

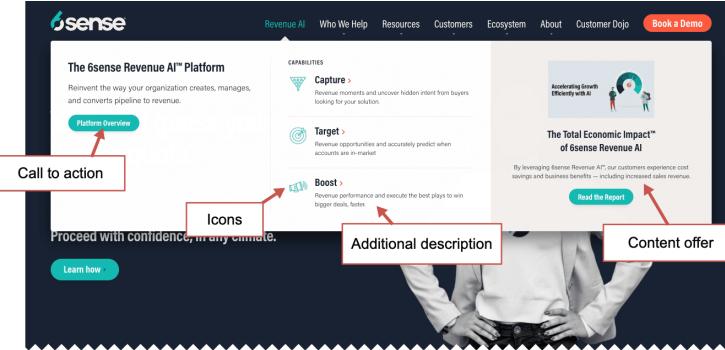


# Scaling Multimodal Evaluation to over 500 Real-World Tasks

Jiacheng Chen<sup>\*†</sup>, Tianhao Liang\*, Sherman Siu\*, Zhengqing Wang, Kai Wang,  
Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, Dongfu Jiang,  
Xuan He, Yuan Liu, Hexiang Hu<sup>‡</sup>, Xiang Yue<sup>‡</sup>, Wenhui Chen<sup>\*†</sup>

Presenter: Wenhui Chen

# Wide application of Multimodal



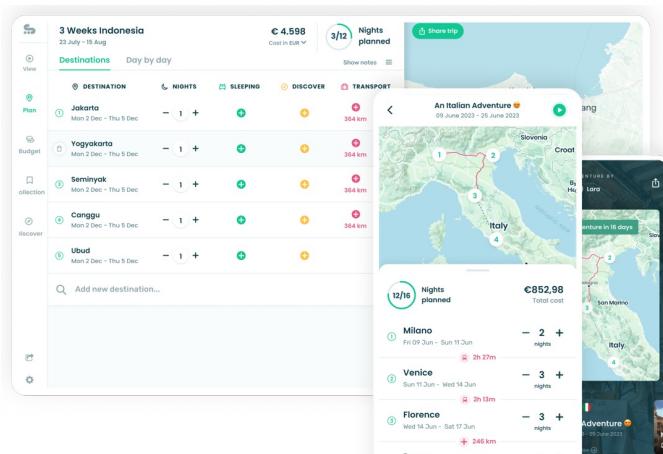
Web navigation



Game playing



Visual navigation



Travel planning



CCTV



Chatbot

# Existing Benchmarks



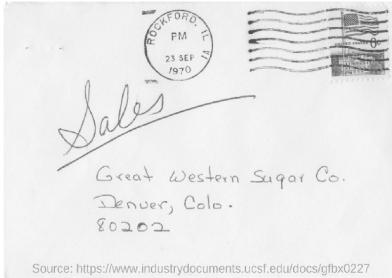
What color are her eyes?  
What is the mustache made of?



How many slices of pizza are there?  
Is this a vegetarian pizza?



Is this person expecting company?  
What is just under the tree?



Source: <https://www.industrydocuments.ucsf.edu/docs/gfbx0227>

Q: Mention the ZIP code written?

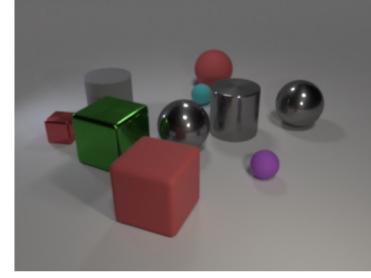
A: 80202

Q: What date is seen on the seal at the top of the letter?

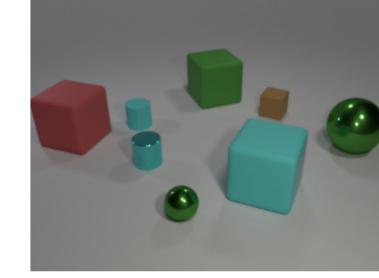
A: 23 sep 1970

Q: Which company address is mentioned on the letter?

A: Great western sugar Co.



(a) (i) Remove all gray spheres. How many spheres are there? (3), (ii) Take away 3 cubes. How many objects are there? (7), (iii) How many blocks must be removed to get 1 block? (2)



(b) Take away all large green metallic spheres. Now remove all cyan objects. How many objects are left? (4)

## Challenges:

- Not Task-driven
- Gap with Real-world applications
- Fundamental
- Imbalance
  - perception-dominated
  - knowledge-dominated
  - reasoning-dominated

# Challenge: Comprehensive Evaluation

- The datasets are highly homogeneous, with a uniform instruction and uniform answer format and types.

Dataset	Annotation	Source	Input	Output	#Metrics	#Tasks
VQA-v2 (Antol et al., 2015)	New	Photo	1 Image	Phrase/Bool/Num	1	1
GQA (Hudson & Manning, 2019)	New	Photo	1 Image	Phrase/Bool/Num	1	1
VizwizVQA (Gurari et al., 2018)	New	Photo	1 Image	Phrase/Bool/Num	1	1
ChartQA (Masry et al., 2022)	New	Chart	1 Image	Bool/Num	1	1
AI2D (Kembhavi et al., 2016)	New	Diagram	1 Image	Multi-choice (MC)	1	1
GeoQA (Chen et al., 2021)	New	Geometry	1 Image	Multi-choice (MC)	1	1
NLVR <sup>2</sup> (Suhr & Artzi, 2019)	New	Photo	2 Images	Bool	1	1
InfoVQA (Mathew et al., 2022)	New	Infographics	1 Image	Phrase/Bool/Num	1	1
DocVQA (Mathew et al., 2021)	New	Document	1 Image	Phrase/Bool/Num	1	1
OCR-VQA (Mishra et al., 2019)	New	Book covers	1 Image	Phrase	1	1

# Challenge: Comprehensive Evaluation

- To lower the burden of evaluation, many advanced multi-task multimodal benchmarks employ multiple-choice questions, deviating from the real use case.

Dataset	Annotation	Source	Input	Output	#Metrics	#Tasks
MMBench (Liu et al., 2023b)	Repurposed	Existing	1 Image	Multi-choice (MC)	1	20
MME (Yin et al., 2023)	Repurposed	Existing	1 Image	Multi-choice (MC)	1	14
MMStar (Chen et al., 2024a)	Repurposed	Existing	1 Image	Multi-choice (MC)	1	18
MMVet (Yu et al., 2024b)	Repurposed	Existing	1 Image	Multi-choice (MC)	1	16
ScienceQA (Lu et al., 2022)	New	K12 Books	$\leq 1$ Image	Multi-choice (MC)	1	26
MMMU (Yue et al., 2024a)	New	Diverse	$\geq 1$ Image	Multi-choice (MC)	1	30
MUIRBench (Wang et al., 2024a)	Hybrid	Existing	$> 1$ Image	Multi-choice (MC)	1	12
MileBench (Song et al., 2024)	Repurposed	Existing	$> 1$ Image	MC or FF	2	28
VideoMME (Fu et al., 2024a)	New	Youtube	Video	Multi-choice (MC)	1	30
MVBench (Li et al., 2024d)	Repurposed	Existing	Video	Multi-choice (MC)	1	20
MMTBench (Ying et al., 2024)	Repurposed	Existing	Image/Videos	Multi-choice (MC)	1	162

# Landscape of Existing Evaluation

- Qwen-VL evaluated on more than 30 benchmarks

Image Benchmarks

Benchmark	Previous SoTA (Open-source LVLM)	Claude-3.5 Sonnet	GPT-4o	Qwen2-VL-72B ( )	Qwen2-VL-7B ( )	Qwen2-VL-2B ( )
MMMU <sub>val</sub>	58.3	68.3	<b>69.1</b>	64.5	54.1	41.1
MMMU-Pro	46.9	51.5	<b>51.9</b>	46.2	43.5	37.6
DocVQA <sub>test</sub>	94.1	95.2	92.8	<b>96.5</b>	94.5	90.1
InfoVQA <sub>test</sub>	82.0	-	-	<b>84.5</b>	76.5	65.5
ChartQA <sub>test</sub>	88.4	<b>90.8</b>	85.7	88.3	83.0	73.5
TextVQA <sub>val</sub>	84.4	-	-	<b>85.5</b>	84.3	79.7
OCRBench	852	788	736	<b>877</b>	845	794
MTVQA	17.3	25.7	27.8	<b>30.9</b>	25.6	18.1
VCR <sub>en easy</sub>	84.67	63.85	91.55	<b>91.93</b>	89.70	81.45
VCR <sub>zh easy</sub>	22.09	1.0	14.87	<b>65.37</b>	59.94	46.16
RealWorldQA	72.2	60.1	75.4	<b>77.8</b>	70.1	62.9
MME <sub>sum</sub>	2414.7	1920.0	2328.7	<b>2482.7</b>	2326.8	1872.0
MMBench-EN <sub>test</sub>	<b>86.5</b>	79.7	83.4	<b>86.5</b>	83.0	74.9
MMBench-CN <sub>test</sub>	86.3	80.7	82.1	<b>86.6</b>	80.5	73.5
MMBench-V1.1 <sub>test</sub>	85.5	78.5	82.2	<b>85.9</b>	80.7	72.2
MMT-Bench <sub>test</sub>	63.4	-	65.5	<b>71.7</b>	63.7	54.5
MMStar	67.1	62.2	63.9	<b>68.3</b>	60.7	48.0
MMVet <sub>GPT-4-Turbo</sub>	65.7	66.0	69.1	<b>74.0</b>	62.0	49.5
HallBench <sub>avg</sub>	55.2	49.9	55.0	<b>58.1</b>	50.6	41.7
MathVista <sub>testmini</sub>	67.5	67.7	63.8	<b>70.5</b>	58.2	43.0
MathVision	16.97	-	30.4	25.9	16.3	12.4

Video Benchmarks

Benchmark	Previous SoTA (Open-source LVLM)	Gemini 1.5-Pro	GPT-4o	Qwen2-VL-72B ( )	Qwen2-VL-7B ( )	Qwen2-VL-2B ( )
MVBench	69.6	-	-	<b>73.6</b>	67.0	63.2
PerceptionTest <sub>test</sub>	66.9	-	-	<b>68.0</b>	62.3	53.9
EgoSchema <sub>test</sub>	62.0	63.2	72.2	<b>77.9</b>	66.7	54.9
Video-MME (wo/w subs)	66.3/69.6	<b>75.0/81.3</b>	71.9/77.2	71.2/77.8	63.3/69.0	55.6/60.4

