

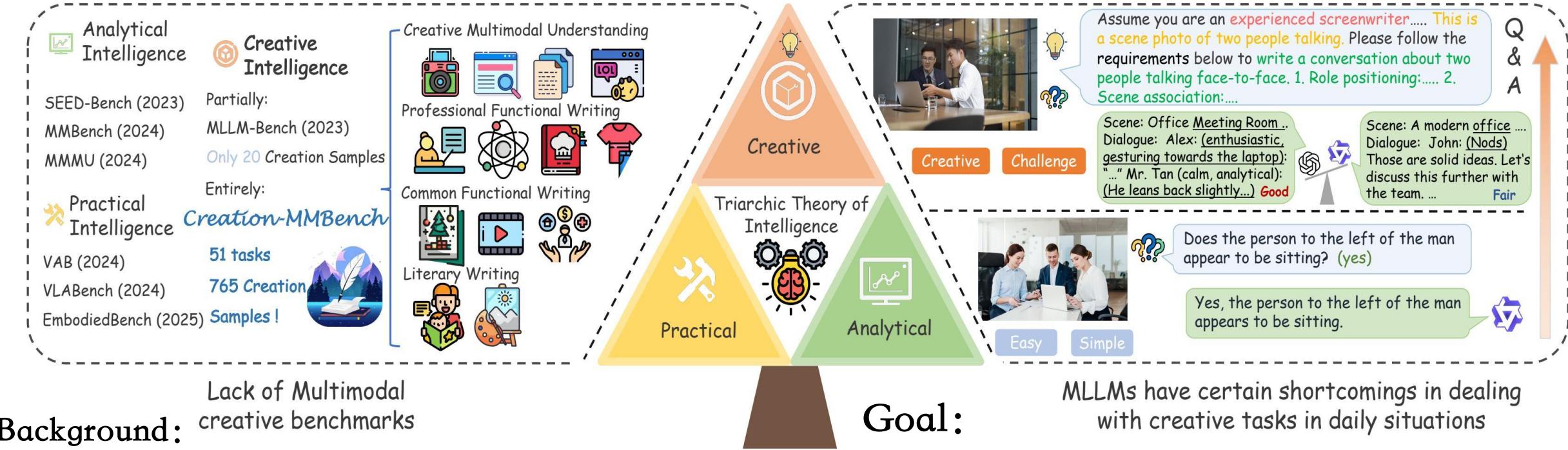
Creation-MMBench: Assessing Context-Aware Creative Intelligence in MLLM

Xinyu Fang

Full Author list: Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, Dahua Lin



Creation-MMBench – Background Introduction



Background:

- Creative intelligence** is the **most difficult and elusive component** of the triarchic theory of intelligence, as it primarily concerns the ability to **generate novel and appropriate solutions/articles across different contexts**.
- Existing evaluation benchmarks** for MLLMs, such as MMBench and MMMU, tend to place greater **emphasis on analytical or practical tasks**, while **overlooking** the “creative tasks” that are frequently encountered in **real-world applications of multimodal AI**.
- Although some benchmarks do incorporate assessments of a model’s creativity, they are generally **small in scale, often restricted to single-image inputs, and situated in overly simple contexts**—even relatively ordinary models can easily provide correct answers.

Goal:

- Construct a multimodal creativity evaluation benchmark that is **closely aligned with real-world scenarios and human preferences**.
- The benchmark aims to assess a model’s ability in **“visual content understanding + contextual adaptation + creative text generation”**, thereby establishing a comprehensive evaluation framework.

Creation-MMBench – Data Annotation & Design



- **Evaluation:** MLLM-as-a-judge, Using GPT-4o generated as reference response, two parts - two rounds:
 - Visual Factuality Score: scoring 1-10, Judge whether the model effectively answers the necessary visual information
 - Reward: Comparing with the reference response, let the Judge Model mark the bias of the response, which is converted into a score

- **Task Design:** Creation-MMBench is designed from top-level tasks. Prototype task sets were first conceived through methods such as brainstorming, then expanded with the assistance of LLMs. These were further refined, summarized, and integrated through human annotation, ultimately resulting in 51 tasks that require image-grounded creative generation in real-world scenarios. The tasks are organized into four major categories.
- **Data Composition:** Each task corresponds to 15 manually annotated examples. Each example contains two main aspects and two categories of evaluation criteria:
 - **Visual Content:** Includes one or more images necessary to complete the use case.
 - **Instructions:** Comprises the **role** (the identity the model is required to assume), **background** (prior knowledge not covered by the visual content and difficult to obtain), **instructions** (the operations the model needs to perform), and **requirements** (constraints or additional considerations).
 - **General Subjective Criteria:** Evaluates the model's expressive ability (structure, style, fluency), its execution of text instructions (alignment with requirements, role, and instructions), and the depth of reflection on the visual content.
 - **Visual Factuality Criteria:** Assesses the model's ability to perceive objective visual content and effectively utilize visual information.

