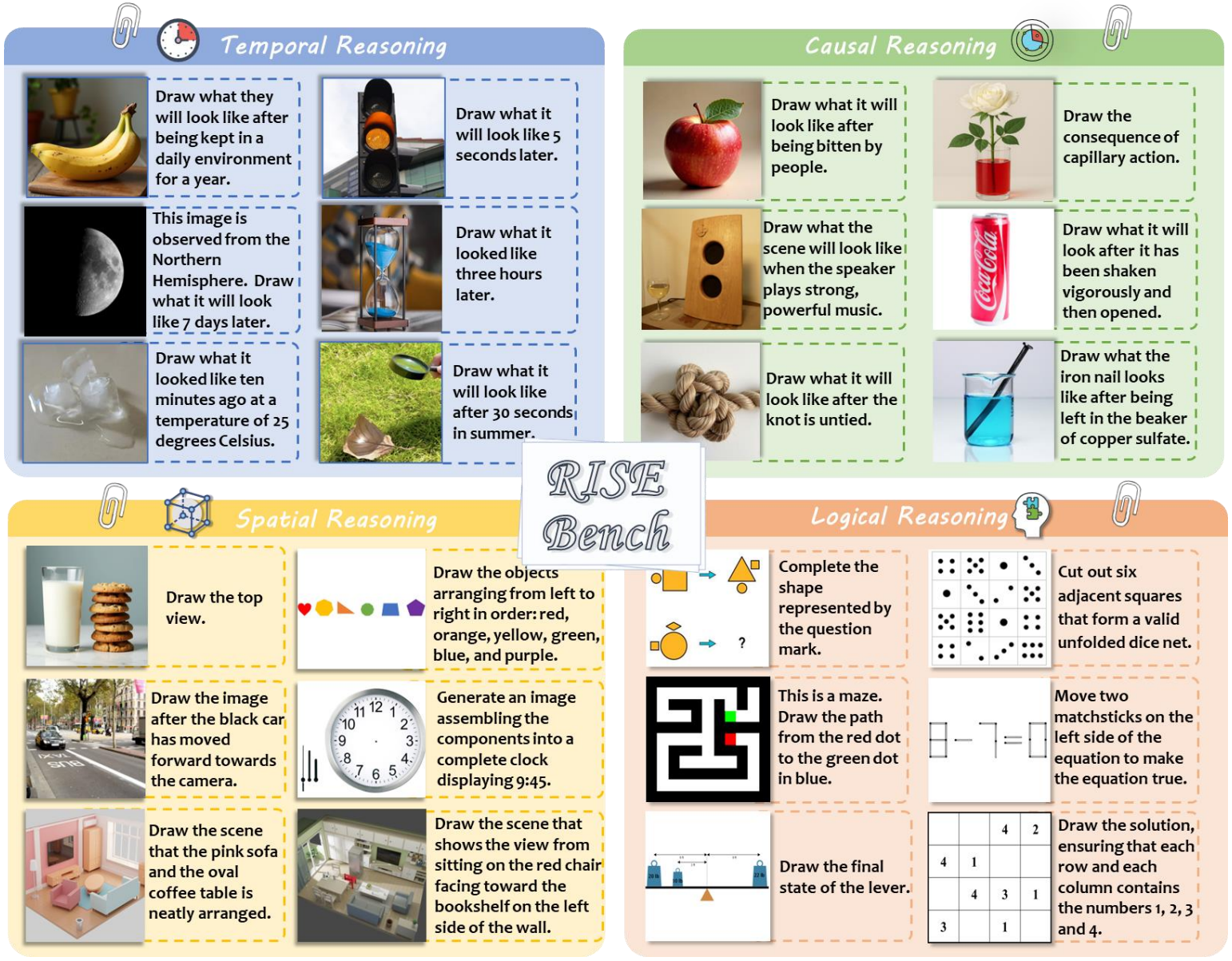
Replace the text 'TRAIN' with 'PLANE'