Agent Benchmarks

	Benchmark	Metric	Previous SoTA	GPT-4o	Qwen2-VL-72B
General	FnCall <sup>[1]</sup>	TM	-	90.2	<b>93.1</b>
		EM	-	50.0	<b>53.2</b>
Game	Number Line	SR	<b>89.4<sup>[2]</sup></b>	91.5	<b>100.0</b>
		SR	<b>40.2<sup>[2]</sup></b>	34.5	<b>42.6</b>
Android	BlackJack	SR	<b>50.0<sup>[2]</sup></b>	85.5	<b>100.0</b>
		SR	<b>2.6<sup>[2]</sup></b>	3.0	<b>4.5</b>
AI2THOR	AITZ	TM	<b>83.0<sup>[3]</sup></b>	70.0	<b>89.6</b>
		EM	<b>47.7<sup>[3]</sup></b>	35.3	<b>72.1</b>
VLN	ALFRED <sub>valid-unseen</sub>	SR	<b>67.7<sup>[4]</sup></b>	-	<b>67.8</b>
		GC	<b>75.3<sup>[4]</sup></b>	-	<b>75.8</b>
	R2R <sub>valid-unseen</sub>	SR	<b>79.0</b>	43.7 <sup>[5]</sup>	51.7
		SR	<b>61.0</b>	31.6 <sup>[5]</sup>	31.0
	REVERIE <sub>valid-unseen</sub>	SR	<b>61.0</b>	31.6 <sup>[5]</sup>	31.0

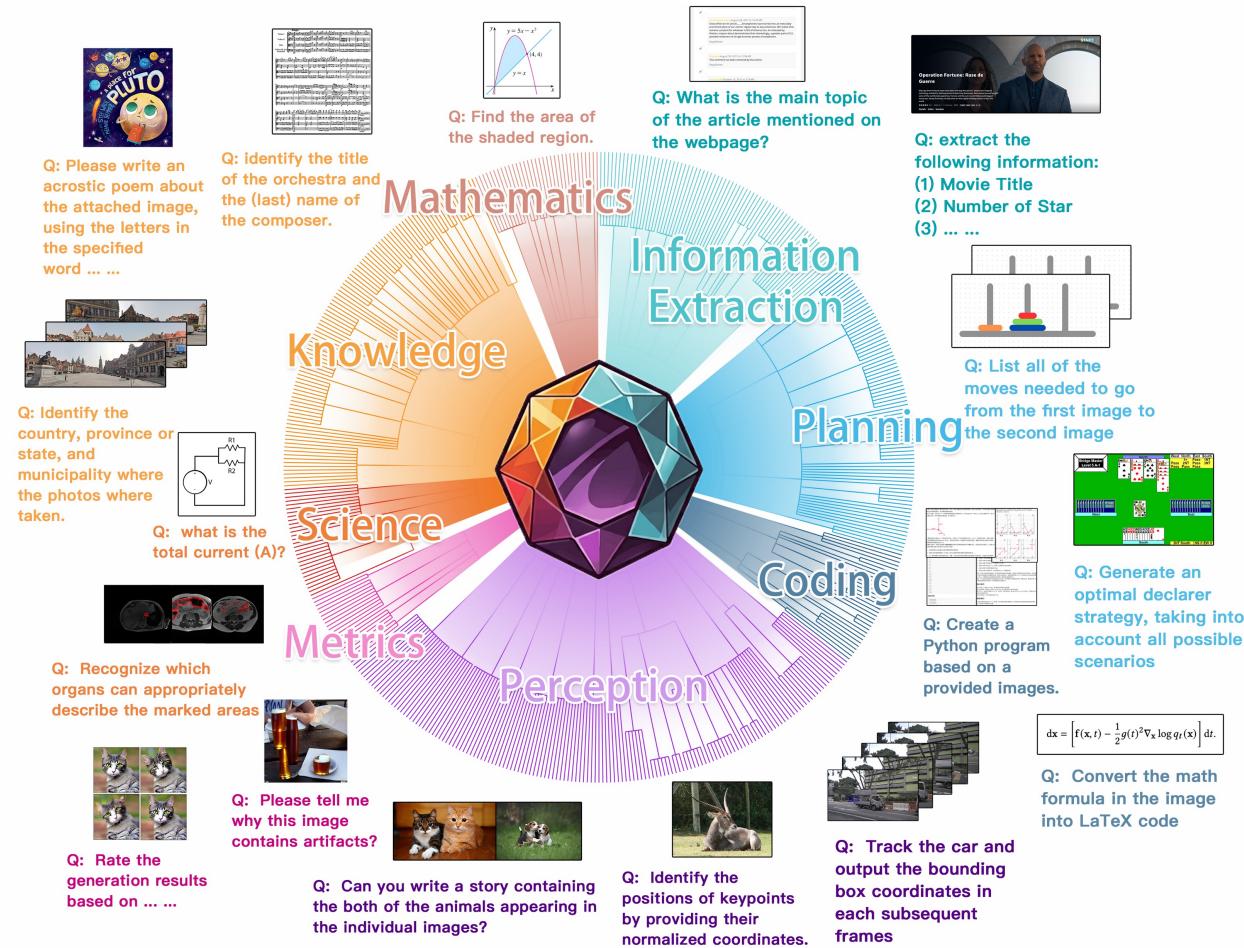
## Challenges:

- Expensive
- Complex
- Intractable
- Irreproducible

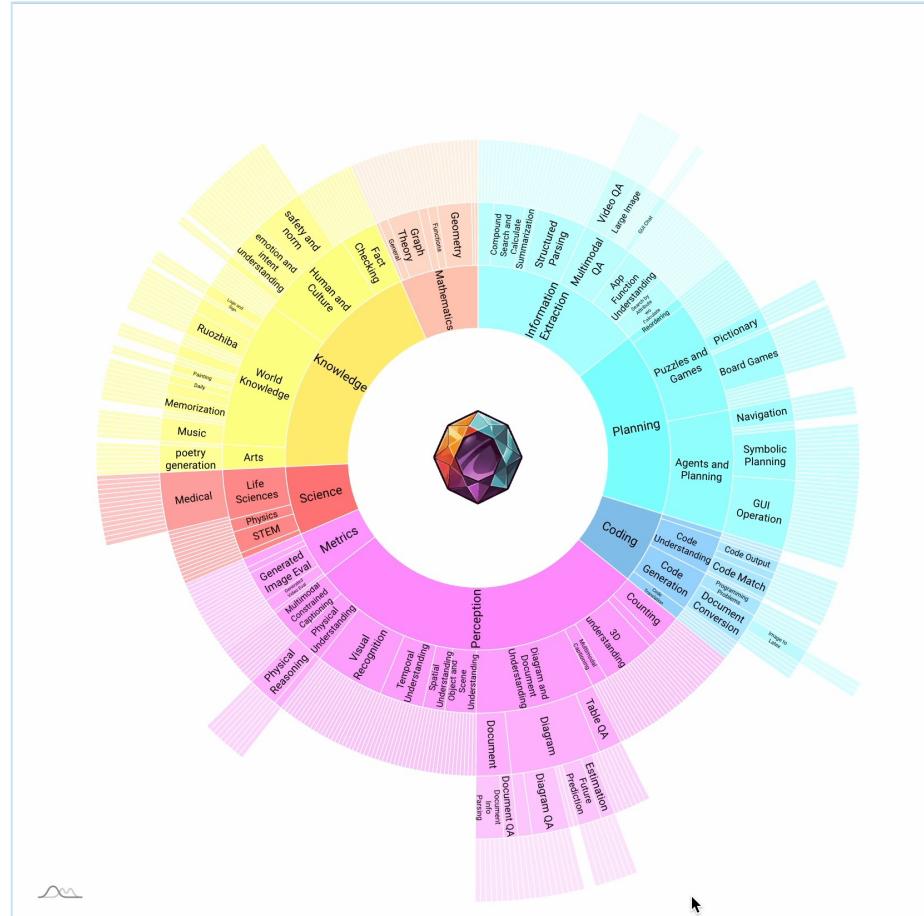
# MEGA-BENCH

- 500+ diverse multimodal tasks with a unified protocol.
- Scales up multi-modal evaluation to maximize the task coverage and the diversity in model outputs.
- Highly diverse tasks from annotator contribution.
- Optimized inference cost, but still provide comprehensive breakdown analysis

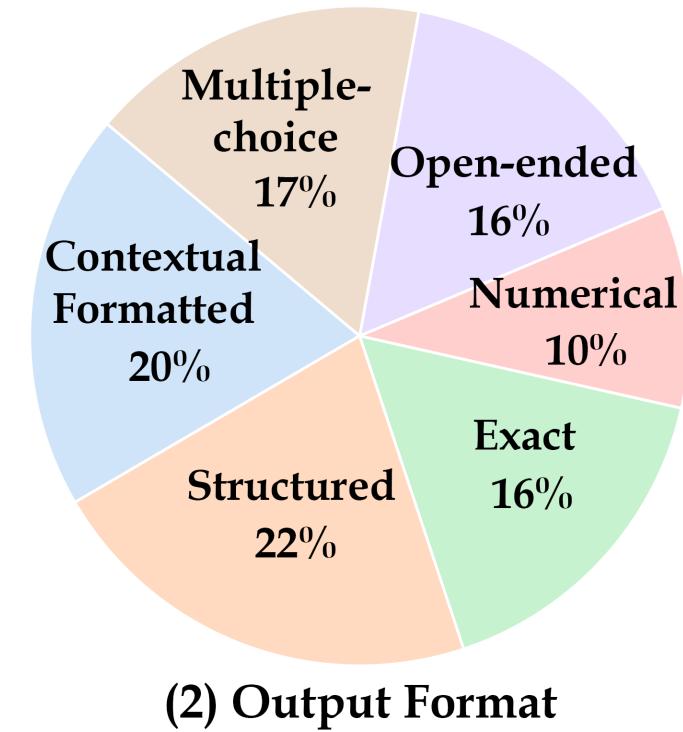
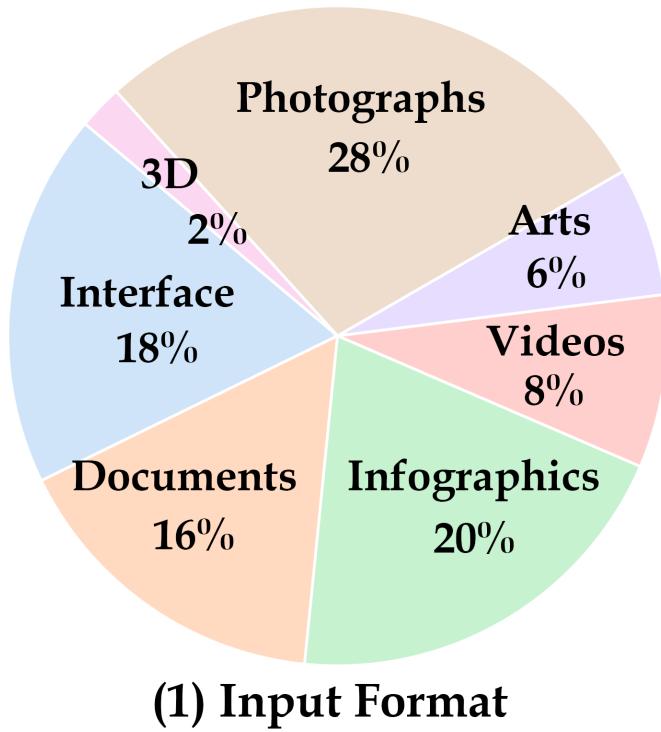
# Coverage



