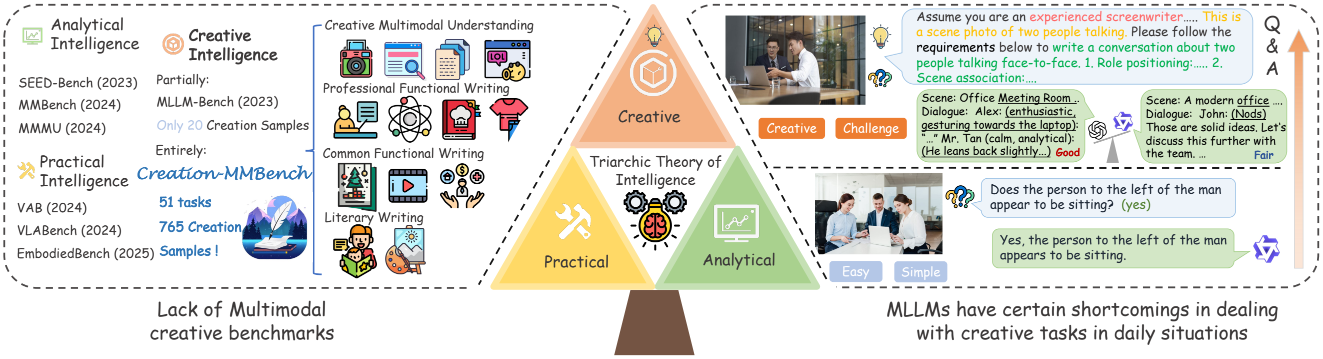
Creation-MMBench: Assessing Context-Aware Creative Intelligence in MLLMs

Xinyu Fang[1,2*], Zhijian Chen[3*], Kai Lan[3], Lixin Ma[3],
Shengyuan Ding[2,4], Yingji Liang[5], Xiangyu Zhao[2,6], Farong Wen[6],
Zicheng Zhang[2,6], Guofeng Zhang[1], Haodong Duan[2†], Kai Chen[2†], Dahua Lin[2,7]
Zhejiang University[1] Shanghai AI Laboratory[2] Tongji University[3] Nanjing University[4]
East China Normal University[5] Shanghai Jiaotong University[6] The Chinese University of Hong Kong[7]



Assume you are an experienced screenwriter..... This is a scene photo of two people talking. Please follow the requirements below to write a conversation about two people talking face-to-face. 1. Role positioning:..... 2. Scene association:....

Scene: Office Meeting Room .. Dialogue: Alex: (enthusiastic, gesturing towards the laptop): "..." Mr. Tan (calm, analytical): (He leans back slightly...)

Scene: A modern office .... Dialogue: John: (Nods) Those are solid ideas. Let's discuss this further with the team. ...

Does the person to the left of the man appear to be sitting? (yes)

Yes, the person to the left of the man appears to be sitting.

Figure 1. **Our Motivation for Creation-MMBench.** The triarchic theory of intelligence divides intelligence into three forms. Current MLLM benchmarks have significant gaps in evaluating visual-creative intelligence compared to the other forms. Additionally, existing benchmarks feature simple questions that fail to assess model performance in real-life creative tasks. Therefore, we proposed Creation-MMBench, which includes four categories, more creative and discriminative questions, and better evaluation of visual creative intelligence.

Abstract

Creativity is a fundamental aspect of intelligence, involving the ability to generate novel and appropriate solutions across diverse contexts. While Large Language Models (LLMs) have been extensively evaluated for their creative capabilities, the assessment of Multimodal Large Language Models (MLLMs) in this domain remains largely unexplored. To address this gap, we introduce Creation-MMBench, a multimodal benchmark specifically designed to evaluate the creative capabilities of MLLMs in real-world, image-based tasks. The benchmark comprises 765 test cases spanning 51 fine-grained tasks. To ensure rigorous evaluation, we define instance-specific evaluation criteria for each test case, guiding the assessment of both general response quality and factual consistency with visual inputs. Experimental results reveal that current open-source MLLMs significantly underperform compared to proprietary models in creative tasks. Furthermore, our analysis demonstrates that visual fine-tuning can negatively impact the base LLM's creative abilities. Creation-MMBench provides valuable insights for advancing MLLM creativity and establishes a foundation for future improvements in multimodal generative intelligence. Full data and evaluation code is released on https://github.com/open-compass/Creation-MMBench.

1. Introduction

Creativity is the ability to generate **novel** and **appropriate** solutions to complex problems across various contexts[1, 17]. With the rapid advancement of Large Language Models (LLMs), numerous benchmarks have been proposed to assess their capabilities across different dimensions of intelligence, including comprehension, reasoning, and creativity [12, 18, 21, 22, 25]. These benchmarks have significantly contributed to a deeper understanding of LLM intelligence and have played a crucial role in driving their improvement. Meanwhile, Multimodal Large Language Models (MLLMs) [2, 4, 14] have also benefited from advancements in LLMs, achieving notable progress in perception, reasoning, and other cognitive abilities [3, 16, 32].