Creation-MMBench – Data Features



Role: an interior space designer, good at finding existing housing problems according to the actual situation of users, and giving relevant opinions.

Query: This picture is the floor plan of the house you have to design. This home is about to welcome a family of four with two children.

Background: This picture is the floor plan of the house you have to design. This home is about to welcome a family of four with two children.

Instruction: analyze the floor plan of the house

Requirements:

1. Combine the characteristics of the occupants and your expertise to list the advantages and disadvantages of the house.
2. Provide suggestions for improving the house's shortcomings along with a preliminary renovation plan.

OCR
Knowledge-based Creation
Spatial Understanding

Role: a renowned writer known for rich imagination and expertise in crafting short stories and micro-novels, capable of drawing inspiration from images for literary creation.

Query: This picture was taken in London in 2015. write a short story or micro-novel based on this image

Background: This picture was taken in London in 2015.

Instruction: write a short story or micro-novel based on this image

Requirements:

1. Develop plot ideas and engage in literary creation centered around the main content of the given image.
2. The setting, background, and character design of the story or novel can be freely crafted.
3. The work must have a clear theme and focus, ...
4. The piece should include an ... title.

Scene Perception
Divergent Thinking
Relation Understanding

Benchmarks	Num of Creative Questions	Criteria Level	multi-images task	Specific Role for each Questions	Visual Factuality Check
VisIT-Bench	65	benchmark	✓	✗	✓
MLLM-Bench	20	instance	✗	✗	✓
Touch-Stone	189	benchmark	✓	✗	✗
AlignMMbench	353	task	✗	✗	✗
Creation-MMBench	765	instance	✓	✓	✓

Table 1. Comparison of Creation-MMBench with other partial-creation MLLM benchmarks.

General Subjective Criteria:

1. Completeness of Analysis: 1.1 Does the response identify ...?
- 1.2 Are the identified advantages and disadvantages reasonable?
2. Feasibility and Relevance of Suggestions: 2.1 Do the suggestions directly address the identified shortcomings of the house? 2.2 Are the suggestions practical, family-friendly, and ...? ...

Visual Factuality Criteria:

1. Consistency with the Floor Plan and Requirements: 1.1 Does the response accurately identify key issues such as insufficient bedrooms, lack of children's spaces, and bathroom accessibility problems? ...

General Subjective Criteria:

1. Completeness and Clarity: 1.1 Does the story create a compelling narrative? 1.2 Is the writing smooth and coherent, allowing the plot to unfold logically and engagingly? 1.3 ...
2. Imagination and Creativity: 2.1 Are the writing style and plot interesting enough to ...? 2.2 ...
3. Clear and Profound Central Theme: 3.1 Does the story maintain a focused and coherent theme? 3.2 ...

Visual Factuality Criteria:

1. Urban Interaction: 1.1 Does the story accurately represent the urban setting of London in 2015? ...? 1.2 Are the central characters in the image developed with realistic emotions and interactions, reflecting their implied roles (i.e. in business attire, ...)? ...? ...

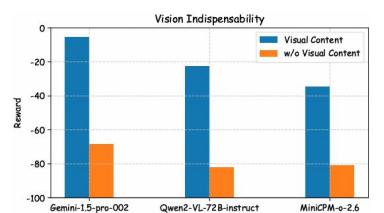
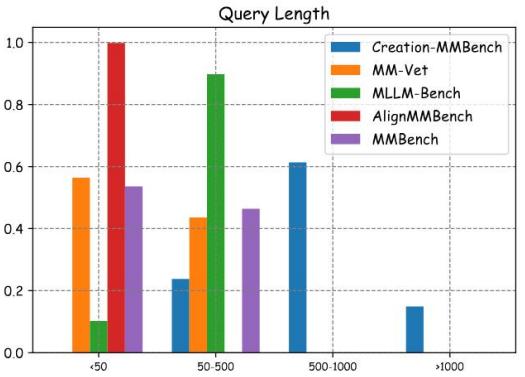
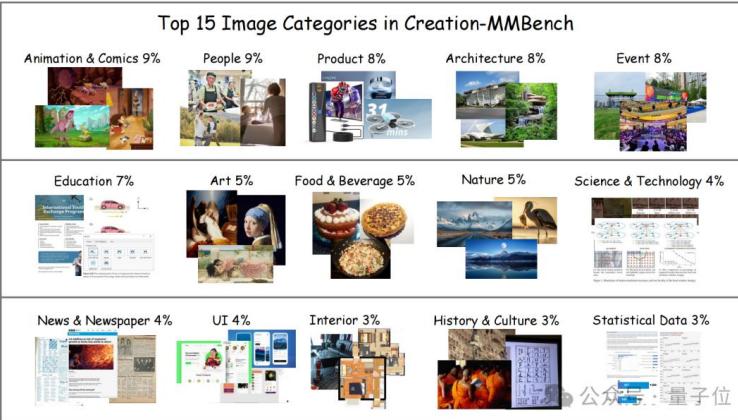