# Task Taxonomy Tree Construction



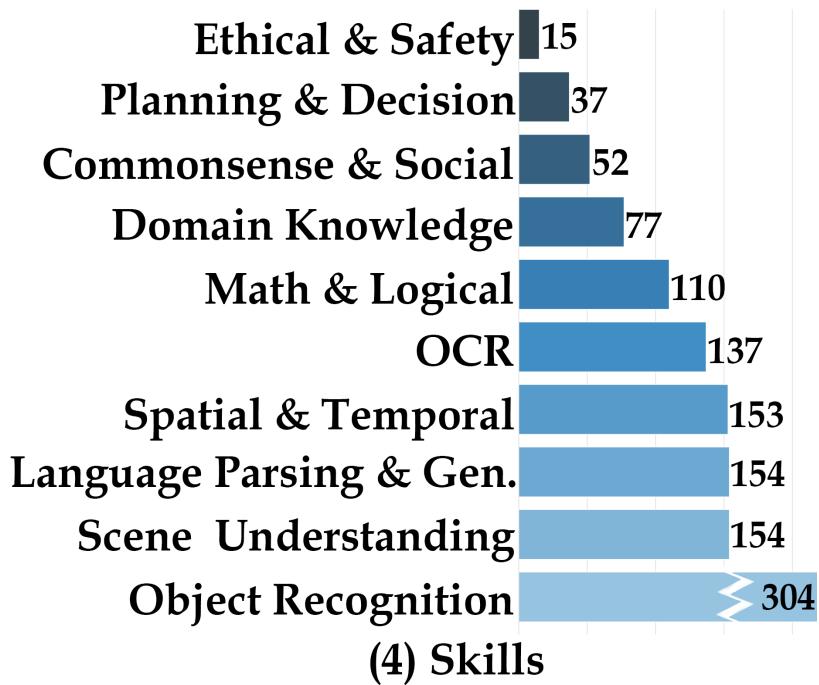
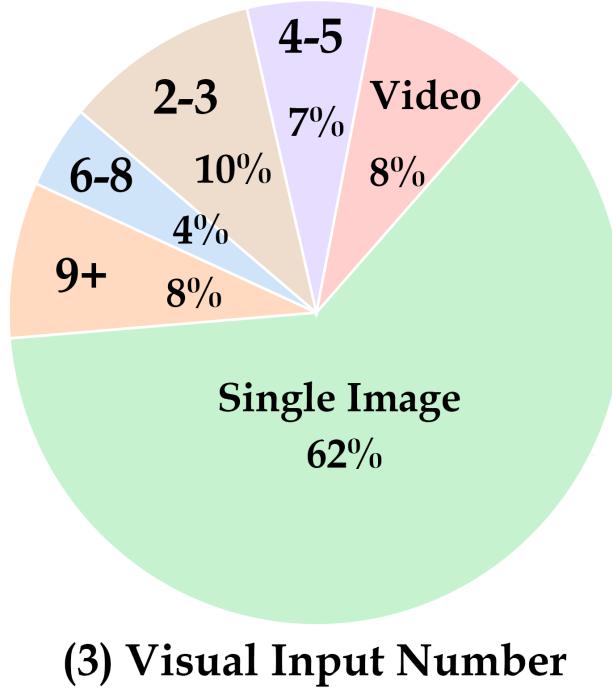
# Diverse input/output formats



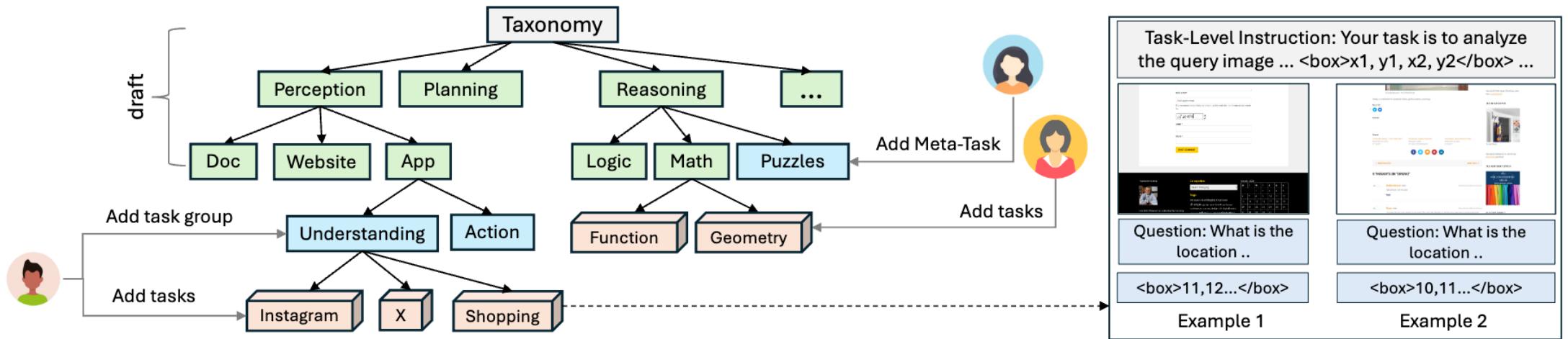
# Diverse evaluation metrics

<p><b>Output Type: Structured</b></p> <pre>```python def f(x, a3, a2, a1, a0):     return a3 * x**3 + a2 * x**2 + a1 * x + a0 def bisection_method(a3, a2, a1, a0, a, b, tolerance=1e-6):     while (b - a) / 2.0 &gt; tolerance:         ... print(f"root: {f(2)}")</pre>  <p><b>Metric:</b> Program Judge    Code Programming Test (easy)</p>	<p><b>Output Type: Structured</b></p> <p>Answer:</p> <pre>(grasp left shot4) (fill-shot shot4 ingredient1 left right dispenser1) (pour-shot-to-clean-shaker shot4 ingredient1 shaker1 left l0 l1) (clean-shot shot4 ingredient1 left right) (fill-shot shot4 ingredient3 left right dispenser3) (pour-shot-to-used-shaker shot4 ingredient3 shaker1 left l1 l2) ... (shake cocktail2 ingredient2 ingredient3 shaker1 left right) (pour-shaker-to-shot cocktail2 shot3 left shaker1 l2 l1)</pre>  <p><b>Metric:</b> Symbolic Planning    Symbolic Planning (Barman)</p>	<p><b>Output Type: Structured</b></p> <p>Answer:</p> <pre>\begin{aligned} \text{\&amp;} \nu_\\mu, \text{\&amp;} \Phi(A, U) = \int_P \\ (\mu \cdot \text{\&amp;} \Phi) \\ (A \cap \varphi_p) \\ (X, U \cap \varphi_p) \\ (X) \cdot d \varphi_p \end{aligned}</pre>  <p><b>Metric:</b> LaTeX Expr Equality    LaTeX Complex Formula Conversion</p>	<p><b>Output Type: Contextual Formatted</b></p> <p>Answer:</p> <pre>426195783 357846192 198237546 ... 735418629 241659738</pre>  <p><b>Metric:</b> Simple String Match    Sudoku</p>
<p><b>Output Type: Numerical</b></p> <p>Answer: [[0.28, 0.48], [0.0, 0.0], [0.28, 0.55], [0.38, 0.55], [0.75, 0.45], [0.45, 0.65], [0.45, 0.75], [0.45, 0.85], [0.55, 0.65], [0.55, 0.75], [0.55, 0.85], [0.7, 0.55], [0.75, 0.65], [0.75, 0.85], [0.75, 0.55], [0.8, 0.65], [0.8, 0.85]]</p>  <p><b>Metric:</b> Sequence Coordinates Similarity    Animal Pose Estimation</p>	<p><b>Output Type: Numerical</b></p> <p>Answer:</p> <pre>[{"x": 0.02, "y": 0.32, "z": 0.12, "score": 0.42}, {"x": 0.0, "y": 0.32, "z": 0.1, "score": 0.42}, {"x": 0.0, "y": 0.32, "z": 0.1, "score": 0.42}, {"x": 0.0, "y": 0.32, "z": 0.1, "score": 0.42}]</pre>  <p><b>Metric:</b> Sequence BBox IoU    Multi-camera Car Tracking</p>	<p><b>Output Type: Structured</b></p> <p>Answer:</p> <pre>{"has_dot_r1c1": true, "has_dot_r1c2": true, "has_dot_r1c3": true, ... "has_dot_r3c4": false, "has_dot_r3c5": false}</pre>  <p><b>Metric:</b> Dict Equality    Mensa IQ Test</p>	<p><b>Output Type: Open-ended</b></p> <p>Answer:</p> <p>West leads the 2S against 5C.  South has 10 top tricks (5 spades, 1 heart, 1 club, and 3 diamonds).  South needs to ...  The key is to ensure that the club suit is played in a way that maximizes the chances of establishing the second club trick while maintaining control of the hand.</p>  <p><b>Metric:</b> GPT-4o as Judge    Bridge Strategies (Worldclass)</p>

# Diverse Skills



# Benchmark Preparation



- Create draft taxonomy tree with the first two levels
- Annotators are guided by the draft tree to keep balanced task distribution
- The taxonomy keeps being refined/edited in annotation process