As a well-established theory in psychology, the Triarchic Theory of Intelligence [23] comprises three subtheories: the

[*] Equal Contribution.
[†] Corresponding Author.

analytical subtheory, the contextual subtheory, and the creative subtheory. The analytical subtheory primarily focuses on information processing and problem-solving skills based on domain-specific knowledge and can be assessed through various knowledge and reasoning benchmarks [10, 32]. The contextual subtheory, on the other hand, emphasizes practical intelligence in real-world scenarios and is typically evaluated using agent-based or embodied AI benchmarks [28, 33]. Despite the significance of the creative subtheory in intelligence, evaluations of MLLMs' creative capabilities remain highly inadequate and lag significantly behind those conducted for LLMs [8, 18]. Moreover, constructing benchmarks to assess visual creativity presents inherent challenges. Cognitive science research suggests that creativity arises from a distributed cortical network involving the coordination of multiple brain regions. As illustrated in Fig. 2, creativity is closely associated with functions of the frontal lobe, such as concentration, planning, and problem-solving [11]. Within the context of MLLM evaluation, assessing creative capabilities requires benchmarks that encompass a broader range of fundamental cognitive abilities compared to those needed for other types of intelligence assessment [15, 31].

To address this significant gap, we introduce **Creation-MMBench**, a novel benchmark specifically designed to assess the creative capabilities of MLLMs in image-based tasks across authentic real-world scenarios. The benchmark consists of 765 test cases spanning 51 fine-grained tasks, which are categorized into four major groups: **Literary Writing**, **Common Functional Writing**, **Professional Functional Writing**, and **Creative Multimodal Understanding**. Additionally, the benchmark is accompanied by rich context to facilitate comprehensive evaluation. In each task, an MLLM is provided with one or more images along with a detailed context specifying the assigned role, necessary background information, and clear task instructions. The model then follow the instruction and leverage the visual input to accomplish various creative tasks, such as composing artwork-inspired prose, developing structured lesson plans, or interpreting the conceptual foundations of advertisements. The approach enables a systematic assessment of MLLMs' capacity to integrate visual perception with creative expression in contextually appropriate ways.

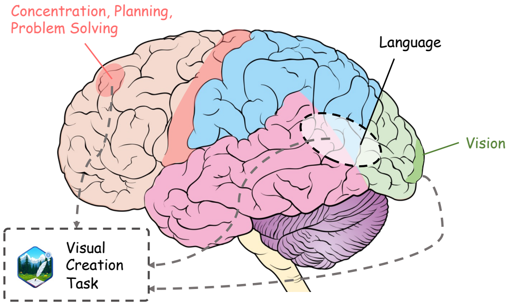


Figure 2. **Brain regions related to creativity and their respective functions [6, 11].**

Unlike ground-truth based evaluations, creative responses generated by models resist rule-based assessment methods. In our evaluation framework, we implement the widely adopted MLLM-as-a-Judge methodology, utilizing GPT-4o to assess the quality of model-generated responses. Given the diverse task types and stylistic variations across Creation-MMBench, a single-criterion evaluation model cannot reliably assess all tasks. To this end, we define instance-level evaluation criteria for each test case, ensuring that responses are assessed based on their ability to integrate contextual and visual information effectively. Using these tailored criteria, an MLLM-generated response is compared against a reference answer, and preferences are assigned accordingly. In addition to the preference obtained through pairwise comparison, we introduce a visual factuality score to evaluate whether the MLLM's response aligns with key facts present in the visual input. This factual score is determined through unitary evaluation conducted by the GPT-4o judge model. Both Unitary Scoring and Pairwise Comparison offer a comprehensive assessment of creative quality and factual accuracy.

Based on Creation-MMBench, we conduct a comprehensive evaluation of mainstream MLLMs. The results indicate that current open-source MLLMs generally underperform compared to advanced proprietary models (*e.g.*, Gemini-2.0-Pro, GPT-4o) in terms of context-aware creativity. To further explore the impact of visual instruction tuning, we transformed Creation-MMBench into a text-only variant, **Creation-MMBench-TO**, by replacing image inputs with corresponding textual descriptions. The results reveal a negative effect of visual fine-tuning on the creative abilities of the base LLM, suggesting potential trade-offs introduced by multimodal adaptation.

In summary, our main contributions are three-fold:

- Development of Creation-MMBench, a multimodal benchmark specifically designed to evaluate the creative capabilities of MLLMs. The benchmark incorporates a diverse set of image sources, spans a wide range of topics and task types across real-world scenarios, and features high-quality, original human-written instructions.
- Design of a robust evaluation methodology that includes carefully crafted instance-specific criteria for each test case, enabling assessment of both general response quality and visual-factual alignment in model-generated content.
- A comprehensive assessment of various MLLMs on Creation-MMBench, providing detailed insights into their performance. The results highlight the current limitations of MLLMs in context-aware creativity and vision-based language generation, offering valuable guidance for future research and development.