NEURAL INFORMATION PROCESSING SYSTEMS 2025

The City of SAN DIEGO

RISEBench: Towards General Editing Intelligence

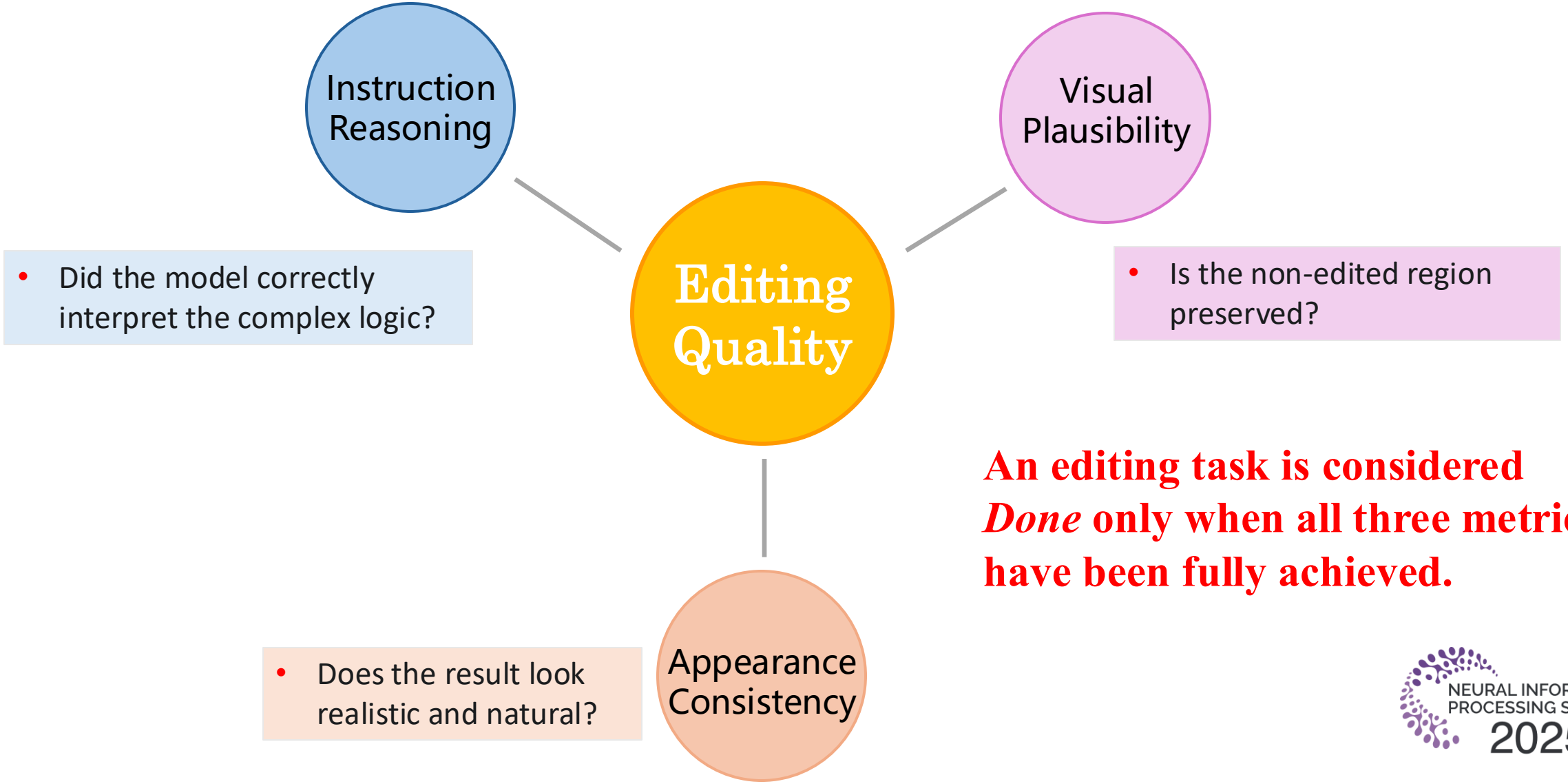
What kind of questions do we have?



