NEURAL INFORMATION PROCESSING SYSTEMS 2025

The City of SAN DIEGO

RISEBench: Benchmarking Reasoning-Informed ViSual Editing

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

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†

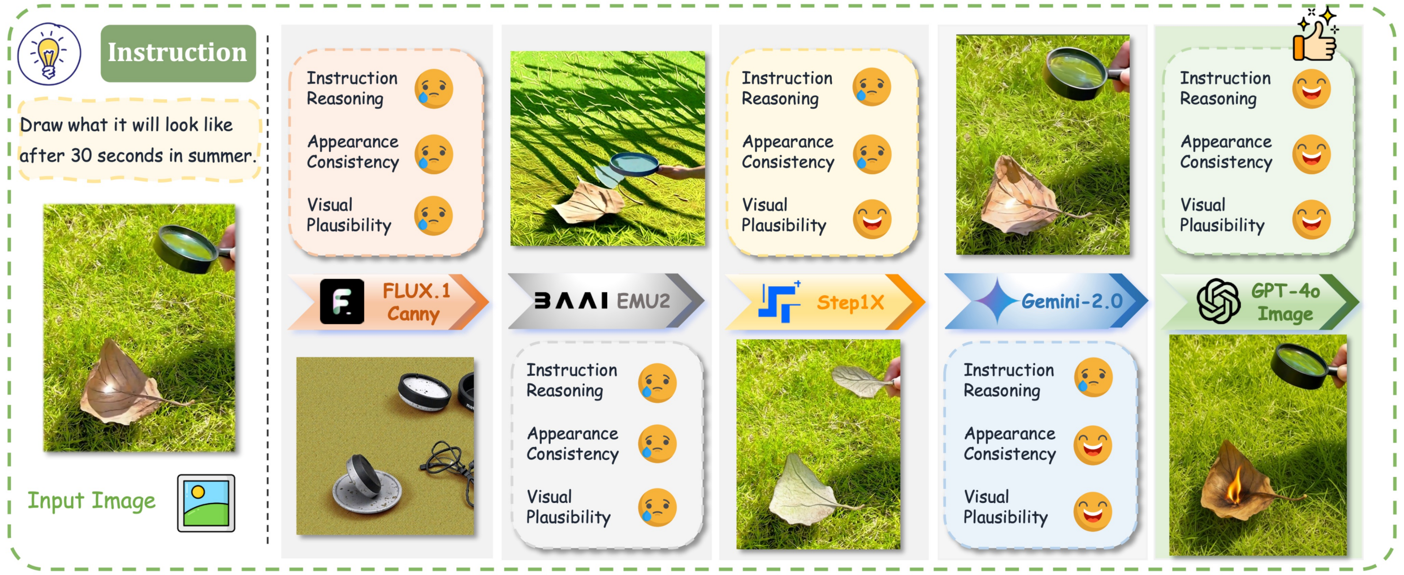
Unified LMMs are popping up (Und&Gen)

Nano Banana Pro   GPT-Image-1   BAGEL   Qwen-Image   ......

RQ: How to benchmark the **emerging capability** of **unified LMMs**?

A: The ability of **Reasoning-Informed Visual Editing** is the key.

An example of RISE (T&I ⇨ I)



GPT-Image-1 Reasoning

It is **sunny**, a magnifying glass will **concentrate the sun light** and finally **ignite the dry leaf**.

Generations are evaluated w. three key dims

1. **Instruction Reasoning:** The model should accurately understand & execute the give instruction.
2. **Appearance Consistency:** Visual elements unrelated to the instruction remain unchanged.
3. **Visual Plausibility:** Generated images should feature high visual quality & realism.

Question Categories & Qualitative Results

Temporal Reasoning

Inst: Draw what it will look like 5 seconds later.

Spatial Reasoning

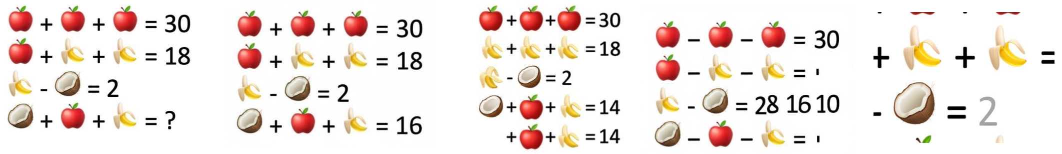
Inst: Draw given objects from large to small (left to right).

Causal Reasoning

Inst: Draw what they will look like when fully inflated.

Logical Reasoning

Inst: Replace the question mark with the correct answer.



Input   Nano Banana Pro   GPT-Image-1   BAGEL   Qwen-Image

Quantitative Results

Model	Temporal	Causal	Spatial	Logical	Overall
Nano Banana Pro	**41.2**	**61.1**	**48.0**	**37.6**	**47.2**
Nano Banana	25.9	47.8	37.0	18.8	32.8
GPT-Image-1	34.1	32.2	37.0	10.6	28.9
GPT-Image-1-mini	24.7	28.9	33.0	9.4	24.4
Bagel w. CoT	5.9	17.8	21.0	1.2	11.9
Seedream 4.0	12.9	12.2	11.0	7.1	10.8
Qwen-Image-Edit	4.7	10.0	17.0	2.4	8.9
Flux.1-Kontext-Dev	2.3	5.5	13.0	1.2	5.8

PS: We insist high standards during evaluation: model needs to achieve 5/5 on all 3 dimensions to pass a test case.

Key Takeaways

1. RISEBench is challenging benchmark for Unified LMMs, even Nano Banana Pro achieves < 50% acc on it.
2. OpenSource LMMs heavily lag behind API ones: Nano Banana Pro 47% vs. Bagel 12%

More Info

Project Homepage   Image Gallery   VLM-EvalKit