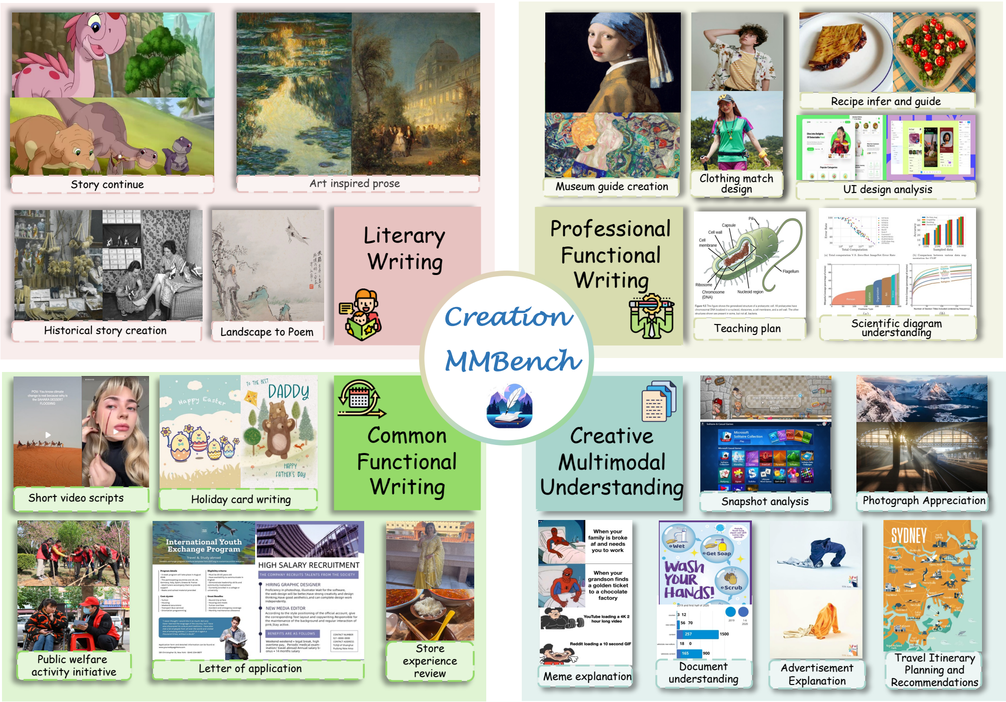


Figure 3. **Overview of Creation-MMBench.** Contains four task categories, each category consists of multiple tasks, and the types of images are diverse. Only a few representative tasks of each category are shown here. Complete list of tasks is detailed in the Appendix A.

2. Related Work

Evaluating Creative Capabilities of LLMs. To evaluate the creative writing capabilities of large language models (LLMs), several benchmark tests have been introduced. One example is the LLM Creative Story-Writing Benchmark [18], where 26 LLMs generate 500 short stories each, incorporating random elements, for a total of 13,000 stories. Six models then assess these stories based on 16 criteria related to character development, plot, and narrative structure. Another test [26] challenges models and humans to create stories based on specific prompts. These benchmarks assess not only the writing quality but also the diversity and complexity of the generated content.

In addition to creative writing tasks, psychological tests commonly used to assess human creativity have also been adapted for evaluating LLMs. The Alternative Uses Test (AUT) evaluates a model's ability to propose novel uses for everyday items within a time limit, as demonstrated in the assessment of GPT-3's creativity [24]. Another benchmark introduces a small-scale test with a leaderboard to evaluate how four LLMs generate alternative uses for objects [20]. The Torrance Tests of Creative Thinking (TTCT) have also been applied to LLMs to assess fluency, flexibility, originality, and elaboration in creative tasks [9].

Brainstorming techniques, commonly used to boost creativity, have been applied to evaluate LLMs' creative abilities. RPGBench [30] uses role-playing games to assess creativity, and LiveIdeaBench [22] evaluates scientific creativity using single-keyword prompts, focusing on novelty, feasibility, fluency, and flexibility. Other benchmarks like LLM-Evolve [29] test problem-solving and adaptability, while SimulBench [13] evaluates creative simulations like acting as a Linux terminal. These benchmarks offer a comprehensive evaluation of LLMs' creative and simulation capabilities, inspiring further exploration of MLLMs' creative potential.

