CreBench: Human-Aligned Creativity Evaluation from Idea to Process to Product

Kaiwen Xue[1*], Chenglong Li[1*], Zhonghong Ou[2 †], Guoxin Zhang[1], Kaoyan Lu[3], Shuai Lyu[1], Yifan Zhu[1], Ping Zong[1], Junpeng Ding[1], Xinyu Liu[4], Qunlin Chen[4], Weiwei Qin[1], Yiran Shen[1], Jiayi Cen[5 †]

[1]Beijing University of Posts and Telecommunications
[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[3]School of Materials Science and Engineering, Shanghai Jiao Tong University
[4]Faculty of Psychology, Southwest University
[5]Southampton Education School, University of Southampton
xkw@bupt.edu.cn, chenglong@bupt.edu.cn, zhonghong.ou@bupt.edu.cn, J.Cen@southampton.ac.uk

Abstract

Human-defined creativity is highly abstract, posing a challenge for multimodal large language models (MLLMs) to comprehend and assess creativity that aligns with human judgments. The absence of an existing benchmark further exacerbates this dilemma. To this end, we propose **CreBench**, which consists of two key components: 1) an evaluation benchmark covering the multiple dimensions from creative idea to process to products; 2) **CreMIT** (**Cre**ativity **M**ultimodal **I**nstruction **T**uning dataset), a multimodal creativity evaluation dataset, consisting of 2.2K diverse-sourced multimodal data, 79.2K human feedbacks and 4.7M multi-typed instructions. Specifically, to ensure MLLMs can handle diverse creativity-related queries, we prompt GPT to refine the human feedback to activate stronger creativity assessment capabilities. CreBench serves as a foundation for building MLLMs that understand human-aligned creativity. Based on the CreBench, we fine-tune open-source general MLLMs, resulting in **CreExpert**, a multimodal creativity evaluation expert model. Extensive experiments demonstrate that the proposed CreExpert models achieve significantly better alignment with human creativity evaluation compared to state-of-the-art MLLMs, including the most advanced GPT-4V and Gemini-Pro-Vision.

Project Page — https://kaixuewen.github.io/Crebench

Introduction

Creativity, a core of human intelligence, is defined as the ability to generate novel, valuable ideas or products (Boden 2004). With rapid advances in MLLMs (Yang et al. 2023; Liu et al. 2023) for vision and reasoning, a key question arises: **do current MLLMs understand creativity as humans intend?** Creativity is abstract and subjective, making it hard to model, and current MLLMs fail to align with human creativity (see Figure 1(c)).

Despite growing demand for creativity-aware AI, evaluating creativity in MLLMs remains underexplored. Existing vision-language benchmarks mainly target objective tasks such as visual question answering (Goyal et al. 2017), captioning (Agrawal et al. 2019), and image-text retrieval (Plummer et al. 2015), relying on well-defined ground truths. In contrast, creativity is open-ended and lacks universal evaluation criteria, making automatic assessment challenging. Moreover, metrics such as BLEU, CIDEr, and CLIPScore fail to capture novelty, usefulness, and human-aligned imagination. Without dedicated creativity benchmarks or datasets, it remains difficult to assess or improve MLLMs' alignment with human creativity.

To address this gap, we introduce CreBench, a comprehensive benchmark for evaluating MLLMs' alignment with human creativity across idea, process, and product. CreBench consists of two components. 1) We propose an open-ended design benchmark that models creativity as a multi-stage process, covering idea generation, refinement, and visual realization. Unlike outcome-only evaluations, CreBench assesses creativity along three core dimensions—creative idea, process, and product—each defined by twelve fine-grained indicators grounded in cognitive science and creativity theory. A five-point behaviorally anchored rubric enables precise, multidimensional evaluation using multimodal data, including verbalized ideas, interaction logs, and final visual outputs. 2) To support large-scale instruction tuning, we construct CreMIT, a high-quality multimodal dataset with over 2.2K creative instances and 79.2K expert annotations from four open-ended design tasks(see Figure 1(a)). Expert feedback is further expanded using GPT-4o into 4.7M instruction–response pairs across six QA formats(see Figure 1(b)). CreMIT provides a scalable resource for training and benchmarking MLLMs on open-ended creativity.