# Benchmark Preparation

## annotation.json

**Task instruction:** <task\_instruction>

**Global media:** [<global\_media>, <global\_media>, ... ...]

### Example

#### Example 1

**Media paths:** [<example\_media>, <example\_media>, ... ...]

**Answers:** <answer\_field>: <answer>  
<answer\_field>: <answer>  
... ...

**Example question:** <example\_question>

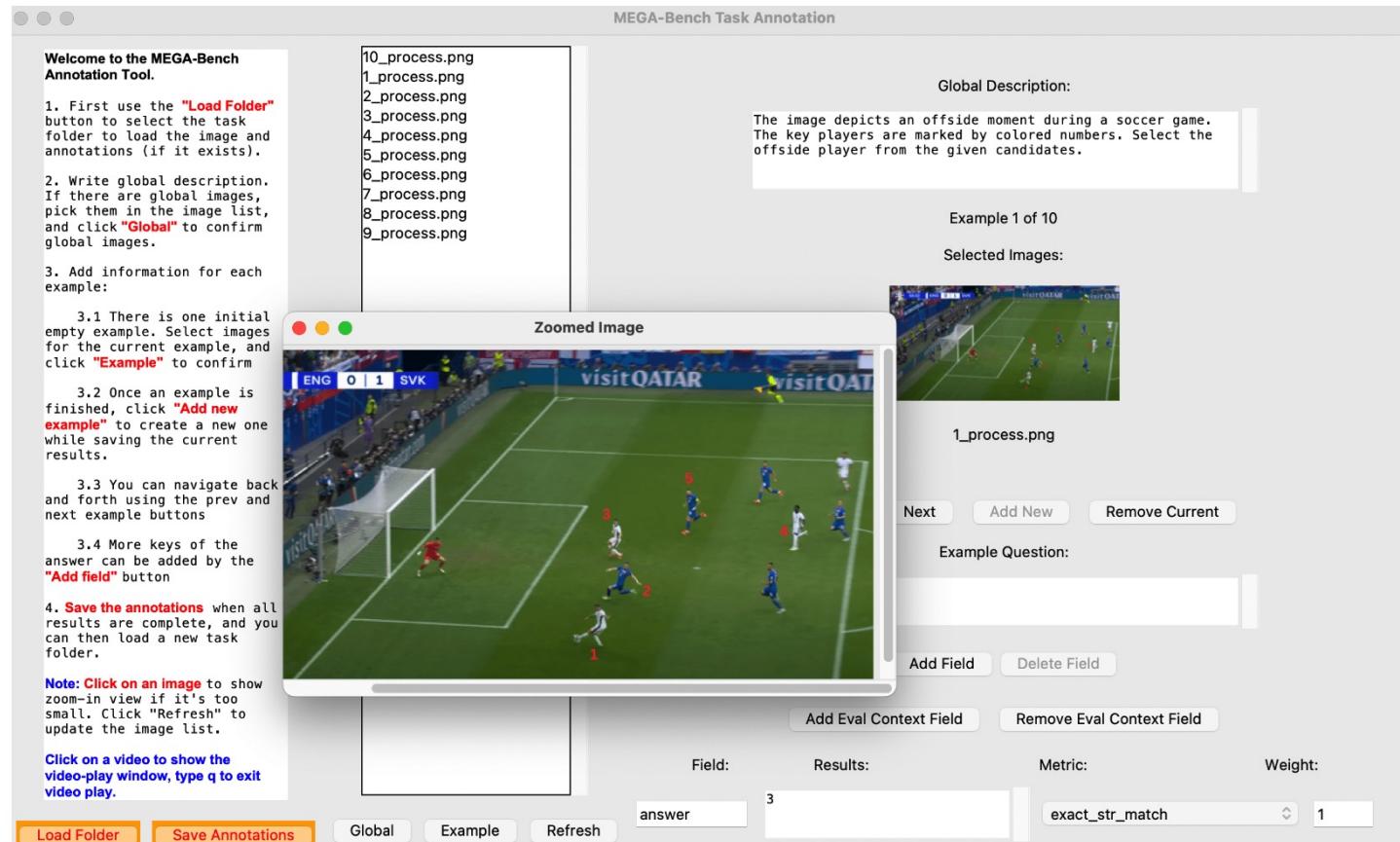
#### Example 2

#### Example 3

... ...

All task annotations follow a  
unified JSON format

# Benchmark Preparation



A GUI annotation tool helps all  
annotators organize the tasks  
into the annotation format

# Benchmark Construction

- [Add new tasks weather map, transit map, and mindmap \(3 tasks\)](#) [task submission](#)  
#311 by kaiw7 was merged on Aug 25 • Approved
- [Add task: Multiple states identify europe \(for geographic bias study\)](#) [task submission](#)  
#310 by XMHZZ2018 was merged on Aug 25 • Approved
- [task: music\\_sheet\\_name](#) [task submission](#)  
#309 by hexuan21 was merged on Aug 25 • Approved
- [task: music\\_sheet\\_note\\_count](#) [task submission](#)  
#308 by hexuan21 was merged on Aug 25 • Approved
- [Music sheet format qa](#) [task submission](#)  
#307 by hexuan21 was merged on Aug 25 • Approved
- [Revamp old tasks and add new autorater tasks \(2 new tasks\)](#) [need better metric](#) [task submission](#)  
#305 by wenhuchen was merged on Aug 25 • Approved
- [Add the Kakuro \(5x5\) task.](#) [task submission](#)  
#292 by shermansiu was merged on Aug 23 • Approved
- [Add rebus task.](#) [task submission](#)  
#287 by shermansiu was merged on Aug 23 • Approved
- [Add mini 5x5 crossword task.](#) [task submission](#)  
#286 by shermansiu was merged on Aug 23 • Approved
- [Constrained generation \(3 tasks\)](#) [need better metric](#) [task submission](#)  
#284 by wenhuchen was merged on Aug 23 • Approved

Annotators submit tasks via  
Pull Requests (PRs)

Task reviewers check the  
quality and merge

# Benchmark Construction

- (1) Completely created by our annotators
- (2) Originated from existing datasets or benchmarks, but augmenting the task with different output formats, task instructions, input constraints, etc.
- (3) Directly converted from existing datasets/benchmarks
  - First two types are encouraged
  - Task reviewers strictly control the number of the third type

# Quality Control

- Evaluate new tasks and visualize the results on a web page
- Annotators can check their own tasks and fix potential bugs

## Task 393: poetry\_haiku

### Description:

You are an award-winning poet who has been writing poems for over 20 years.  
Please write a Haiku poem about the attached image. No title is needed.

### Scores by Model:

#### Scores:

- gpt-4o: 14.0 / 15
- gemini-1.5-pro: 13.0 / 15
- claude3.5: 6.0 / 15
- internvl-76B: 0.0 / 15

### Metric Configuration by Model:

#### Raw metric config:

```
{  
  "field_score_function": {  
    "#poem": "constrained_generation"  
  },  
  "aggregation": {  
    "function": "mean",  
    "field_weights": {  
      "#poem": 1  
    }  
  },  
  "response_parse_function": "json"  
}
```

# Quality Control

- Check the results of the top proprietary models. If all models get 0, double-check the annotation correctness.

Task 11: TRANCE\_physics\_reasoning\_basic

1 in-context + 17 query examples

(gpt-4o:0.88; gemini-1.5-pro:0.76; claude3.5:0.88; internvl-76B:0.41; )

Task 13: logical\_reasoning\_fit\_pattern

1 in-context + 14 query examples

(gpt-4o:0.29; gemini-1.5-pro:0.14; claude3.5:0.43; internvl-76B:0.21; )

Task 15: knowledge\_sign\_recognition

1 in-context + 9 query examples

(gpt-4o:0.44; gemini-1.5-pro:0.33; claude3.5:0.33; internvl-76B:0.00; )

Task 17: autonomous\_driving\_scene\_analysis

1 in-context + 14 query examples

(gpt-4o:1.00; gemini-1.5-pro:0.71; claude3.5:1.00; internvl-76B:0.93; )

Task 19: geometry\_reasoning\_count\_line\_intersections

1 in-context + 14 query examples

(gpt-4o:0.36; gemini-1.5-pro:0.39; claude3.5:0.50; internvl-76B:0.39; )

Task 12: geometry\_reasoning\_overlapped\_circle

1 in-context + 14 query examples

(gpt-4o:0.75; gemini-1.5-pro:0.61; claude3.5:0.79; internvl-76B:0.46; )

Task 14: GUI\_Act\_Mobile\_tap

1 in-context + 14 query examples

(gpt-4o:0.14; gemini-1.5-pro:0.21; claude3.5:0.36; internvl-76B:0.36; )

Task 16: relative\_depth\_of\_different\_points

1 in-context + 14 query examples

(gpt-4o:0.86; gemini-1.5-pro:0.64; claude3.5:0.71; internvl-76B:0.43; )

Task 18: booking\_web\_rating

1 in-context + 14 query examples

(gpt-4o:1.00; gemini-1.5-pro:0.89; claude3.5:1.00; internvl-76B:0.32; )

Task 20: ishihara\_test

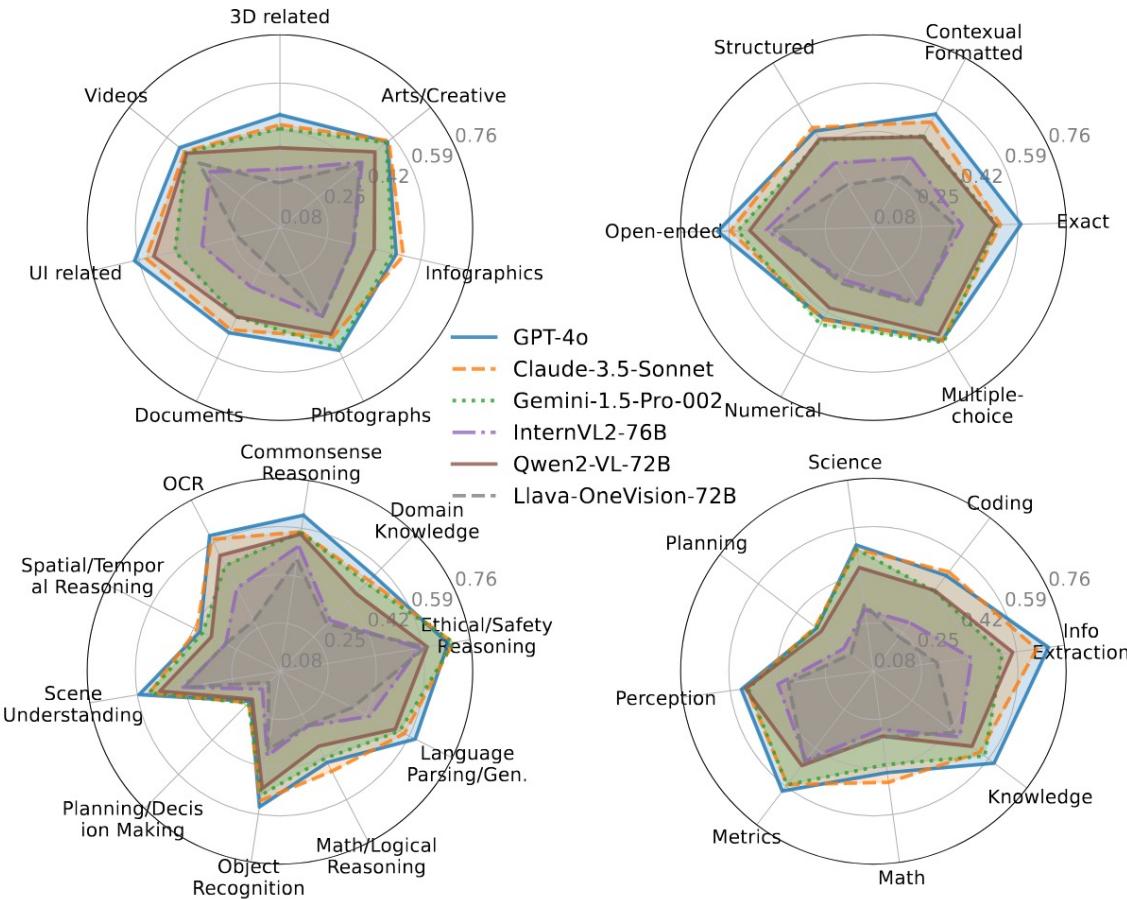
1 in-context + 14 query examples

(gpt-4o:0.56; gemini-1.5-pro:0.31; claude3.5:0.07; internvl-76B:0.00; )