Advancing the Evaluation of Creative Intelligence in MLLMs. The advancement of MLLMs has led to the development of various benchmarks to evaluate their intelligence. MMBench [15] covers 20 distinct ability dimensions, focusing on MLLMs' general capability. MMMU [32] evaluates advanced perception and reasoning with domain-specific knowledge, featuring 11,500 multimodal questions across 6 disciplines. These benchmarks mainly focus on the analytical intelligence of MLLMs. For assessing MLLMs' contextual intelligence, agent-based or embodied AI benchmarks are commonly used. VLABench [33] provides 100 categories of tasks to evaluate robotics' language-conditioned manipulation ability, while EmbodiedBench [28] offers a comprehensive evaluation on models' problem-solving ability with 1,128 tasks across 4

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.**

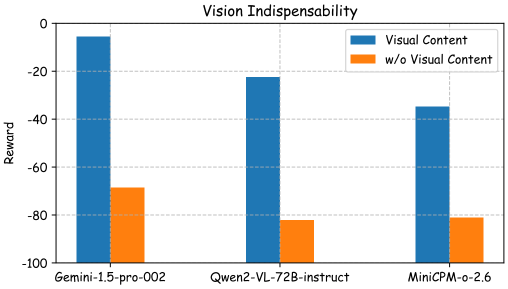


Figure 4. **Evaluation Result of MLLMs w/o visual input.**

environments.

While the evaluation of MLLMs' analytical and contextual intelligence has become relatively mature, the assessment of their creative intelligence remains insufficient. Existing partial-creation benchmarks, such as MLLM-Bench [7] and AlignMMBench [27], lack a systematic and comprehensive evaluation, often failing to assess models' capabilities in complex, real-world scenarios. Furthermore, a dedicated benchmark designed specifically to evaluate MLLMs' creativity has yet to be developed. Therefore, there is a pressing need for a comprehensive and practical benchmark to bridge this gap. Creation-MMBench aims to establish a dedicated benchmark for creative ability evaluation by incorporating a diverse set of real-world tasks, offering a novel perspective on evaluating MLLMs' creative intelligence.

3. Creation-MMBench

This section describes the construction process of Creation-MMBench, covering aspects such as task design, data collection, annotation, quality control, and evaluation. As shown in Fig. 3, the dataset includes diverse categories, reflecting the complexity and breadth of the tasks involved. Additionally, we introduce the data format and the indicators used to assess model capabilities.

3.1. Benchmark construction

Task Design. We began with a brainstorming session to explore creative tasks in daily scenarios and designed a prototype task set encompassing both routine (e.g., writing common emails) and professional tasks (e.g., designing teaching plans). Leveraging a large language model, we then expanded this set to generate a diverse range of candidate tasks. Finally, through manual refinement and integration, a well-defined set of 51 tasks was established.

Task Categorization. We divided the 51 tasks into four main categories:

1. Literary Writing: Focus on literary creation (poetry, dialogues, stories, etc.)
2. Common Functional Writing: Focus on functional writing in daily life (social media writing, daily affairs inquiry, etc.)
3. Professional Functional Writing: Focus on functional writing and creative problem-solving in professional domains (analyzing design, developing lesson plans, etc.)
4. Creative Multimodal Understanding: Focus on the integration of visual understanding and creativity (formatted visual content analysis, image appreciation, etc.)

Data Composition. For each task, 15 carefully crafted test cases are collected. Each test case comprises two major components:

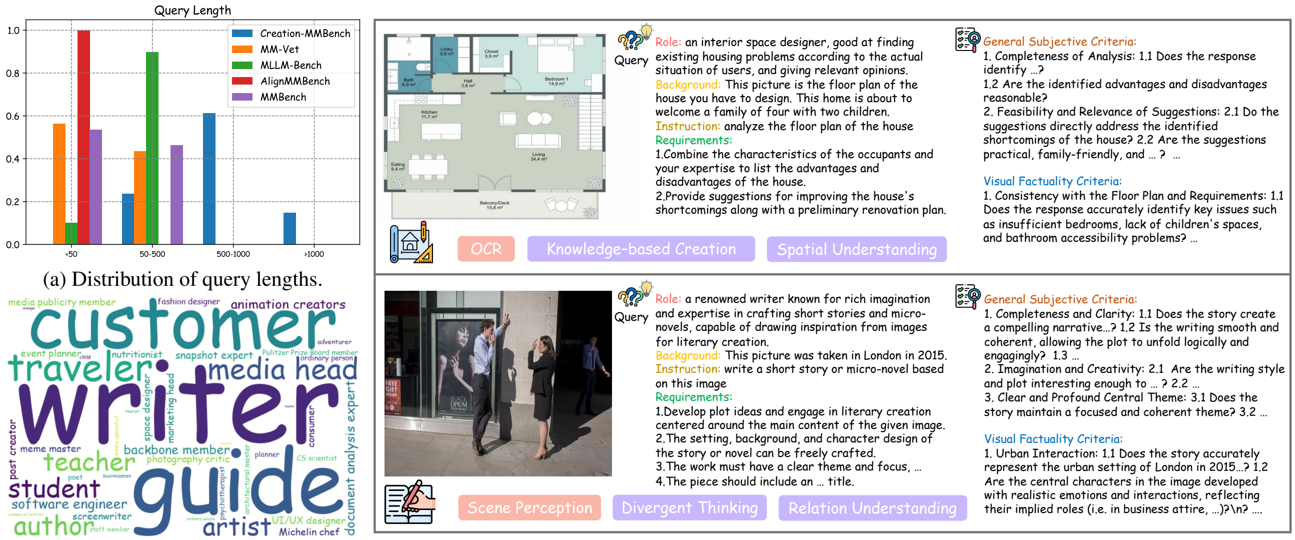
- **Visual Content**: One or more images that contain the necessary information required to accomplish the test case.
- **Query**: Include Role (the identity models need to play), Background (prior knowledge that is not duplicated by the visual content and is difficult to acquire, Instruction (operations that models need to perform), and Requirement (constraints or additional considerations).