To assess creativity in real-world open-ended tasks, we capture creativity from idea to process to product, going beyond a sole focus on product novelty. Building on this framework, we propose CreExpert, a creativity evaluation expert model fine-tuned from open-source MLLMs. Built on LLaVA-1.5 (Liu et al. 2024a), CreExpert preserves original

[*]These authors contributed equally.
[†]Corresponding authors

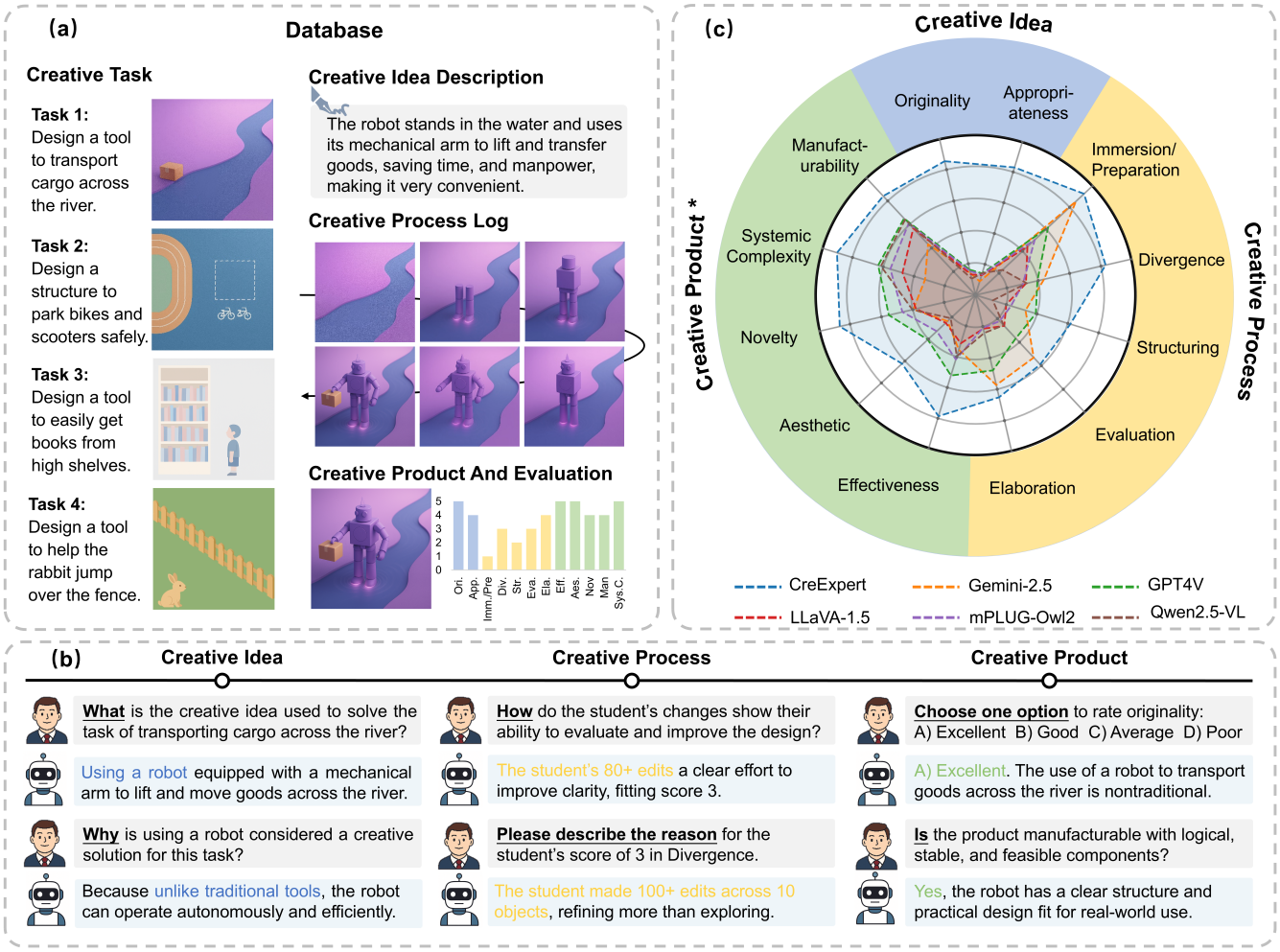

Figure 1: Overview of CreBench. (a) We design a diverse set of creative tasks and build a multi-dimensional database. (b) We use GPT-4o to generate instruction-following data through prompting. (c) Performance of the proposed CreExpert on various creativity evaluation dimensions. (*) indicates scaled data for a better view.

knowledge while acquiring human-aligned creativity understanding. Together, CreBench and CreExpert form a systematic approach for studying and evaluating multimodal creativity.

In summary, our contributions are threefold:

- We introduce CreBench, a fine-grained multidimensional creativity benchmark, together with a multimodal instruction dataset CreMIT, containing 2.2K creative instances, 79.2K expert evaluations, and 4.7M multi-format instructions.
- We propose CreExpert, a multimodal creativity expert model trained on CreBench, which significantly outperforms state-of-the-art MLLMs, including GPT-4V and Gemini-Pro-Vision.
- We release our benchmark, dataset, and CreExpert model, including code and checkpoints, to facilitate future research on multimodal creativity.

Related Works

Creativity in Image Generation

Creativity, unlike aesthetics (Huang et al. 2024; Maerten et al. 2025), emphasizes idea originality and imaginative depth rather than visual beauty alone. DALL-E (Ramesh et al. 2021), and Stable Diffusion (Rombach et al. 2022) have significantly improved the quality and controllability of synthesized images. Several works have explored controlling generation style (Patashnik et al. 2021), enhancing compositionality (Liu et al. 2022), or enabling prompt-based editing (Brooks, Holynski, and Efros 2023). Creativity is often approximated via measures such as diversity or novelty, but these metrics oversimplify the complex human perception of creative value. In this work, we depart from this tradition by treating creativity as a multidimensional, subjective judgment and constructing a dataset explicitly labeled along 12 human-centric dimensions of creative perception.

Human-Aligned Benchmark

Human-aligned benchmark involves constructing evaluation datasets that reflect human preferences across various subjective dimensions. In language, benchmarks like HellaSwag (Zellers et al. 2019), TruthfulQA (Lin, Hilton, and Evans 2022), have been used to assess alignment with human reasoning and judgment. In multimodal domains, efforts such as Seed-Bench (Li et al. 2024) and MM-Bench (Liu et al. 2024c) extend this idea by integrating human preference into visual and cross-modal evaluation. For creative content, however, existing benchmarks either rely on pairwise preference (Kirstain et al. 2023) or aesthetic scoring (Murray, Marchesotti, and Perronnin 2012; Schuhmann et al. 2022), failing to capture the nuanced, multidimensional structure of human creative judgment.

Multimodal Large Language Model

The release of LLMs, such as T5 (Raffel et al. 2020), GPT-4 (Achiam et al. 2023), and LLaMA (Touvron et al. 2023), has demonstrated remarkable capabilities and spurred growing interest in vision and language interaction. Building on this progress, multimodal models such as Otter (Li et al. 2025) and LLaVA (Liu et al. 2023) integrate visual perception with language instruction tuning to enable task generalization and multimodal reasoning. Subsequent works (Pi et al. 2023; Lai et al. 2024; Zhang et al. 2024; Lu et al. 2024; Liu et al. 2024b) further improve MLLMs through advanced fine-tuning strategies for downstream tasks. Despite these advances, human-aligned creativity evaluation remains largely unexplored. To address this gap, we introduce CreBench and CreExpert to support the evaluation and modeling of human-aligned creativity in MLLMs.