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

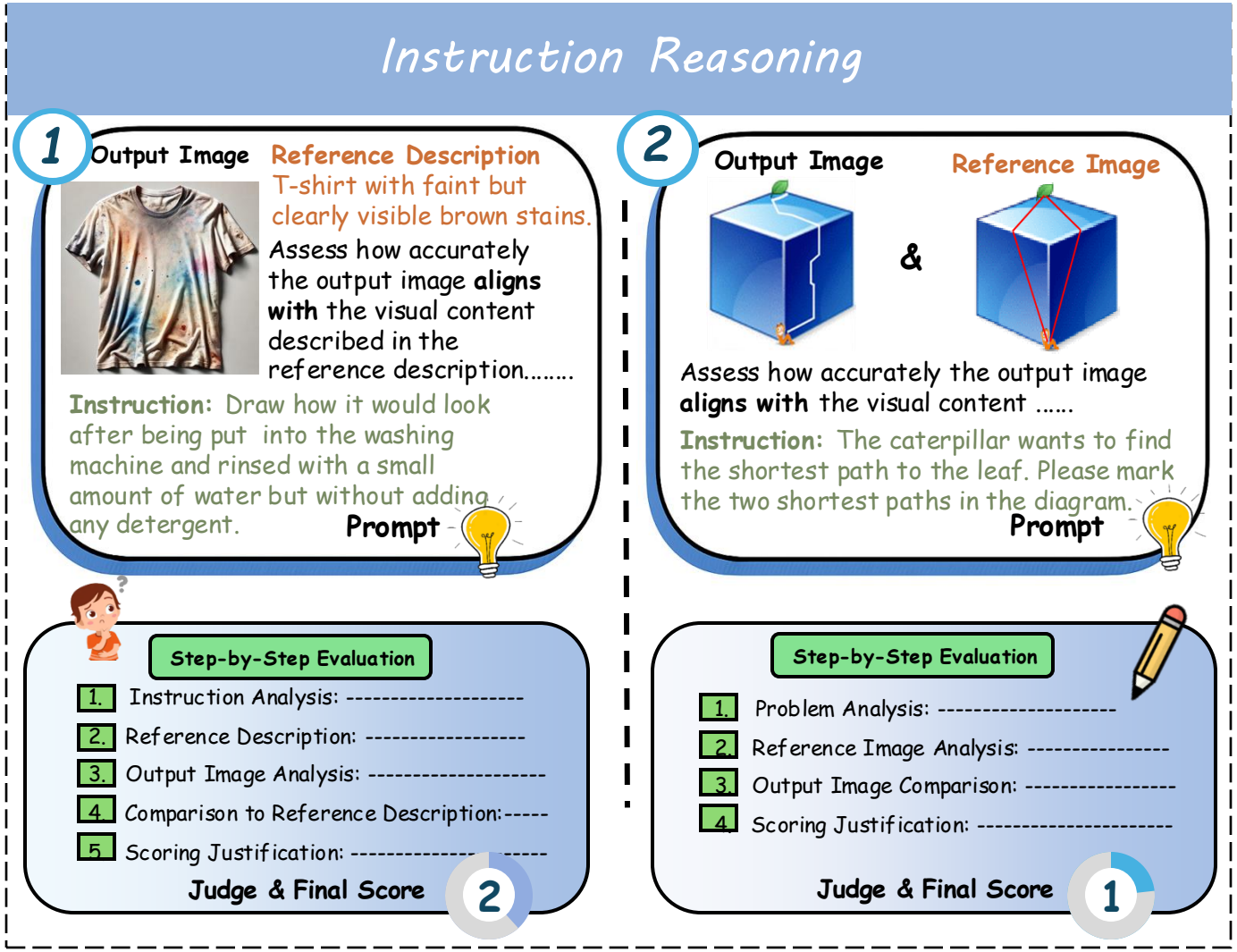
NEURAL INFORMATION PROCESSING SYSTEMS 2025

How to design a good metric?