Figure 4. Evaluation Result of MLLMs w/o visual input.

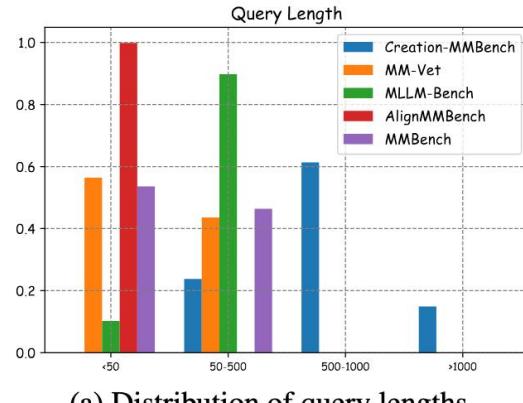


(a) Distribution of query lengths.

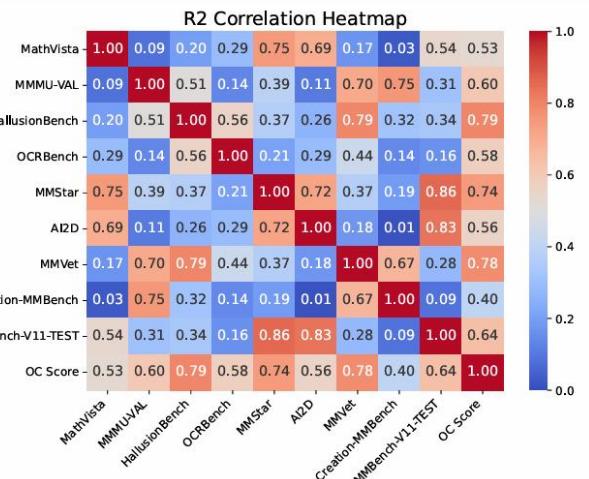
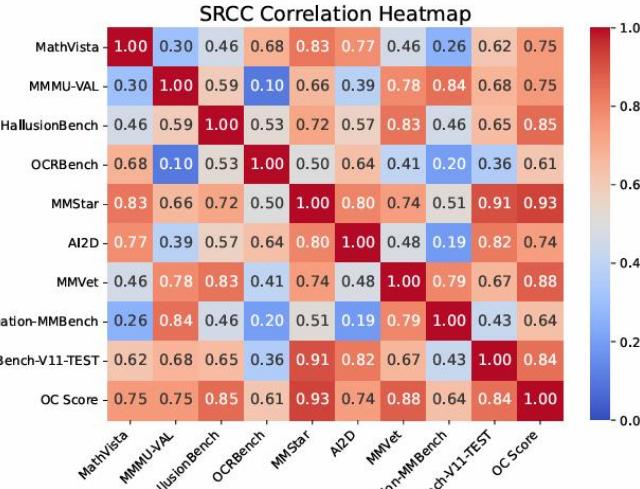
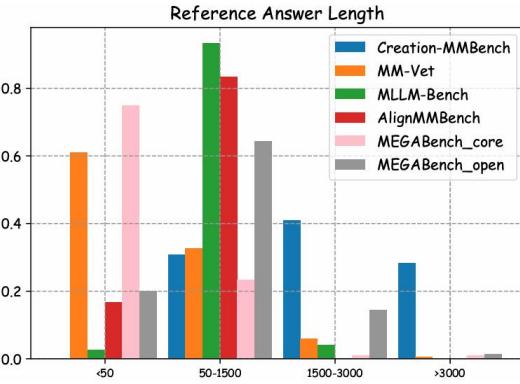


- Four Major Task Categories:** Creation-MMBench consists of 51 tasks, which can be grouped into four primary categories: **literary writing**, **common functional writing**, **professional functional creation**, and **creative multimodal understanding**.
- Thousands of Cross-domain Images:** In terms of visual content, Creation-MMBench spans **nearly 30 categories**, encompassing **thousands of distinct images**. Each task supports up to 9 image inputs, realistically simulating authentic creative environments.
- Complex Real-world Scenarios:** Each instance is annotated based on real images and is paired with four complementary components—**clear role**, **specific background**, **task instructions**, and **additional requirements**—together forming a complete problem.