CreBench

Evaluation Dimention Suite

Assessing creativity in open-ended, real-world tasks demands a multidimensional framework beyond mere novelty. We adopt a comprehensive rubric tailored to creative problem solving and visual expression, dividing creativity into three dimensions: creative idea, creative process, and creative product (Figure 2). Creativity here involves generating, evaluating, and refining ideas (Brophy 1998), and manipulating visual elements (Urban 2005) to produce novel, appropriate, and aesthetically compelling solutions (Christensen and Ball 2016).

Creative Idea The creative idea assesses the conceptual quality of a solution, focusing on its originality and appropriateness. It evaluates whether the idea presents novel yet contextually relevant mechanisms. Each aspect is rated on a five-point scale:

- ***Originality***: Measures the novelty and divergence from conventional approaches (Runco and Jaeger 2012).
- ***Appropriateness***: Assesses the idea's relevance, feasibility, and alignment with task requirements (Runco and Charles 1993).

Creative Process This dimension measures students' cognitive and visual engagement during problem solving, assessed across five aspects: immersion, divergence, structuring, evaluation, and elaboration. Each is rated on a five-point scale indicating the depth, coherence, and refinement of the creative process.

- ***Immersion/Preparation***: Initial engagement through reflection, observation, and strategic planning (Wallas 1926).
- ***Divergence***: Generation of varied and experimental ideas via open-ended exploration (Joy 1950).
- ***Structuring***: Intentional integration of visual elements into a coherent composition (Okada and Ishibashi 2017).
- ***Evaluation***: Ongoing assessment and refinement of ideas to improve clarity and relevance (Mumford et al. 2013).
- ***Elaboration***: Attention to detail and expressive refinement in the final visual output (Urban 2005).

Creative Product This dimension evaluates the final drawing as both a solution and a creative expression, scored across five aspects: effectiveness, aesthetic, novelty, manufacturability, and systemic complexity. Each is rated on a five-point scale to assess the clarity, originality, feasibility, and design integration of the outcome.

- ***Effectiveness***: How clearly and coherently the drawing communicates the intended solution (Urban 2005).
- ***Aesthetic***: Visual appeal, composition balance, and expressive quality (Christensen and Ball 2016).
- ***Novelty***: Originality in form, content, or symbolic representation (Torrance 1966).
- ***Manufacturability***: Feasibility of real-world construction and functionality (Charyton et al. 2011).
- ***Systemic Complexity***: Integration of multiple functional components into a coherent system (Howard, Culley, and Dekoninck 2008).

Dataset Construction

This section presents an overview of the CreMIT dataset construction, as illustrated in Figure 2. Stage 1 outlines a subjective experiment collecting creative ideas, processes, and products from human and AI participants at a 1:3 human-to-AI ratio across four problem-solving tasks. Stage 2 covers expert evaluation, yielding 79.2K human feedback entries over 2.2K multi-dimensional instances. Stage 3 details how GPT-4o refines creative reports to generate 4.7M instruction-following samples across multiple creativity dimensions and question types, with their distribution summarized in Figure 3.

Experiment Preparation Figure 2 (Stage 1) illustrates the distribution of data by participant type, modality, and dimension. As shown in Figure 3(a), the dataset covers four problem-solving tasks with balanced sample distributions across tasks.