# Experimental Results

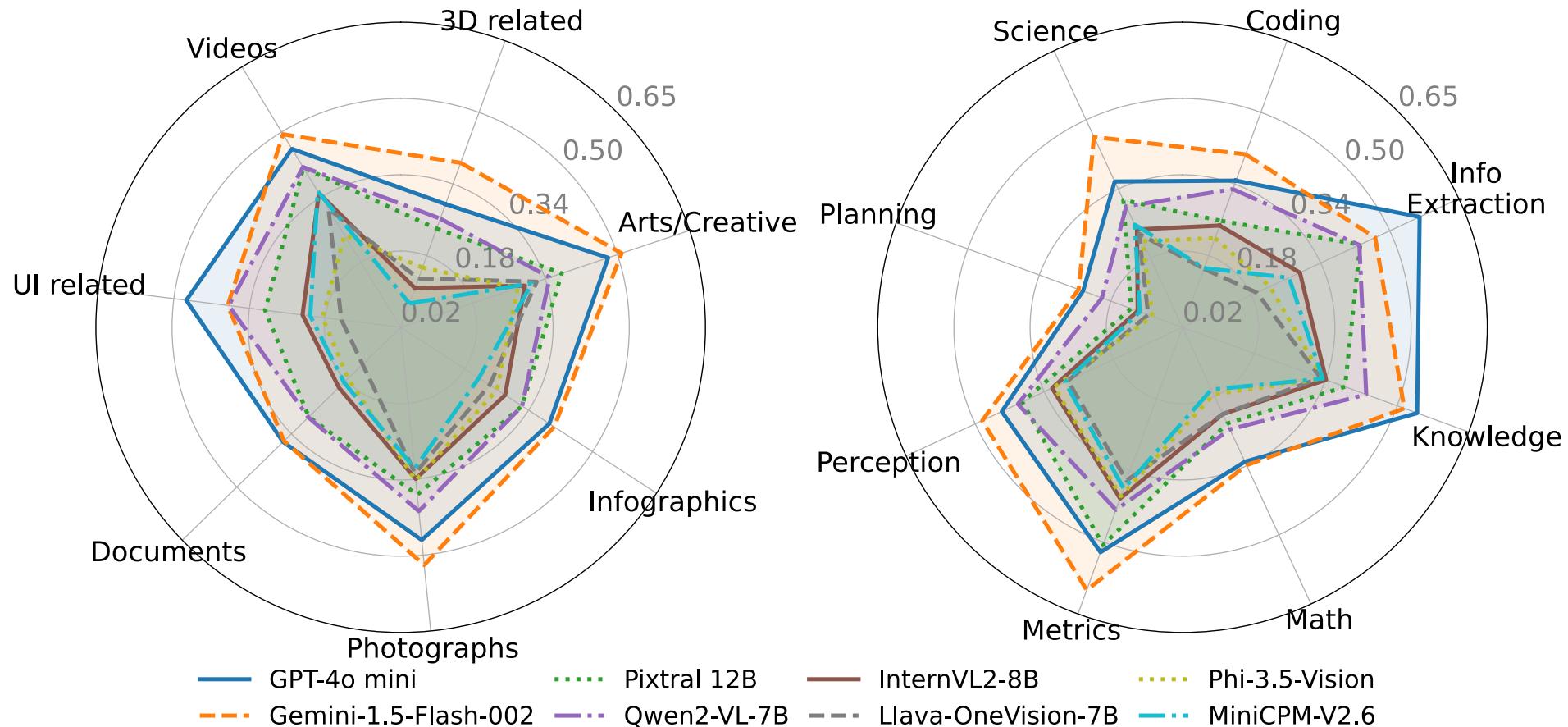
Model	Eval Tier	Open Source	Core (rule-based eval)		Open-ended (GPT eval)	Overall
			w/o CoT	w/ CoT		
Claude-3.5-Sonnet (1022) (Anthropic, 2024b)	Flagship	No	49.20	52.59	65.63	54.26
GPT-4o (0513) (OpenAI, 2024a)	Flagship	No	<b>52.03</b>	<b>52.65</b>	<b>64.78</b>	<b>54.10</b>
Claude-3.5-Sonnet (0620) (Anthropic, 2024a)	Flagship	No	48.63	50.24	63.74	51.97
Gemini-1.5-Pro-002 (Google, 2024b)	Flagship	No	46.89	48.14	58.58	49.48
Gemini-1.5-Flash-002 (Google, 2024b)	Efficiency	No	<b>41.84</b>	<b>41.84</b>	56.91	<b>43.78</b>
GPT-4o mini (OpenAI, 2024b)	Efficiency	No	<u>39.74</u>	<u>40.71</u>	<b>58.65</b>	<u>43.02</u>
Qwen2-VL-72B (Alibaba, 2024)	Flagship	Yes	<b>46.24</b>	<b>45.28</b>	<b>56.40</b>	<b>47.55</b>
InternVL2-Llama3-76B (Chen et al., 2024d)	Flagship	Yes	<u>34.98</u>	<u>35.54</u>	<u>51.93</u>	<u>37.65</u>
LLaVA-OneVision-72B (Li et al., 2024a)	Flagship	Yes	<u>31.96</u>	<u>29.73</u>	<u>45.99</u>	<u>33.77</u>
Qwen2-VL-7B (Alibaba, 2024)	Efficiency	Yes	<b>34.73</b>	<b>32.84</b>	43.96	<b>35.91</b>
Pixtral 12B (Mistral, 2024)	Efficiency	Yes	<u>31.87</u>	<u>31.32</u>	<b>45.66</b>	<u>33.64</u>
Aria (Li et al., 2024d)	Efficiency	Yes	<u>30.49</u>			
InternVL2-8B (Chen et al., 2024d)	Efficiency	Yes	25.92	24.06	39.79	27.71
Phi-3.5-Vision (Abdin et al., 2024)	Efficiency	Yes	23.24	22.95	39.48	25.33
MiniCPM-V2.6 (Yao et al., 2024)	Efficiency	Yes	22.84	22.90	41.73	25.32
LLaVA-OneVision-7B (Li et al., 2024a)	Efficiency	Yes	22.39	21.35	33.98	23.88
Llama-3.2-11B (Meta, 2024)	Efficiency	Yes	10.04	15.98	31.73	18.01
Idefics3-8B-Llama3 (Laurençon et al., 2024)	Efficiency	Yes	11.12	8.96	32.11	13.82

# Breakdown analysis: flagship models

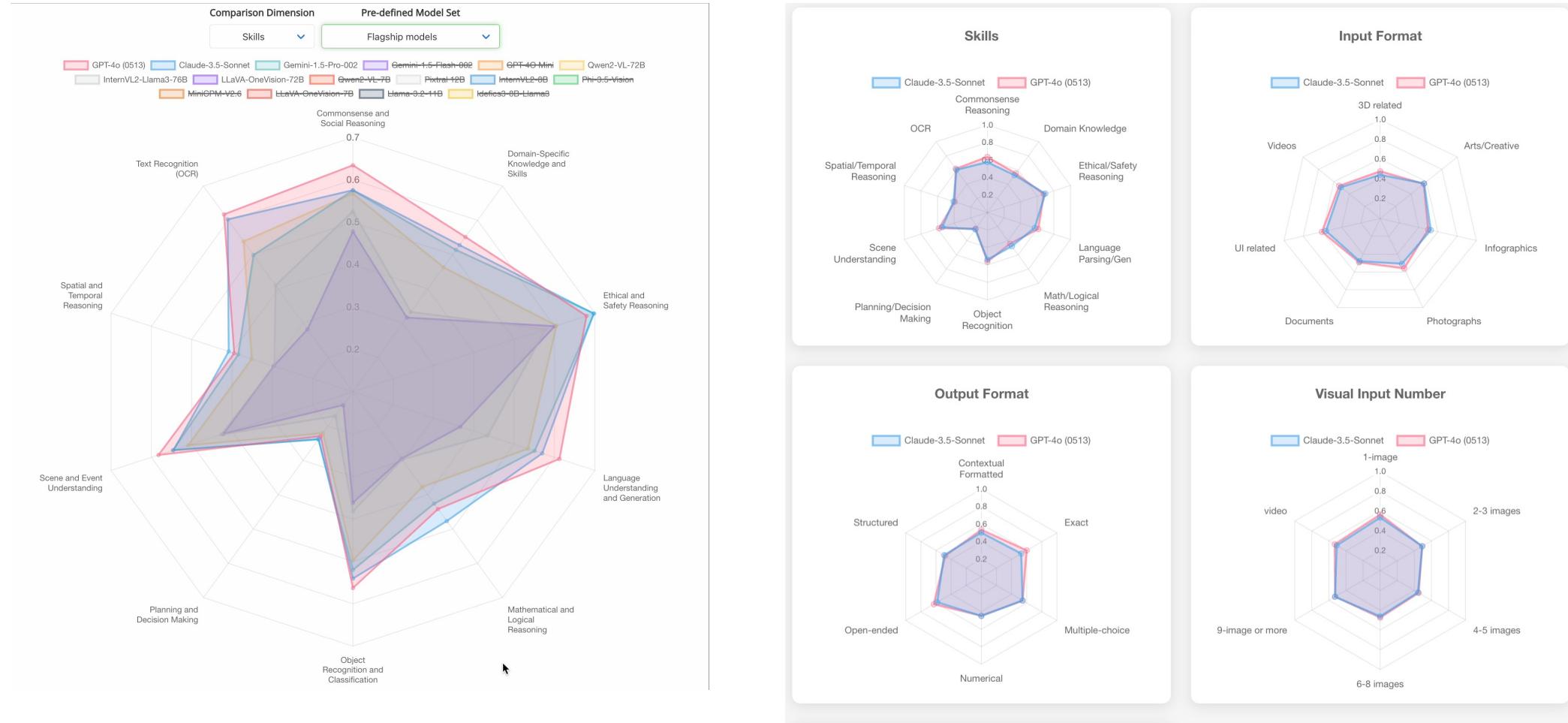


- GPT-4o wins over Claude-3.5-Sonnet in most multimodal tasks, except in tasks related to math, coding, and planning
- Qwen2VL is dominant in open-source flagship models. Results on some dimensions are close to top proprietary models

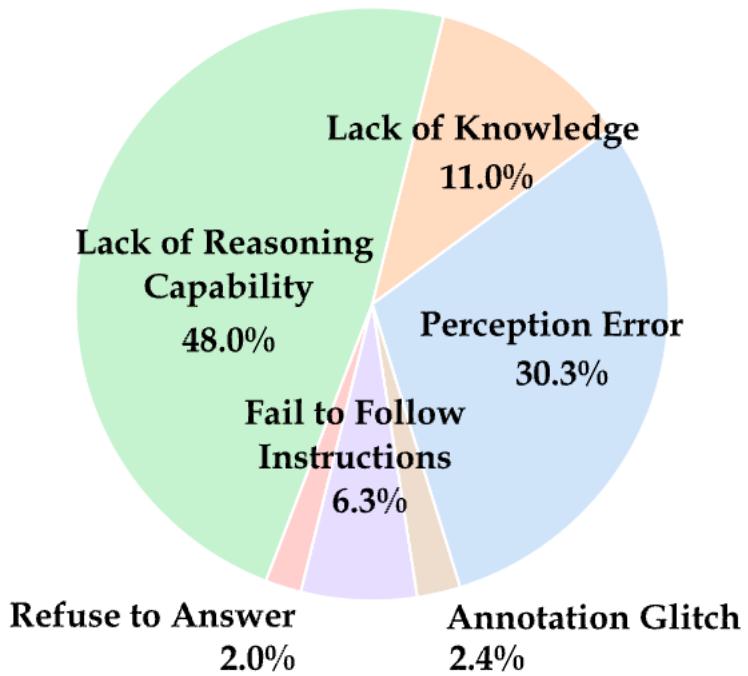
# Breakdown analysis: efficiency models



# Breakdown analysis: interactive



# Error Analysis



- The lack of various reasoning capability is the major failure mode for most models.
- Perception is also a major source of errors
- Sometimes the model has trouble following the instruction on output format

GPT-4o, a subset of 255 tasks

# Error Analysis: reasoning

## Coding: Code Debugging

### Task Name: Code Error Line Identification

Given a piece of program code, identify the most obvious error. Output the line number on the left where the error occurs. Attention: Do not invoke any external compilers or related programs.