Creation-MMBench – Further Data Analysis



(a) Distribution of query lengths.



(b) Instructions in Creation-MMBench.

Creation-MMBench – Experiment (Main Results)



Model	Overall		LW		CFW		PFW		CMU		OC Score	Avg Tokens
	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward		
<i>Proprietary MLLMs</i>												
Gemini-2.0-pro-exp	8.53	4.48	8.66	-1.88	8.98	12.71	8.01	3.33	8.65	-8.06	73.4	718
GPT-4o-1120[Baseline]	8.72	0.00	8.86	0.00	8.93	0.00	8.26	0.00	9.38	0.00	72.0	497
Gemini-1.5-pro-002	8.41	-5.49	8.66	-6.04	8.59	-2.04	8.05	-4.82	8.75	-17.22	72.2	444
GPT-4.5-0227	8.54	-5.88	8.63	-4.38	8.76	-8.33	8.05	-5.88	9.29	-0.56	/	394
GPT-4o-mini	8.07	-13.56	8.30	-4.38	8.44	-15.28	7.50	-16.05	8.40	-12.78	64.1	436
Doubao-VL	8.38	-14.09	8.28	-19.17	9.01	-3.33	7.65	-18.72	8.77	-25.00	/	516
Claude-3.5-Sonnet	7.96	-15.46	8.44	-16.46	7.45	-21.57	7.98	-11.14	8.88	-9.44	70.6	336
Moonshot-v1-32k-vision	7.43	-20.58	7.30	-21.46	8.20	-8.80	6.91	-26.50	6.91	-36.11	/	485
<i>Open-Source MLLMs</i>												
Qwen2.5-VL-72B-Instruct	8.33	-5.82	8.04	-10.83	8.91	4.44	7.68	-11.49	8.86	-11.94	76.1	553
InternVL2.5-78B-MPO	8.06	-12.55	8.22	-9.17	8.60	-5.00	7.45	-16.32	8.22	-27.78	77.0	461
InternVL2.5-8B-MPO	7.65	-15.10	8.09	-16.25	8.30	-3.80	6.80	-23.95	7.88	-19.44	70.3	548
InternVL2.5-78B	7.91	-16.43	8.05	-17.50	8.45	-7.69	7.26	-20.53	8.18	-28.33	75.2	473
Qwen2-VL-72B-instruct	7.87	-22.45	7.75	-24.58	8.17	-15.56	7.42	-26.84	8.43	-26.39	74.8	439
InternVL2.5-8B	7.38	-25.42	7.91	-23.33	7.95	-15.83	6.62	-33.95	7.45	-30.00	68.1	500
Qwen2.5-VL-7B-Instruct	7.55	-29.80	7.34	-39.38	8.40	-21.67	6.71	-33.25	7.78	-30.56	70.9	510
MiniCPM-o-2.6	7.49	-34.77	7.79	-35.42	7.95	-27.31	6.76	-40.88	8.08	-36.94	70.2	389
DeepSeek-VL2	7.24	-38.52	7.58	-33.75	7.58	-32.50	6.61	-44.02	7.81	-45.56	66.4	440
LLaVA-OneVision-72B	7.16	-39.87	7.26	-36.32	7.72	-30.61	6.43	-47.98	7.62	-46.37	68.0	315
LLaVA-OneVision-7B	6.75	-43.49	7.36	-43.54	7.27	-31.85	6.04	-50.53	6.82	-56.11	60.2	373
Qwen2-VL-7B-instruct	7.12	-43.76	6.99	-55.83	7.67	-36.30	6.57	-45.26	7.25	-45.28	67.1	456
VITA-1.5	6.43	-53.31	6.77	-46.19	7.23	-46.50	5.70	-57.43	6.22	-69.72	63.3	385

Table 2. **Evaluation Result of MLLMs on Creation-MMBench.** VFS stands for Visual Factuality Score. LW, CFW, PFW, and CMU stand for four categories in Creation-MMBench: Literary Writing, Common Functional Writing, Professional Functional Writing, and Creative Multimodal Understanding. OC Score represents the average score of the OpenVLM Leaderboard and mainly demonstrates the objective performance of the model. The token number is calculated with tiktoken GPT-4o-1120 tokenizer.