All queries are organized into a complete format using a unified template and sent to MLLMs with visual content. Instance-specified criteria are defined to make the evaluation more reasonable. The criteria can be mainly divided into two groups:

- **General Subjective Criteria**: Assess models' expressive capability (structure, style, fluency), execution ability for queries (compliance with requirements, roles, and instructions), and deep reflection on visual content.
- **Visual Factuality Criteria**: Assess models' ability to perceive objective visual content and utilize visual information effectively.

Data Annotation and Quality Control. After task design and definition of data composition, we proceeded with data annotation (including questions and criteria) and quality control. To make the annotator easier to understand, we first built an example question for each task with detailed annotation, then asked volunteers to annotate 15 sample questions for each task with the example and guideline provided below:

1. The **visual content** of questions should be semantic rich, and the query should not contain any explicit information in the visual content.



(a) Distribution of query lengths.

(b) Roles in Creation-MMBench.

Query

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

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? ...

Query

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.
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.

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, ...)?\n? ....

(c) Example Case of Creation-MMBench.

Figure 5. **Statistics and Cases of Creation-MMBench.** Compared to other widely used MLLM benchmarks, Creation-MMBench features a more comprehensive query design to capture abundant creative contexts. Diverse roles are introduced into the queries to stimulate MLLMs' utilization of disciplinary and prior knowledge. As an MLLM benchmark, Creation-MMBench includes a rich variety of images to thoroughly evaluate multiple capabilities of MLLMs.

2. You are encouraged to formulate **diverse queries within the task scope**, like diverse roles and background settings, matching the visual content.
3. The ideal answer should be **open-ended, creative**, but the quality of the response can be assessed using criteria.
4. Ensure each **requirement is clear and avoids redundancy**. Keep the Visual Factuality Criteria concise and direct.

After initial labeling, we conducted cross-verification among volunteers, followed by expert review to ensure data quality.

Evaluation Strategy. We employ the MLLM-as-a-judge approach, which consists of two forms: Unitary Scoring and Pairwise Comparison. In Unitary Scoring, the judging model assigns a score between 1 and 10 to the response of the evaluated model based on the Visual Factuality Criteria. The **Visual Factuality Score** is the average score across all questions. In Pairwise Comparison, the evaluated model is designated as model A, while the baseline model (GPT-4o-1120) is designated as model B. The judging model assesses the responses based on General Subjective Criteria and visual content, selecting from the set {A>>B, A>B, A=B, A<B, A<<B}. To facilitate further computation, we assign numerical values to the pairwise comparison results: {A>>B = +2, A>B = +1, A=B = 0, A<B = -1, A<<B = -2}. For better interpretability, we multiply this average score by 50 and normalize it to the range of -100 to +100, forming a metric as **Reward**. To mitigate the inherent position bias in the MLLM-as-a-judge approach, we conduct a Dual Evaluation, swapping the response positions. The final result is obtained by averaging the outcomes of both evaluations. Detailed evaluation prompt is shown in Appendix B.

3.2. Dataset Statistics

To better understand the composition of Creation-MMBench, we conducted a statistical analysis.

Benchmark Comparison Tab. 1 shows the comparison result of Creation-MMBench and four widely used partial-creation MLLM benchmarks. As a dedicated benchmark for evaluating creativity, Creation-MMBench features a significantly richer set of creative questions and adopts a multi-image format. Each question is designed with specific roles to stimulate MLLMs' creative capabilities. Unlike other benchmarks that apply the same evaluation criteria across an entire benchmark or task, Creation-MMBench customizes assessment criteria for each question, taking into account both subjective creativity and visual factuality. This tailored approach enables a more comprehensive evaluation of MLLMs' creative abilities.

Statistics and Cases Fig. 5 presents several statistics and cases of Creation-MMBench. As depicted in Fig. 5a, we analyzed the query length distributions of Creation-MMBench in comparison with two partial-creation benchmarks (MLLM-Bench, AlignMMBench) and two widely used general benchmarks (MM-Vet, MMBench). The results indicate that our benchmark features more comprehensive and complex query designs. The majority of queries

Model	Overall		LW		CFW		PFW		CMU		OC Score	Avg Tokens
	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward		
Proprietary MLLMs												
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
Open-Source MLLMs												
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.

exceed a length of 500 tokens, which facilitates models in capturing richer creative contexts. Fig. 5b illustrates the diversity of roles present in the queries (e.g., writer, artist, Michelin chef, etc.), reflecting the richness of the questions. As an MLLM benchmark, our dataset contains a total of 1,001 images spanning more than 25 different categories, with some questions incorporating up to 9 images. Fig. 5c displays the example cases in Creation-MMBench.