Question: Where the error occurs?

Keywords:

Domain-Specific Knowledge and Skills, Text-Based Images and Documents, Mathematical and Logical Reasoning, Multiple Choice

```
1 void Topsort(int a[NUM][NUM], int TopNum[NUM])
2 {
3     int S[NUM], Indegree[NUM];
4     int Counter = 0, top, n, i, j;
5     int V;
6     top = -1;
7     n=NUM;
8     for (j=0; j<n; j++) {
9         Indegree[j]=0;
10        for (i=0; i<n; i++)
11            if (a[i][j]==0) Indegree[j]++;
12        if ( Indegree[j] == 0 ) S[++top]=j;
13    }
14    while (top>=0) {
15        V = S[top--];
16        TopNum[V] = ++Counter;
17        for (j=0; j<n; j++)
18            if ( a[V][j]!=0 )
19                if (Indegree[j] == 0 ) S[++top]=j;
20    }
21 }
```

Claude-3.5: 1. The function declaration and variable initializations seem correct. 2. The first nested loop (lines 7-12) appears to be properly structured. 3. The while loop (lines 13-19) also seems structurally correct. 4. However, within this while loop, there's an issue on line 18: 5. This error is the most obvious and significant in the code, as it will likely cause runtime errors or incorrect results. Therefore, the most obvious error occurs on line 18.

Answer: 18 (score: 1.0)

Error Category: Lack of Capability

GPT-4o: The most obvious error in the code is the use of `printf` instead of `printf` in line 20. The correct function name should be `printf`.

Answer: 20 (score: 0.0)

Correct Answer: 18

## Coding: Code Understanding

### Task Name: Symbolic Graphics Programs Computer Aided Design

Given a symbolic graphics program, the task is to answer a related question by understanding the program.

Question: How many visible screws or screw holes are there on the CAD object?

Keywords:

Mathematical and Logical Reasoning, Contextual Formatted Text, Domain-Specific Knowledge and Skills, Text-Based Images and Documents

```
SOL;
Arc:(130,26,64,1);Line:(221,126);Arc:(223,128,64,1);
Line:(223,176);Arc:(221,176,64,1);Line:(130,178);
Arc:(128,176,64,1);Line:(128,128);
SOL;
Circle:(139,138,3);
SOL;
Circle:(139,161,3);
SOL;
Circle:(176,161,10);
SOL;
Circle:(212,138,3);
SOL;
Circle:(212,161,3);
Ext:(192,64,192,34,128,62,189,207,128,Newbody,One-sided);
SOL;
Arc:(131,125,64,1);Line:(220,125);Arc:(223,128,64,1);
Line:(223,217);Arc:(220,220,64,1);Line:(131,220);
Arc:(128,217,64,1);Line:(128,128);
SOL;
Circle:(176,173,18);
Ext:(192,64,192,74,128,78,107,224,128,Join,One-sided);EOS
```

Claude-3.5: 1. In CAD programs, screws or screw holes are often represented by circles.

2. Looking at the program, we can see several Circle commands: ...

Answer: 4 (score: 1.0)

Error Category: Lack of Capability

GPT-4o: To determine the number of visible screws or screw holes in the CAD object, we need to identify the instances of circles in the program, as circles typically represent screws or screw holes in CAD drawings ...

Answer: 6 (score: 0.0)

Error Category: Lack of Capability

Gemini-1.5-pro: The code shows six circles that could be screw holes and two "Ext" commands that specify extrusions. Extrusions can be used to create screw holes.

Answer: 8 (score: 0.0)

Correct Answer: 4

# Error Analysis: perception

**Information Extraction: Multimodel QA**

Task Name: Object Interaction Video

The task involves identifying objects a person interacts with in various videos based on the provided questions. You need to output the object name.

**Question:** Which object was taken by the person?

**Keywords:** Contextual Formatted Text, Scene Understanding, Scene and Event Understanding, Videos



GPT-4o: Answer: pillow (score: 1.0)

Error Category: Perception Error  
Claude-3.5: After carefully analyzing the sequence of images, I can see that the person in the video is interacting with a hat. In the earlier frames, the person is holding what appears to be a white or light-colored object, which is likely laundry or bedding. However, as the sequence progresses, we see the person in front of a mirror, and they appear to be adjusting or handling a hat. In the final frames, there's a motion that suggests the person is tossing or throwing the hat upwards.  
Answer: hat (score: 0.0)

Correct Answer: pillow

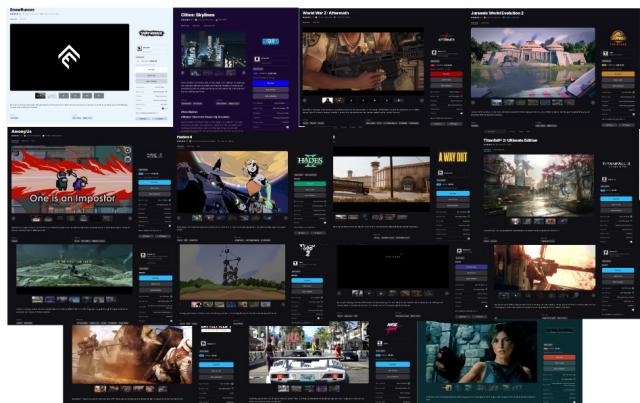
**Information Extraction: Search by Attribute without Calculate**

Task Name: Game Info Retrieval

Given a set of screenshots from a game information page and a specific query, retrieve the names of the games that match the query. If there are multiple matching games, provide a list of game names separated by commas.

**Question:** Find the game available on 'MacOS'

**Keywords:** Object Recognition and Classification, OCR, Language Understanding and Generation, User Interface Screenshots



Claude-3.5 : Let's go through this step-by-step:  
1. We need to look for a game that specifically mentions Mac or MacOS as a platform. ...  
Answer: World of Goo 2, SnowRunner (score: 1.0)

Error Category: Perception Error  
GPT-4o: To find the game available on 'MacOS', I will examine each screenshot for the platform information.  
...  
6. World of Goo 2: No mention of MacOS.  
...  
Answer: SnowRunner(score: 0.0)

Correct Answer: World of Goo 2, SnowRunner

# Error Analysis: knowledge

**Knowledge: World Knowledge**

Task Name: Insect Order Classification

This task involves the classification of insects into their respective taxonomic orders based on visual input. The goal is to accurately identify and categorize an insect from an image into one of the pre-defined orders. Each order is represented by a specific label corresponding to the insect's taxonomic classification, including orders such as Diptera, Hymenoptera, Coleoptera, Hemiptera, Lepidoptera, Psocidea, Thysanoptera, Trichoptera, Orthoptera, Blattodea, Neuroptera, Ephemeroptera, Dermaptera, Archaeognatha, Plecoptera, or Embioptera.

**Keywords:** Object Recognition and Classification, Photographs, Domain-Specific Knowledge and Skills, Contextual Formatted



**Error Category: Lack of Knowledge**  
GPT-4o: 1. The insect has a pair of wings. 2. The antennae are feathery, which is a characteristic feature of some insects. 3. The body structure is relatively small and delicate ...  
Answer: Trichoptera (score: 0.0)

**Error Category: Lack of Knowledge**  
Claude-3.5: 1. The insect is small and appears to be in a circular container, likely for observation purposes. 2. The insect has an elongated body shape ...  
Answer: Lepidoptera (score: 0.0)

**Error Category: Lack of Knowledge**  
Gemini-1.5-pro: The insect has long, slender antennae and its wings are held roof-like over its body. These are characteristic of the order Trichoptera, or caddisflies.  
Answer: Trichoptera (score: 0.0)

**Correct Answer:** Hymenoptera

**Knowledge: World Knowledge**

Task Name: Memorization Chinese Celebrity

Please answer the name and nationality of the celebrity. For the name, please put last name in the front.

**Keywords:** Object Recognition and Classification, Photographs, Language Understanding and Generation, Structured



**Error Category: Lack of Knowledge**  
GPT-4o: I don't know who they are.  
Answer: null (score: 0.0)

**Error Category: Refuse to Answer**  
Claude-3.5: I apologize, but I'm not able to identify or name specific individuals in images. However, I can describe what I see in this photo:...  
Answer: null (score: 0.0)

**Error Category: Lack of Knowledge**  
Gemini-1.5-pro: The person in the image is Eric Chou. He is a Taiwanese singer.  
Answer: {"name": "Chou Chih-chieh", "nationality": "Taiwanese"} (score: 0.0)

**Correct Answer:** {"name": "Lin Junjie", "nationality": "Singaporean"}

# Error Analysis: instruction following

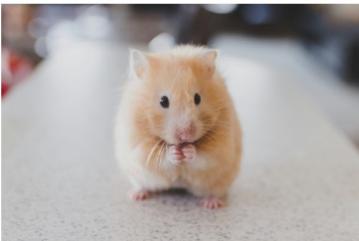
**Knowledge: Arts**

Task Name: Poetry Generation (Acrostic)

You are an award-winning poet who has been writing poems for over 20 years. Please write an acrostic poem about the attached image, using the letters in the specified word or words to start off the first letter of each line in the poem. A space between the words indicates a newline. If there are no spaces, there should not be any blank lines. No title is needed.

Question: Word: hamster

**Keywords:** Object Recognition and Classification, Language Understanding and Generation, Photographs



**GPT-4o: Answer:**  
Huddled in a cozy corner,  
A tiny heart beats with gentle grace.  
Mischief sparkles in those bright eyes,  
Soft fur, a warm embrace.  
Tiny paws clutching a hidden treasure,  
Eagerly nibbling, lost in pleasure.  
Radiating charm, a small wonder. (score: 1.0)

Error Category: Fail to Follow Instruction  
Qwen2-VL-72B: Answer:  
Humble hamster, with fur so fine,  
A creature small, yet full of might.  
Majestic in its tiny form,  
Standing tall, without a care.  
Eyes so bright, a curious gleam,  
Ruling its world, in a tiny dream. (score: 0.0) (acrostic: 0, contain: 1)

Eval Context: acrostic: hamster,  
contain: [hamster, paw, tail, whisker, seed, rodent, scurry, fur]"

**Perception: Multimodal Constrained Captioning**

Task Name: Constrained Generation Contain Length

Can you write a story about the animal in the figure with more than 10 but less than 20 words?