An editing task is considered *Done* only when all three metrics have been fully achieved.

How to design a good metric?



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.

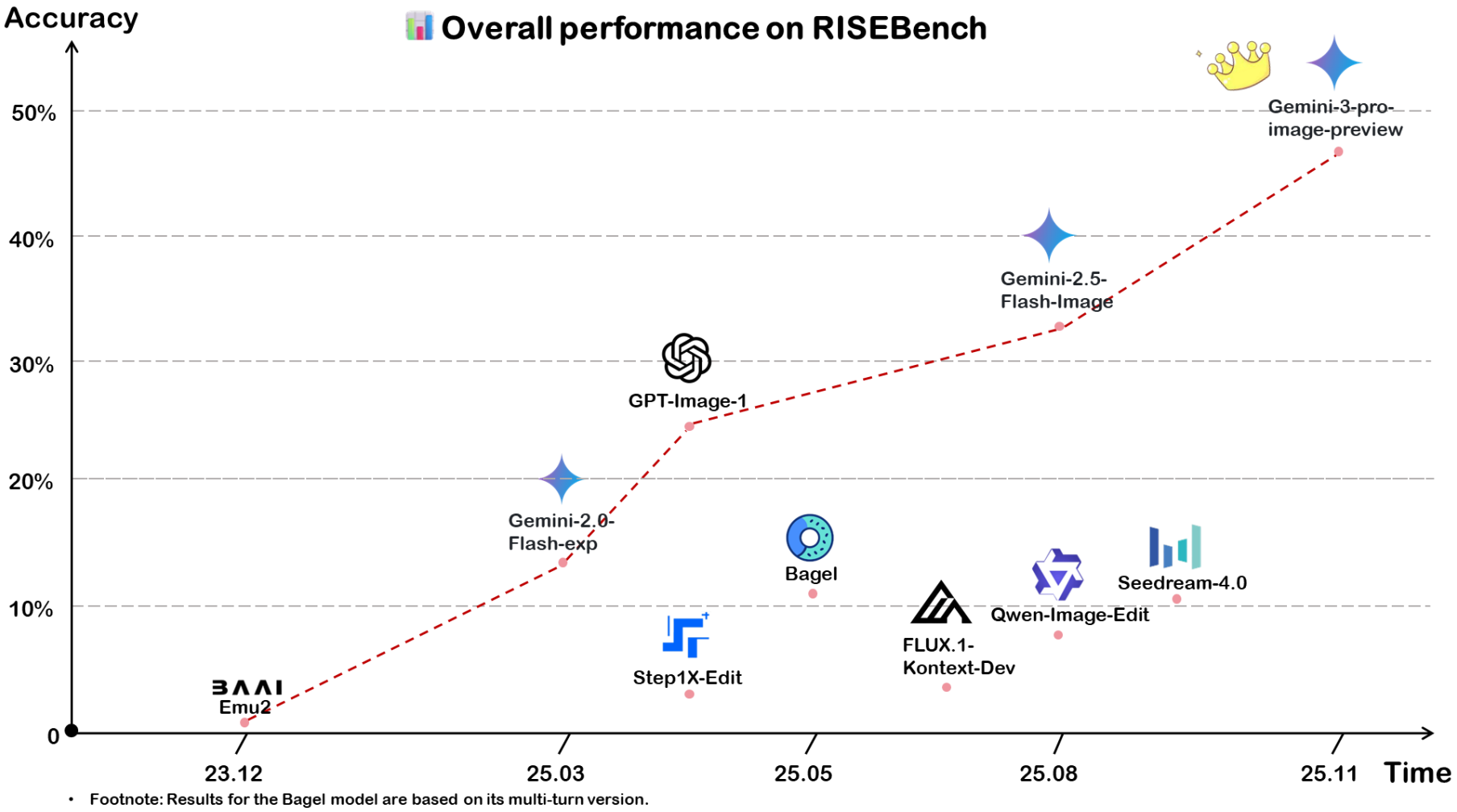
NEURAL INFORMATION PROCESSING SYSTEMS 2025