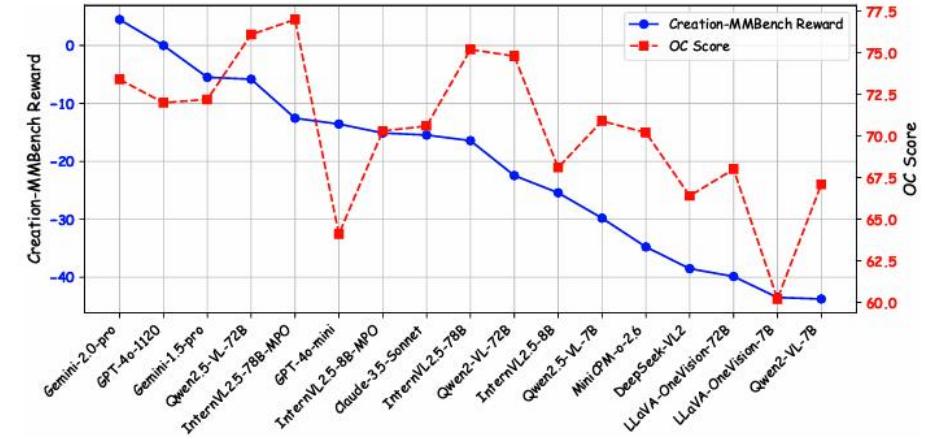
- **Gemini-2.0-Pro** demonstrates outstanding multimodal creative writing abilities. In contrast, **GPT-4.5**, despite being positioned as a model specialized in creative writing, performs overall weaker than **Gemini-Pro** and **GPT-4o**, though it shows notable strengths in multimodal content understanding and creative tasks.
- **Open-source models** such as Qwen2.5-VL-72B and InternVL2.5-78B-MPO also exhibit creative capabilities comparable to those of proprietary models. However, overall, they still lag behind **proprietary MLLMs** to some extent.

Creation-MMBench – Experiment (Main Results)



Model	Overall		LW		CFW		PFW		CMU		OC Score	Avg Tokens
	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward		
<i>Proprietary MLLMs</i>												
Gemini-2.0-pro-exp	8.53	4.48	8.66	-1.88	8.98	12.71	8.01	3.33	8.65	-8.06	73.4	718
GPT-4o-1120[Baseline]	8.72	0.00	8.86	0.00	8.93	0.00	8.26	0.00	9.38	0.00	72.0	497
Gemini-1.5-pro-002	8.41	-5.49	8.66	-6.04	8.59	-2.04	8.05	-4.82	8.75	-17.22	72.2	444
GPT-4.5-0227	8.54	-5.88	8.63	-4.38	8.76	-8.33	8.05	-5.88	9.29	-0.56	/	394
GPT-4o-mini	8.07	-13.56	8.30	-4.38	8.44	-15.28	7.50	-16.05	8.40	-12.78	64.1	436
Doubao-VL	8.38	-14.09	8.28	-19.17	9.01	-3.33	7.65	-18.72	8.77	-25.00	/	516
Claude-3.5-Sonnet	7.96	-15.46	8.44	-16.46	7.45	-21.57	7.98	-11.14	8.88	-9.44	70.6	336
Moonshot-v1-32k-vision	7.43	-20.58	7.30	-21.46	8.20	-8.80	6.91	-26.50	6.91	-36.11	/	485
<i>Open-Source MLLMs</i>												
Qwen2.5-VL-72B-Instruct	8.33	-5.82	8.04	-10.83	8.91	4.44	7.68	-11.49	8.86	-11.94	76.1	553
InternVL2.5-78B-MPO	8.06	-12.55	8.22	-9.17	8.60	-5.00	7.45	-16.32	8.22	-27.78	77.0	461
InternVL2.5-8B-MPO	7.65	-15.10	8.09	-16.25	8.30	-3.80	6.80	-23.95	7.88	-19.44	70.3	548
InternVL2.5-78B	7.91	-16.43	8.05	-17.50	8.45	-7.69	7.26	-20.53	8.18	-28.33	75.2	473
Qwen2-VL-72B-instruct	7.87	-22.45	7.75	-24.58	8.17	-15.56	7.42	-26.84	8.43	-26.39	74.8	439
InternVL2.5-8B	7.38	-25.42	7.91	-23.33	7.95	-15.83	6.62	-33.95	7.45	-30.00	68.1	500
Qwen2.5-VL-7B-Instruct	7.55	-29.80	7.34	-39.38	8.40	-21.67	6.71	-33.25	7.78	-30.56	70.9	510
MiniCPM-o-2.6	7.49	-34.77	7.79	-35.42	7.95	-27.31	6.76	-40.88	8.08	-36.94	70.2	389
DeepSeek-VL2	7.24	-38.52	7.58	-33.75	7.58	-32.50	6.61	-44.02	7.81	-45.56	66.4	440
LLaVA-OneVision-72B	7.16	-39.87	7.26	-36.32	7.72	-30.61	6.43	-47.98	7.62	-46.37	68.0	315
LLaVA-OneVision-7B	6.75	-43.49	7.36	-43.54	7.27	-31.85	6.04	-50.53	6.82	-56.11	60.2	373
Qwen2-VL-7B-instruct	7.12	-43.76	6.99	-55.83	7.67	-36.30	6.57	-45.26	7.25	-45.28	67.1	456
VITA-1.5	6.43	-53.31	6.77	-46.19	7.23	-46.50	5.70	-57.43	6.22	-69.72	63.3	385