**Keywords:** Object Recognition and Classification, Open-Ended, Language Understanding and Generation, Photographs



**GPT-4o:** Two curious cats, one tabby and one ginger, explored their cozy home, discovering hidden treasures together. (score: 1.0)

**Error Category: Fail to Follow Instruction**  
Claude-3.5: Two feline friends, a tabby and a ginger, lounged side by side, their curious eyes gleaming with mischief and companionship. (score: 0.0) (contain: 0, length1: 1, length2: 0)

**Eval Context:** contain: [cat, kitten], length1: >10, length2: <20



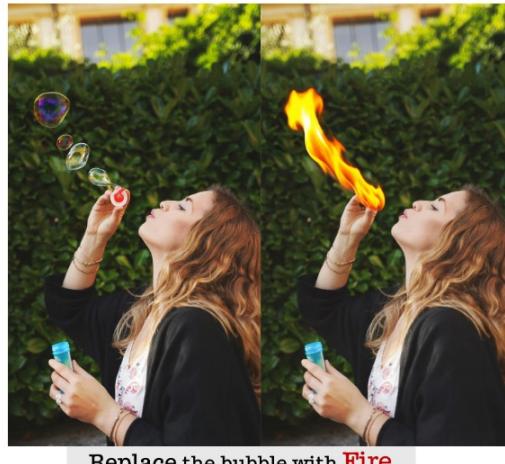
# OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, Wenhui Chen

Presenter: Wenhui Chen

# Image Editing

- Image Editing
- Multi-Skill
- Multi-Resolution
- Instruction



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

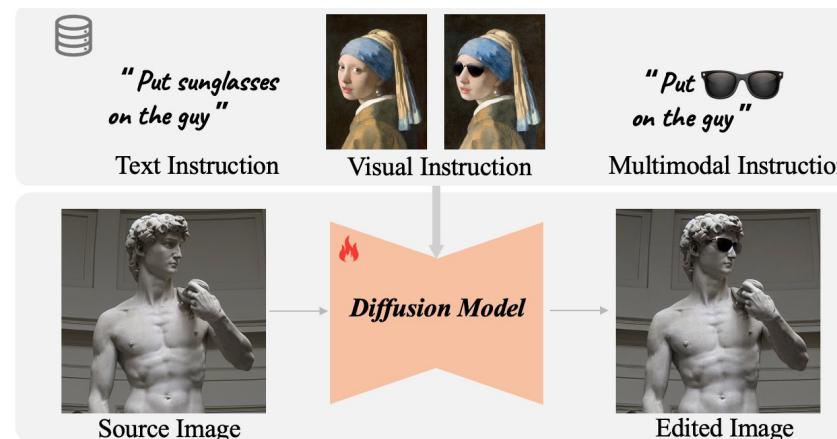
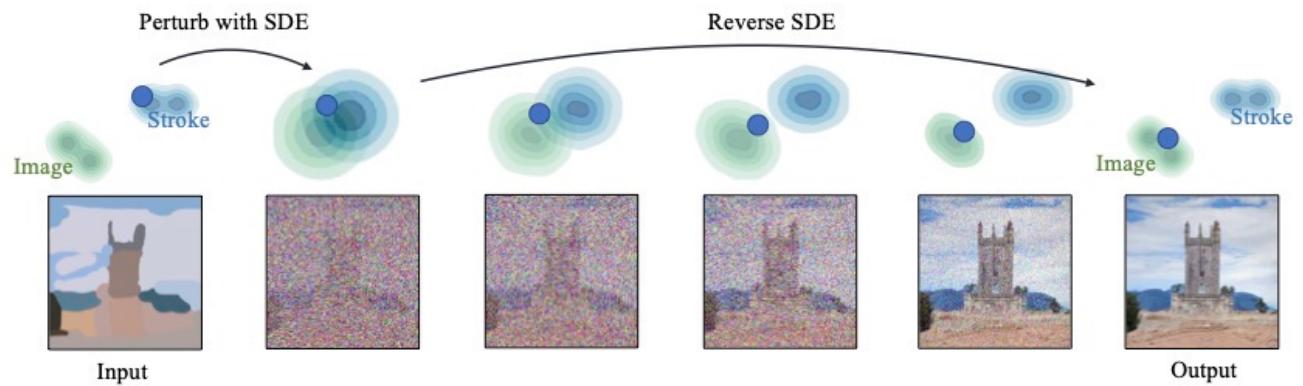


Turn the environment into a snowy landscape



# Background

- Inversion-based methods
  - SDEdit
  - Prompt2Prompt
  - Plug-and-Play
  - Null Text Inversion
- End-to-end methods
  - InstructPix2Pix
  - MagicBrush
  - HQEdit
  - HIVE
  - UltraEdit



# Pro and Cons

- Inversion-based methods
  - Zero-shot without tuning
  - Slow
  - Sensitive to hyper-parameters
- End-to-end methods
  - Requires Tuning
  - Fast
  - Stable w.r.t hyper-parameters

# Issues with end-to-end methods

- Trained entirely with synthetic data, thus sensitive to the synthesis approach and framework.
  - InstructPix2Pix was synthesized with Prompt2Prompt, which can barely handle local editing like addition/removal/swap.
  - MagicBrush was synthesized with DALLE-inpainting, which can barely handle global editing like style/environment.
  - The same issues exist in UltraEdit, CosXL-Edit, HQ-Edit, etc.
- Quality Control is very crude
  - Mostly using CLIP-score, DINO-score ,etc.
  - These scores can only measure the semantic correspondence.

# End-to-end Methods

Preliminary Results on 40 prompts across different editing tasks

Table 1: Comparison of OMNI-EDIT with all the existing end-to-end image editing models.

Property	InstructP2P	MagicBrush	UltraEdit	MGIE	HQEdit	CosXL	OMNI-EDIT
Training Dataset Properties							
Real Image?	✗	✓	✓	✓	✗	✗	✓
Any Res?	✗	✗	✗	✗	✗	✗	✓
High Res?	✗	✗	✗	✗	✓	✗	✓
Fine-grained Image Editing Skills							
Obj-Swap	★★★	★★★	★★★	★★★	★★★	★★★	★★★
Obj-Add	★★★	★★★	★★★	★★★	★★★	★★★	★★★
Obj-Remove	★★★	★★★	★★★	★★★	★★★	★★★	★★★
Attribute	★★★	★★★	★★★	★★★	★★★	★★★	★★★
Back-Swap	★★★	★★★	★★★	★★★	★★★	★★★	★★★
Environment	★★★	★★★	★★★	★★★	★★★	★★★	★★★
Style	★★★	★★★	★★★	★★★	★★★	★★★	★★★

# Motivation

- We need to build more omnipotent image editing models to handle all the editing tasks
  - Balanced Skillset
  - Better Quality Control
  - Support different aspect ratio
  - Support high resolution images.
- Instead of restricting to a fixed data synthesis process, we could learn from data synthesized from multiple specialist.

# Formulation

Assuming that we have an optimal distribution  $p(x'|x, c)$

$$L(\theta) := \sum_{\mathbf{x}, c} D_{KL}(p(\mathbf{x}'|\mathbf{x}, c) \| p_\theta(\mathbf{x}'|\mathbf{x}, c)) = - \sum_{\mathbf{x}, c} \sum_{\mathbf{x}'} p(\mathbf{x}'|\mathbf{x}, c) \log p_\theta(\mathbf{x}'|\mathbf{x}, c) + C \quad (4)$$

where,  $\mathbf{x}'$  is the target image,  $\mathbf{x}$  is the source image and  $c$  is the instruction

# Formulation

Assuming that we have an optimal distribution  $p(x'|x, c)$

$$L(\theta) := \sum_{\mathbf{x}, c} D_{KL}(p(\mathbf{x}'|\mathbf{x}, c) \| p_\theta(\mathbf{x}'|\mathbf{x}, c)) = - \sum_{\mathbf{x}, c} \sum_{\mathbf{x}'} p(\mathbf{x}'|\mathbf{x}, c) \log p_\theta(\mathbf{x}'|\mathbf{x}, c) + C \quad (4)$$

where,  $\mathbf{x}'$  is the target image,  $\mathbf{x}$  is the source image and  $c$  is the instruction

Importance Sampling Approach with a proposal distribution  $q(x'|x, c)$

$$\begin{aligned} L(\theta) &= - \sum_{\mathbf{x}, c} \sum_{\mathbf{x}'} q(\mathbf{x}'|\mathbf{x}, c) \frac{p(\mathbf{x}'|\mathbf{x}, c)}{q(\mathbf{x}'|\mathbf{x}, c)} \log p_\theta(\mathbf{x}'|\mathbf{x}, c) \\ &\approx -\mathbb{E}_{(\mathbf{x}, c) \sim D} [\mathbb{E}_{\mathbf{x}' \sim q(\mathbf{x}'|\mathbf{x}, c)} [\lambda(\mathbf{x}', \mathbf{x}, c) \log p_\theta(\mathbf{x}'|\mathbf{x}, c)]] \\ &\approx -\mathbb{E}_{(\mathbf{x}, c) \sim D} [\mathbb{E}_{\mathbf{x}' \sim q_s(\mathbf{x}'|\mathbf{x}, c)} [\lambda(\mathbf{x}', \mathbf{x}, c) \log p_\theta(\mathbf{x}'|\mathbf{x}, c)]] \end{aligned}$$

$q(x'|x, c)$  can be represented as an ensemble of lots of specialists  $q_s(x'|x, c)$

# Formulation

- We adopt importance sampling function  $\lambda(x', x, c)$

$$\begin{aligned} L(\theta) &= - \sum_{\mathbf{x}, c} \sum_{\mathbf{x}'} q(\mathbf{x}' | \mathbf{x}, c) \frac{p(\mathbf{x}' | \mathbf{x}, c)}{q(\mathbf{x}' | \mathbf{x}, c)} \log p_\theta(\mathbf{x}' | \mathbf{x}, c) \\ &\approx - \mathbb{E}_{(\mathbf{x}, c) \sim D} \left[ \mathbb{E}_{\mathbf{x}' \sim q(\mathbf{x}' | \mathbf{x}, c)} [\lambda(\mathbf{x}', \mathbf{x}, c) \log p_\theta(\mathbf{x}' | \mathbf{x}, c)] \right] \\ &\approx - \mathbb{E}_{(\mathbf{x}, c) \sim D} \left[ \mathbb{E}_{\mathbf{x}' \sim q_s(\mathbf{x}' | \mathbf{x}, c)} [\lambda(\mathbf{x}', \mathbf{x}, c) \log p_\theta(\mathbf{x}' | \mathbf{x}, c)] \right] \end{aligned}$$

- We use large multimodal models (GPT-4o) to approximate  $\lambda$

$$\lambda(\mathbf{x}', \mathbf{x}, c) = \begin{cases} 1, & \text{if LMM(prompt, } \mathbf{x}', \mathbf{x}, c) \geq 9 \\ 0, & \text{otherwise} \end{cases}$$

# Building Swap Specialist

- Mask Generation: Use GroundingDINO and SAM to generate a mask for swapping candidate
- Mask Dilation: Dilate the boundary
- Image Editing: Apply **BrushNet (our own version)** to generate the edited image  $x_{edit}$  by replacing the source image with:

$$\mathbf{x}_{edit} = q_{\text{obj\_replace}}(\mathbf{x}_{\text{src}} \odot (1 - M_{\text{src\_obj}}), M_{\text{src\_obj}}, C_{\text{trg\_obj}})$$

- Rewriting instruction: We apply large multimodal models to rewrite the “replace A with B” with more detailed instruction.