NEURAL INFORMATION PROCESSING SYSTEMS 2025

The City of SAN DIEGO

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

GPT-Image-1

Bagel

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

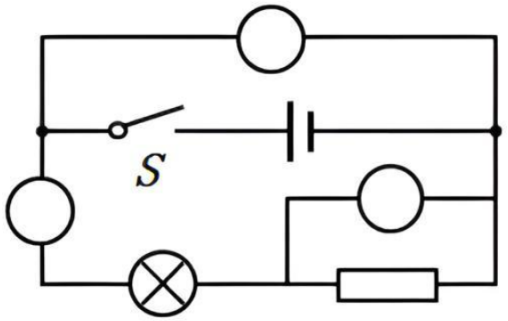
GPT-Image-1

Bagel

Seedream4.0

Comparison

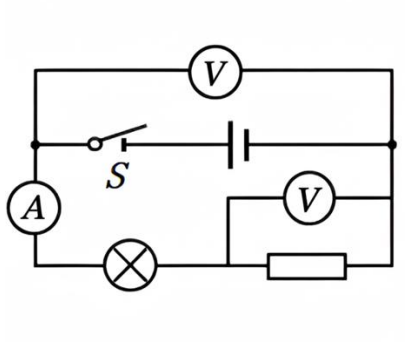
Logical Reasoning



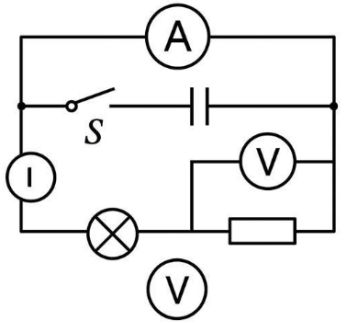
Complete a picture via filling in the Ammeter "A" and Voltmeter "V" at the circle in the circuit to make it the correct circuit.

Reference Description

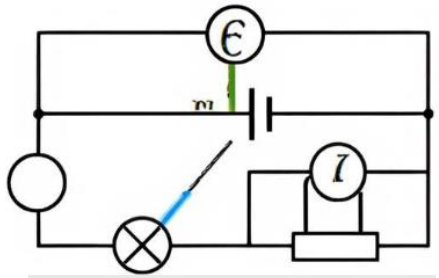
The circle at the very top is filled with "V", the circle at the far left is filled with "A", and the circle at the bottom right is filled with "V".



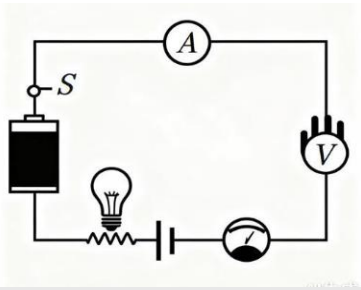
Banana pro



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

GPT-Image-1

Bagel

Seedream4.0

Future Work: What could we do?