Table 2. **Evaluation Result of MLLMs on Creation-MMBench.** VFS stands for Visual Factuality Score. LW, CFW, PFW, and CMU stand for four categories in Creation-MMBench: Literary Writing, Common Functional Writing, Professional Functional Writing, and Creative Multimodal Understanding. OC Score represents the average score of the OpenVLM Leaderboard and mainly demonstrates the objective performance of the model. The token number is calculated with tiktoken GPT-4o-1120 tokenizer.



- Professional functional writing poses greater challenges for models due to its high demand for specialized knowledge and its deeper requirements for understanding visual content. By contrast, common functional writing, being closely tied to daily social life with relatively simple contexts and visual inputs, allows even weaker models to perform reasonably well.
- Although most models achieve relatively high scores in visual factuality within creative multimodal understanding tasks, their ability to creatively re-generate content based on visual inputs still faces significant bottlenecks.

Creation-MMBench – Experiment (Creativity on Text)



VLM	Corresponding LLM	Text Input w. LLM		Text Input w. VLM		Vision+Text Input w. VLM	
		VFS	Reward	VFS	Reward	VFS	Reward
GPT-4o-1120	GPT-4o-1120	8.71	6.96	8.71	6.96	8.72	0.36
Gemini-2.0-pro-exp	Gemini-2.0-pro-exp	8.49	4.08	8.49	4.08	8.53	4.48
Qwen2.5-VL-72B-Instruct	Qwen2.5-72B-Instruct	8.55	0.82	8.51	-4.05	8.33	-5.82
Qwen2.5-VL-7B-Instruct	Qwen2.5-7B-Instruct	8.18	-19.18	7.97	-27.50	7.55	-29.80
MiniCPM-o-2.6	Qwen2.5-7B-Instruct	8.18	-19.18	7.78	-36.57	7.49	-34.77
InternVL2.5-8B	InternLM2.5-7B-Chat	7.83	-22.19	7.92	-28.73	7.38	-25.42

- Current LLM benchmarks for creativity are mostly focused on specific themes (e.g., LiveIdeaBench) and fail to capture their abilities across diverse everyday scenarios. To address this, we constructed **Creation-MMBenchTO**, using GPT-4o to generate image descriptions.
- As shown in the results, **proprietary LLMs generally exhibit slightly better creative performance than open-source ones. Larger LLMs also perform better** in contextual understanding and idea expression. Moreover, overall visual factuality scores improve, since **GPT-4o's image descriptions help LLMs interpret images more effectively compared to MLLMs**.
- After Visual-SFT, VLMs consistently perform worse than their LLM counterparts on Creation-MMBenchTO. This may be due to **instruction tuning with fixed-length prompts, which limits handling of longer texts, leading to lower visual factuality**. In addition, the lack of creative image-text training data further contributes to a significant drop in reward scores.

LLM	Generic		Query-Specific	
	VFS	Reward	VFS	Reward
GPT-4o-1120	8.71	6.96	8.88	3.33
Qwen2.5-72B-Instruct	8.55	0.82	8.82	4.80
InternLM2.5-7B-Chat	7.83	-22.19	8.33	-15.29

Description

Generic Instruction: Please carefully describe the content of each incoming image, starting with the number of images. For each image, first provide a general introduction to the content, then describe the image type, characters and objects, scene and atmosphere, the relationships between people and objects, and any text on the image.