Vision Indispensability To verify the necessity of visual content in Creation-MMBench, we selected three MLLMs with varying capability levels (Gemini-1.5-Pro-002, Qwen2-VL-72B-instruct, and MiniCPM-o-2.6) and examined their performance after removing visual input. In Fig. 4, we observe that when the visual information is removed, the same models exhibit significant declines in Reward. This finding verifies the necessity of visual content in evaluating model performance.

4. Experiment

Using Creation-MMBench, we evaluate various Multimodal Large Language Models (MLLMs), with a focus on image-based MLLMs that support multiple image inputs. Additionally, we adapted our benchmark into a text-only version (Creation-MMBench-TO) by replacing the visual inputs with corresponding textual descriptions and tested multiple Large Language Models (LLMs) to gain deeper insights into their creative capabilities. All evaluations were conducted based on VLMEvalKit [5], employing greedy decoding during inference with the maximum output tokens set to 4096.

4.1. Main Results

We evaluated 20 current powerful MLLMs on Creation-MMBench, results are shown on Tab. 2.

Proprietary MLLMs. Gemini-2.0-Pro performs similarly to GPT-4o, particularly in common functional writing, where it excels in producing content with a conversational tone and effectively integrates images. Its strong pre-existing knowledge also helps in professional functional writing tasks, but there is a slight gap in perception, especially in tasks like document and snapshot analysis. The smaller GPT-4o-mini outperforms proprietary models like Claude but struggles with professional functional writing due to its limited disciplinary knowledge. DoubaoVL stands out in common functional writing tasks, achieving the highest visual factuality score in this area.

Open-Source MLLMs. Among open-source MLLMs, Qwen2.5-VL-72B stands out, performing similarly to advanced proprietary models like Gemini-1.5-Pro and outperforming GPT-4o-mini across all four major categories. This

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

Table 3. **LLM performance on Creation-MMBench-TO and Visual Instruction Tuning Impact on VLM creation capability.** The image descriptions provided by GPT-4o are general. For the proprietary models, we point to themselves as corresponding LLM and report the performance with image descriptions and questions.

highlights the potential of open-source models in visual creation. The InternVL series also shows strong performance across different model sizes, indicating potential advantages in data and training strategies. The mixed preference optimized (MPO) model demonstrates impressive results in smaller models, with particular strengths in creative multimodal understanding, suggesting that MPO can effectively guide models to better align with human preferences.

Category-level Evaluation Results. Across all four categories, professional functional writing shows relatively weaker performance, while common functional writing performs the best. This may be due to the greater difficulty of tasks in the former, which require extensive disciplinary knowledge and a deeper understanding of image content. These tasks are more complex and demand higher cognitive abilities. In contrast, common functional writing typically involves simpler, everyday tasks that require less advanced image understanding, making them easier to complete. In the Multimodal Content Understanding and Creation category, while all models show basic content understanding, their ability to generate more creative content is limited. This highlights the gap between the models' objective interpretation abilities and their human-aligned visual creativity, further qualitative cases are provided in Appendix G.

Comparison of Model Performance on Objective Tasks and Creation-MMBench. To better compare the models' objective performance with their visual creativity, we use the OC Score to represent the overall objective performance. As shown in Fig. 6, proprietary models perform well both in objective tasks and visual creativity. However, some open-source models, despite showing strong objective performance, struggle with open-ended visual creativity tasks. These models tend to excel in tasks with definitive answers but fall short in generating creative, contextually relevant content. This discrepancy emphasizes the need for a more comprehensive evaluation approach, as traditional objective metrics alone may not fully capture a model's creative abilities in complex, real-world scenarios.

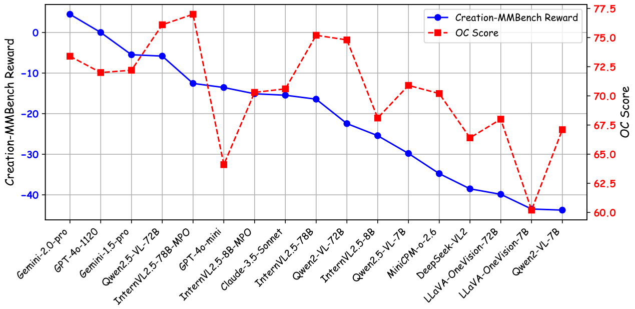


Figure 6. **Comparing OC Score and Creation-MMBench Reward.** This figure shows the model performance on the OpenVLM Leaderboard and Creation-MMBench, highlighting a significant gap between objective performance and visual creativity in some open-source models.