Future Work

Scaling

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

Evaluation Pipeline

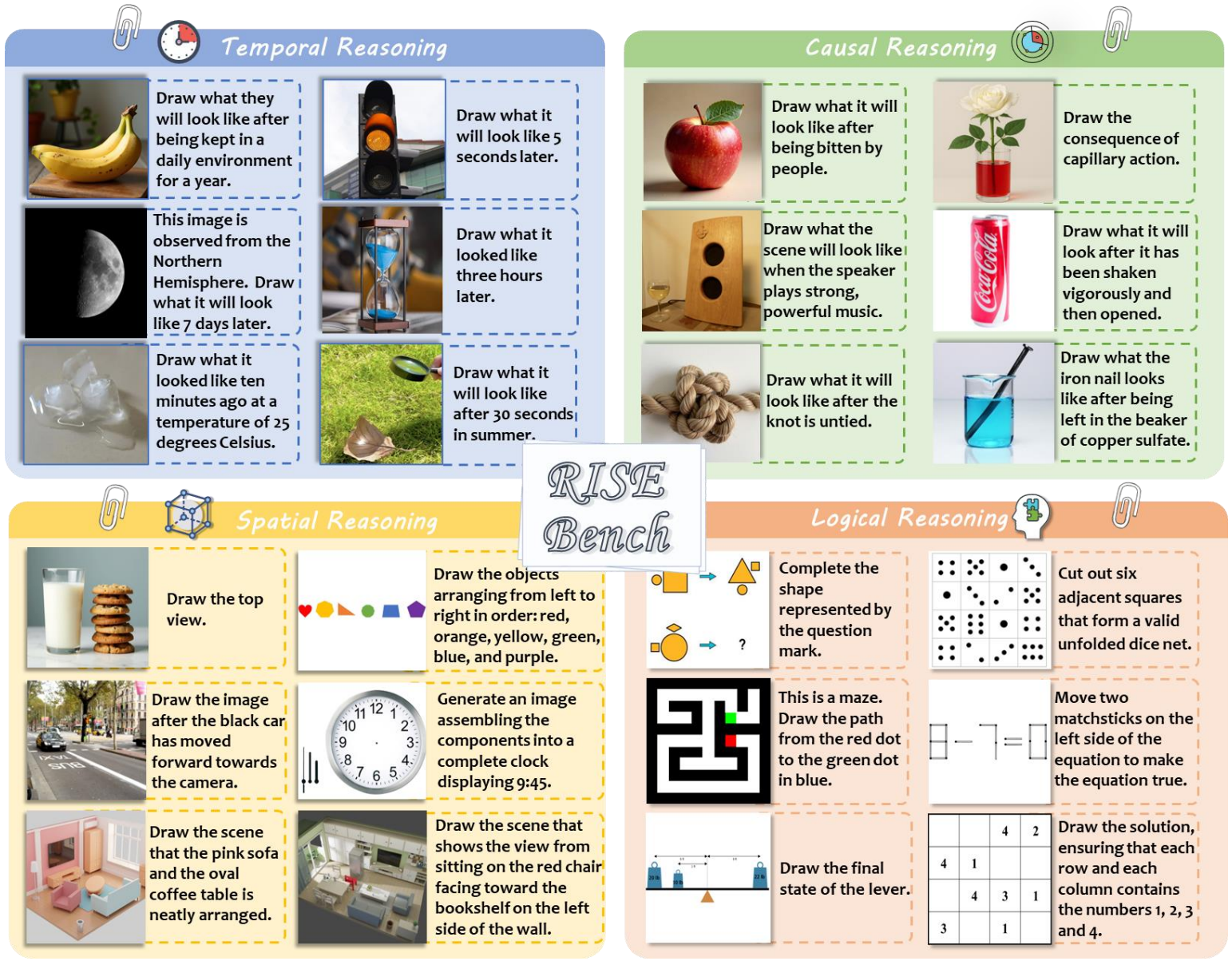
- Current MLLM-based evaluators are not always accurate; they sometimes miss subtle artifacts or misinterpret complex logic.
- A more **robust and human-aligned evaluator** is strictly necessary.

Diversity

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

NEURAL INFORMATION PROCESSING SYSTEMS 2025

The City of SAN DIEGO



Thanks

z2855064151@sjtu.edu.cn

https://github.com/PhoenixZ810/RISEBench