Query-Specific Instruction: Please carefully describe the content of each incoming image, starting with the number of images. For each image, first provide a general introduction to the content, then describe the image type, characters and objects, scene and atmosphere, the relationships between people and objects, and any text on the image. Please pay special attention to the following aspects: <query-specific part>.

Creation-MMBench – Experiment (Judging Strategy)



Judging Method	Judging Model/Human	MLLM	Dual Evaluation			Non-Dual Evaluation		
			MAE↓		Consistency↑	MAE↓		Consistency↑
LLM-as-a-judge	Gemini-2.0-Pro	Gemini-1.5-pro-002	0.67	83.17	84.16	0.75	77.23	
		Qwen2-VL-72B	0.59	84.16	84.16	0.65	78.22	79.21
		MiniCPM-o-2.6	0.61	85.15		0.67	82.18	
	GPT-4o-mini	Gemini-1.5-pro-002	0.67	83.17	86.23	0.79	74.26	
		Qwen2-VL-72B	0.59	85.29	86.23	0.67	76.47	77.38
		MiniCPM-o-2.6	0.52	90.20		0.66	81.37	
	Claude-3.5-Sonnet	Gemini-1.5-pro-002	0.63	89.11	91.80	0.73	78.22	
		Qwen2-VL-72B	0.46	94.12	91.80	0.58	82.35	81.97
		MiniCPM-o-2.6	0.46	92.16		0.58	85.29	
	GPT-4o	Gemini-1.5-pro-002	0.56	93.07	91.48	0.56	90.10	
		Qwen2-VL-72B	0.46	92.16	91.48	0.54	87.25	87.54
		MiniCPM-o-2.6	0.51	89.22		0.58	85.29	
MLLM-as-a-judge	Gemini-2.0-Pro	Gemini-1.5-pro-002	0.65	82.83	86.67	0.78	74.75	
		Qwen2-VL-72B	0.51	91.00	86.67	0.67	80.00	78.67
		MiniCPM-o-2.6	0.61	86.14		0.69	81.19	
	GPT-4o-mini	Gemini-1.5-pro-002	0.64	84.16	89.51	0.71	76.24	
		Qwen2-VL-72B	0.53	93.14	89.51	0.65	82.35	80.33
		MiniCPM-o-2.6	0.49	91.18		0.61	82.35	
	Claude-3.5-Sonnet	Gemini-1.5-pro-002	0.56	89.90	90.60	0.61	83.84	
		Qwen2-VL-72B	0.46	92.00	90.60	0.59	85.00	85.23
		MiniCPM-o-2.6	0.47	89.90		0.57	86.87	
	GPT-4o	Gemini-1.5-pro-002	0.53	92.08	92.13	0.57	89.11	
		Qwen2-VL-72B	0.42	96.08	92.13	0.46	91.18	88.85
		MiniCPM-o-2.6	0.53	88.24		0.59	86.27	
Human-as-a-judge	H1	Gemini-1.5-pro-002	/	/	/	0.65	84.16	
		Qwen2-VL-72B	/	/	/	0.60	90.20	
		MiniCPM-o-2.6	/	/		0.66	87.25	
	H2	Gemini-1.5-pro-002	/	/	/	0.82	74.26	
		Qwen2-VL-72B	/	/	/	0.72	82.35	78.69
		MiniCPM-o-2.6	/	/		0.73	79.41	
	H3	Gemini-1.5-pro-002	/	/	/	0.74	76.24	
		Qwen2-VL-72B	/	/	/	0.62	80.39	82.30
		MiniCPM-o-2.6	/	/		0.72	90.20	
	H4	Gemini-1.5-pro-002	/	/	/	0.64	87.13	
		Qwen2-VL-72B	/	/	/	0.61	89.22	87.87
		MiniCPM-o-2.6	/	/		0.65	87.25	

- 4 Volunteers, 3 models (Gemini-1.5-pro-002, Qwen2-VL-72B, MiniCPM-o-2.6) as model A, 使 reference model (GPT – 4o -1120) as model B , using MAE and Consistency as metrics。
- The results show that the MLLM-as-a-judge approach is optimal (further highlighting the importance of visual content), while Dual-Evaluation plays a key role in improving overall consistency. Among these, GPT-4o emerges as the most reliable judge model.
- Human preferences, however, are not highly consistent with each other, reflecting the inherent subjectivity of our benchmark.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\mathcal{J}_i - \mathcal{P}_i|$$

$$\text{Consistency} = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1, & \text{if } |\mathcal{J}_i - \mathcal{P}_i| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Creation-MMBench – Qualitative Analysis



Question

Nobody: The Australian who threw his cigarette into the grass:



Assume you are a master of meme interpretation with a good sense of humor. Here is an Internet meme picture. Please follow the requirements below to analyze it. 1. First, please describe the content of the given image in as much detail as possible. 2. Second, if ..., please provide an introduction to that knowledge or background.