4.2. Evaluating LLMs on Creation-MMBench-TO

Current creation benchmarks for Large Language Models mostly focus on specific topics (e.g., LiveIdeaBench [22]), but fail to reveal their creation capability in multiple daily scenarios. To investigate it, we build **Creation-MMBench-TO** and GPT-4o was used to make the image descriptions with the prompt shown in Appendix E. As shown in Tab. 3, proprietary LLMs showed slightly better contextual creativity than open-source LLMs, though the gap was smaller than that between MLLMs. Large-scale language models performed better at understanding context and expressing ideas compared to smaller models. Additionally, the visual factuality score improved because GPT-4o's image descriptions helped LLMs better interpret the image in comparison to MLLMs. Surprisingly, GPT-4o performed better in visual creativity on Creation-MMBench-TO. This could be because the model can focus more on divergent thinking and creation with the help of descriptions, which may minimize the negative impact of the basic visual content on creativity.

4.3. Impact of Visual instruction tuning on creation capability of MLLM

Existing research indicates that visual instruction tuning procedures may adversely affect the language encoder's ca-

Judger	MLLM	Dual Eval				Single Eval			
		MAE↓		Cons.↑		MAE↓		Cons.↑	
Gemini-2P	Gemini	0.65	0.59	82.83	86.67	0.78	0.72	74.75	78.67
	Qwen	0.51		91.00		0.67		80.00	
	MiniCPM	0.61		86.14		0.69		81.19	
Claude-3.5	Gemini	0.56	0.50	89.90	90.60	0.61	0.59	83.84	85.23
	Qwen	0.46		92.00		0.59		85.00	
	MiniCPM	0.47		89.90		0.57		86.87	
GPT-4o	Gemini	0.53	**0.50**	92.08	**92.13**	0.57	**0.54**	89.11	**88.85**
	Qwen	0.42		96.08		0.46		91.18	
	MiniCPM	0.53		88.24		0.59		86.27	

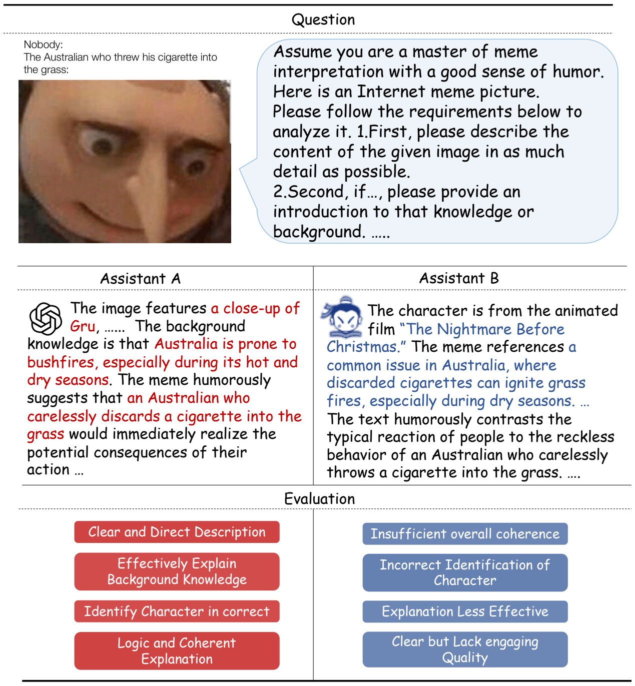
Table 4. **The Alignment Between Different Evaluation Strategies and Human Preference.**

pacity to process and model text-only inputs. To further investigate this, we conducted three experiments under different settings, as shown in Tab. 3. The results indicate that the open-source MLLM, after visual instruction tuning, consistently performs worse compared to the corresponding LLM on Creation-MMBench-TO. This could be due to the instructions used during tuning being of similar length, which restricts the model's ability to grasp detailed content in longer texts, resulting in a lower visual factuality score. The lack of creative data that combines images further contributes to a significant drop in the reward score. Although some proprietary models have shown stronger performance on Creation-MMBench, the performance gap of most MLLMs on Creation-MMBench-TO and Creation-MMBench highlights the need for improvement in the perceptual capabilities of MLLMs.

4.4. Evaluation Strategy Selection

The goal of MLLM-as-a-judge is always to achieve a higher alignment with human preferences. Therefore, we randomly sampled a subset of questions (51 tasks × 2 questions) and recruited four volunteers to do the pairwise comparison. We selected three models (Gemini-1.5-pro-002, Qwen2-VL-72B, MiniCPM-o-2.6) as Model A, used the baseline model (GPT-4o-1120) as Model B, randomizing the responses' position to avoid human biases. Details of the human evaluation process are provided in Appendix F.