# Building Removal Specialist

- Mask Generation: Use GroundingDINO and SAM to generate a mask for swapping candidate
- Prompt GPT-4o to predict what background should fill in that swapped bounding box
- Image Editing: Apply **BrushNet (our own version)** to generate the edited image  $x_{edit}$  by replacing the source image with:
$$\mathbf{x}_{edit} = q_{\text{obj\_removal}} (\mathbf{x}_{\text{src}} \odot (1 - M_{\text{src\_obj}}), M_{\text{src\_obj}}, C_{\text{trg\_background}})$$
- Rewriting instruction: We apply large multimodal models to rewrite the “Remove A from image” with more detailed instruction.

# Building Attribute Modify Specialist

- We first let the image generation model to generate a source image  $x_{src}$  using  $c_{src}$
- Mask Generation: Use GroundingDINO and SAM to generate a mask for modifying candidate
- Utilizing Prompt2Prompt with mask guidance to generate edited images for attribute modification

$$M_{obj} \odot x_{\text{edited},t} + (1 - M_{obj}) \odot x_{\text{input},t}$$

- Rewriting instruction: We apply large multimodal models to rewrite the “make ...” with more detailed instruction.

# Data Synthesis

- The input images are sampled from LAION-5B and OpenImageV6. We sample images with aspect ratios of 1:1, 2:3, 3:2, 3:4, 4:3, 9:16 and 16:9.
- We utilize these specialists with designated pipelines to generate 200K – 1M pairs as the candidates.
- However, we do observe plenty of artifacts and low-quality pairs.

# Strict Quality Control

- Unlike previous research to use CLIP-Score, we prompt large multimodal models to assign quality score.
- We adopt the VIEScore prompt to assign “semantic consistency”, “visual quality” scores to each pairs with GPT-4o.
- To save cost, we distill GPT-4o rationale to InternVL2-8B with 200K examples and then apply that to filter.

# Data Statistics

<b>Task</b>	<b>Pre-Filtering Number</b>	<b>After-Filtering Number</b>
Object Swap	1,500,000	150,000
Object Removal	1,000,000	100,000
Object Addition	1,000,000	100,000
Background Swap	500,000	50,000
Environment Change	500,000	50,000
Style Transfer	250,000	25,000
Object Property Modification	300,000	30,000
<b>Total</b>	<b>5,050,000</b>	<b>505,000</b>

# EditNet

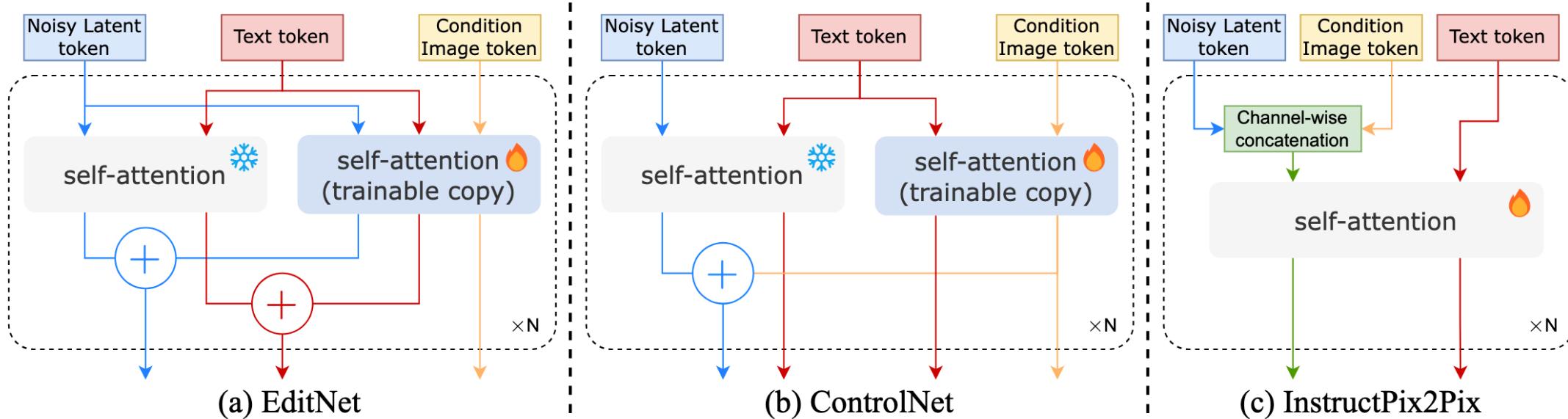


Figure 3: Architecture Comparison between **EditNet(ours)**, ControlNet and InstructPix2Pix for DiT models. Unlike ControlNet's parallel execution, EditNet allows adaptive adjustment of control signals by intermediate representations interaction between the control branch and the original branch. EditNet also updates the text representation, enabling better task understanding.

# Evaluation Set

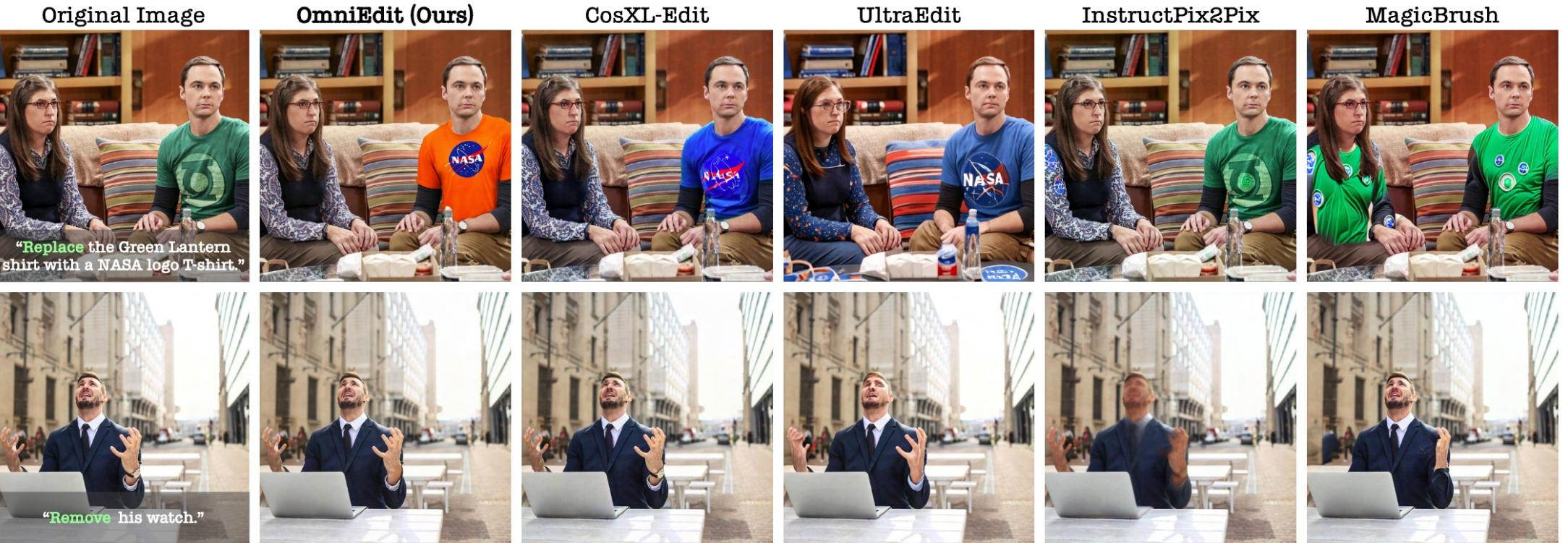
- To create a high-resolution, multi-aspect ratio, multi-task benchmark for instruction-based image editing, we manually collected 62 images from pexels and LAION-5B.
- These images cover a variety of aspect ratios, including 1:1, 2:3, 3:2, 3:4, 4:3, 9:16, and 16:9.
- We write prompts to cover all the mentioned 7 tasks.

# Experimental Results

Table 3: Main evaluation results on Omni-Edit-Bench. In each column, the highest score is bolded, and the second-highest is underlined.

Models	VIEScore (GPT4o)			VIEScore (Gemini)			Human Evaluation			
	$PQ_{avg} \uparrow$	$SC_{avg} \uparrow$	$O_{avg} \uparrow$	$PQ_{avg} \uparrow$	$SC_{avg} \uparrow$	$O_{avg} \uparrow$	$PQ_{avg} \uparrow$	$SC_{avg} \uparrow$	$O_{avg} \uparrow$	$Acc_{avg} \uparrow$
Inversion-based Methods										
DiffEdit	5.88	2.73	2.79	6.09	2.01	2.39	-	-	-	-
SDEdit	6.71	2.18	2.78	6.31	2.06	2.48	-	-	-	-
End-to-End Methods										
InstructPix2Pix	7.05	3.04	3.45	6.46	1.88	2.31	-	-	-	-
MagicBrush	6.11	3.53	3.60	6.36	2.27	2.61	-	-	-	-
UltraEdit(SD-3)	6.44	4.66	4.86	6.49	4.33	4.45	0.72	0.52	0.57	0.20
HQ-Edit	5.42	2.15	2.25	6.18	1.71	1.96	0.80	0.27	0.29	0.10
CosXL-Edit	<u>8.34</u>	<u>5.81</u>	<u>6.00</u>	<u>7.01</u>	<u>4.90</u>	<u>4.81</u>	<u>0.82</u>	<u>0.56</u>	<u>0.59</u>	<u>0.35</u>
HIVE	5.35	3.65	3.57	5.84	2.84	3.05	-	-	-	-
<b>OMNI-EDIT</b>	<b>8.38</b>	<b>6.66</b>	<b>6.98</b>	<b>7.06</b>	<b>5.82</b>	<b>5.78</b>	<b>0.83</b>	<b>0.71</b>	<b>0.69</b>	<b>0.55</b>
$\Delta$ - Best baseline	+0.04	+0.85	+0.98	+0.05	+0.92	+0.97	+0.01	+0.15	+0.10	+0.20

# Visualization



# Visualization