Assistant A

 The image features a close-up of Gru, The background knowledge is that Australia is prone to bushfires, especially during its hot and dry seasons. The meme humorously suggests that an Australian who carelessly discards a cigarette into the grass would immediately realize the potential consequences of their action ...

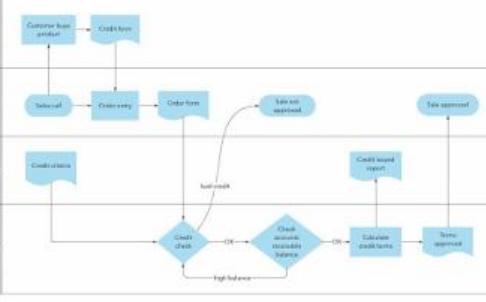
Assistant B

 The character is from the animated film "The Nightmare Before Christmas." The meme references a common issue in Australia, where discarded cigarettes can ignite grass fires, especially during dry seasons. ... The text humorously contrasts the typical reaction of people to the reckless behavior of an Australian who carelessly throws a cigarette into the grass.

Evaluation

Clear and Direct Description	Insufficient overall coherence
Effectively Explain Background Knowledge	Incorrect Identification of Character
Identify Character in correct	Explanation Less Effective
Logic and Coherent Explanation	Clear but Lack engaging Quality

Question



Assume you are an experienced software development engineer,... This is a software engineering diagram that you encounter in your work. Please follow the requirements below to draft a document to ensure that ... can understand the type of diagram, the business process, and the division of work after reading it.

1. Ensure the document is concise and clear.
2. The document should clearly specify the type of diagram, its purpose, and the stage in software engineering where it is typically used.

Assistant A

 1. The diagram shown is a **Swimlane Diagram**, a type of ... It visually represents a business process by dividing it into lanes, where each lane corresponds to a specific role, ... 2. **Purpose of the Diagram:** The purpose of this Swimlane Diagram is to: Illustrate the flow of a business process: It shows how tasks and decisions move between different roles or departments. Clarify responsibilities: Each lane represents a ... 3. **Stage in Software Engineering:** This diagram is typically used during the Requirements Analysis or System Design stage of software engineering. It helps: Understand the current ... 4. **Explanation of the Business Process:** The diagram represents a credit approval process for a customer purchasing a product. Below is a step-by-step explanation: Roles/Departments ... 6. **Key Takeaways:** This Swimlane Diagram provides a clear, visual representation of the credit approval process. By following this document, even a beginner can understand the purpose, structure, and workflow depicted in the diagram.

Assistant B

 1. ... The diagram is a Data Flow Diagram (DFD). 2. **Business Process Illustrated:** This specific DFD illustrates the credit approval process for a sales transaction. Here's a step-by-step breakdown of the process: 1. Customer buys product: The process begins when a customer decides to buy a product. 2. Credit form: The customer fills out a credit form, which is then sent to the sales department ... 3. **Division of Work:** The DFD also shows how the work is divided among different departments:
- **Customer:** Initiates the process by buying a product and filling out a credit form.
- **Sales Department:** Handles the sales call, order entry, and order form creation. ... By following this DFD, each department knows its role and responsibilities in the credit approval process, ensuring a smooth and efficient workflow.

Evaluation : A Much Better Than B

Assistant A correctly identifies the diagram as a **Swimlane Diagram**, which is accurate based on the image. The response is clear and concise, making it accessible to beginners. It effectively explains the purpose of the diagram, highlighting its role in illustrating business processes, clarifying responsibilities, and identifying dependencies. The document also accurately places the diagram in the Requirements Analysis or System Design stage of software engineering, which is appropriate. The explanation of the business process is detailed and aligns well with the image, covering the roles involved and the flow of tasks. The language is professional and structured, suitable for both technical and non-technical readers.

Assistant B Evaluation:\n\nAssistant B incorrectly identifies the diagram as a Data Flow Diagram (DFD), which does not align with the image. This misidentification affects the clarity and accuracy of the response. While the explanation of the business process is detailed, it is based on the incorrect assumption of the diagram type. The response does cover the division of work and provides a step-by-step breakdown, but the initial error in identifying the diagram type undermines the overall quality. The language is clear and professional, but the misidentification impacts the completeness and accuracy of the document.

Q&A

Thanks for listening!