We then selected three advanced MLLMs (Gemini-2.0-Pro, Claude-3.5-Sonnet, GPT-4o) as judging models, and used MAE and Consistency as metrics to reflect the alignment degree. Tab. 4 presents the alignment degree between different evaluation strategies and human preferences. The results indicate that for all judging models, Dual Evaluation outperforms Single Evaluation, verifying the necessity of Dual Evaluation. Among all the judging models, GPT-4o achieves the best performance in terms of MAE and Consistency, exhibiting the highest alignment with human preferences. Finally, we selected Dual Evaluation, and GPT-4o as the evaluation strategy for Creation-MMBench.



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. .....

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 …

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

Figure 7. **Qualitative study Case between InternVL-2.5-78B and Reference Answer (GPT4o-1120).**

4.5. Qualitative Study

To further explore the differences between models on Creation-MMBench, we conducted a detailed qualitative study by combining model responses with evaluations. As shown in Fig. 7, InternVL2.5 exhibited limitations in visual perception, particularly in accurately identifying characters due to insufficient latent knowledge. Additionally, InternVL2.5 showed certain weaknesses in the fluency and engagement of its language expression. In contrast, GPT-4o was favored by the evaluation model, which provided a more balanced assessment. This highlights that open-source models still have considerable space for improvement, particularly in visual creativity tasks.

5. Conclusion

We present Creation-MMBench, a novel benchmark designed to assess the creative capabilities of MLLMs in real-world scenarios. The benchmark consists of 765 cases across 51 detailed tasks. For each case, we develop instance-specific criteria to evaluate both the subjective quality of responses and visual-factual alignment. Additionally, we create a text-only version, Creation-MMBench-TO, by substituting image inputs with corresponding textual descriptions. Extensive experiments on both benchmarks enable a thorough assessment of mainstream MLLMs' creative abilities and allow us to examine the negative impact of visual instruction tuning.

Acknowledgement

This project is supported by the National Key R&D Program of China (No.2022ZD0161600), the Shanghai Postdoctoral Excellence Program (No.2023023), and China Postdoctoral Science Fund (No.2024M751559, No.2025T180412), and Shanghai Artificial Intelligence Laboratory.

References

[1] Teresa M Amabile. *Creativity in context: Update to the social psychology of creativity*. Routledge, 2018. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1

[5] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 6

[6] Zhenni Gao, Xiaojin Liu, Delong Zhang, Ming Liu, and Ning Hao. Subcortical structures and visual divergent thinking: a resting-state functional mri analysis. *Brain Structure and Function*, 226(8):2617–2627, 2021. 2

[7] Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, et al. Mllm-bench: evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*, 2023. 4

[8] Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. *arXiv preprint arXiv:2411.02429*, 2024. 2

[9] Erik E Guzik, Christian Byrge, and Christian Gilde. The originality of machines: Ai takes the torrance test. *Journal of Creativity*, 33(3):100065, 2023. 3

[10] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025. 2

[11] Kenneth M Heilman. Possible brain mechanisms of creativity. *Archives of Clinical Neuropsychology*, 31(4):285–296, 2016. 2

[12] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 1

[13] Qi Jia, Xiang Yue, Tianyu Zheng, Jie Huang, and Bill Yuchen Lin. Simulbench: Evaluating language models with creative simulation tasks. *arXiv preprint arXiv:2409.07641*, 2024. 3

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[15] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2, 3

[16] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1

[17] Richard E Mayer. Fifty years of creativity research. *Handbook of creativity*, pages 449–460, 1999. 1

[18] Lech Mazur. Llm creative story-writing benchmark. https://github.com/lechmazur/writing, 2025. 1, 2, 3

[19] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2025. 2

[20] Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. Do llms agree on the creativity evaluation of alternative uses? *arXiv preprint arXiv:2411.15560*, 2024. 3

[21] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. 1

[22] Kai Ruan, Xuan Wang, Jixiang Hong, and Hao Sun. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context. *arXiv preprint arXiv:2412.17596*, 2024. 1, 3, 7

[23] Robert J Sternberg. The triarchic theory of intelligence. 1997. 1

[24] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*, 2022. 3

[25] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1

[26] Paul Williams and Carlos Gómez-Rodríguez. A confederacy of models: A comprehensive evaluation of llms on creative

writing. In *UniSC Research Conference*. University of the Sunshine Coast, 2024. 3

[27] Yuhang Wu, Wenmeng Yu, Yean Cheng, Yan Wang, Xiaohan Zhang, Jiazheng Xu, Ming Ding, and Yuxiao Dong. Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models. *arXiv preprint arXiv:2406.09295*, 2024. 4

[28] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 2, 3

[29] Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Llm-evolve: Evaluation for llm's evolving capability on benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16937–16942, 2024. 3

[30] Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, et al. Rpgbench: Evaluating large language models as role-playing game engines. *arXiv preprint arXiv:2502.00595*, 2025. 3

[31] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[32] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1, 2, 3

[33] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024. 2, 3

[34] Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. Redundancy principles for mllms benchmarks, 2025. 1