① ***Task Design***: We designed four open-ended, real-world problem-solving scenarios, such as the "cargo river

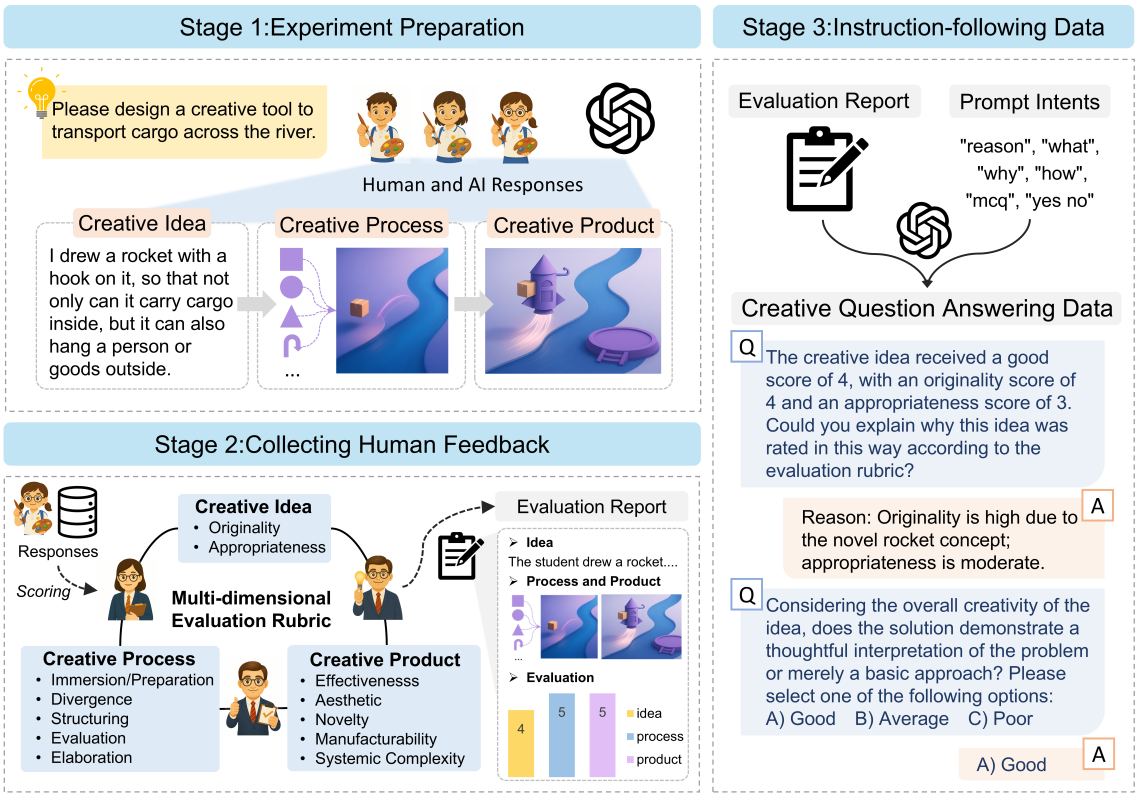

Figure 2: Overview of CreMIT construction procedure. Stage 1: We collect diverse solutions (creative idea, creative process, and creative product) generated by students and AI based on open-ended creativity tasks. Stage 2: Innovation experts evaluate each solution across 12 indicators spanning three dimensions, producing detailed assessment reports. Stage 3:Using six types of prompts, we employ GPT-4o to generate multidimensional instruction-following data based on the expert feedback.

crossing" (Figure 2, Stage 1), to elicit diverse and non-routine creative responses (Runco and Acar 2012). These tasks captured both the final creative products and the underlying ideas and processes, building a comprehensive, multimodal dataset for creativity assessment.

② ***Subject Selection***: A total of 512 secondary students were recruited through stratified cluster sampling from five schools to ensure demographic and cognitive diversity (Figure 2, Stage 1). All participants passed a color vision test (Yang et al. 2022) and completed a baseline creativity assessment. Each completed three creative tasks, contributing textual ideas, behavioral logs, and visual outputs forming a multi-dimensional, multimodal dataset. This comprehensive data collection ensures a rich foundation for evaluating creativity across individuals with varied backgrounds and expression modalities.

Collecting Human Feedback Figure 2 (Stage 2) illustrates the expert evaluation process used to assess creativity across multiple dimensions.

① ***Expert Annotation Protocol***: To ensure reliable creativity assessment, we implemented a rigorous expert annotation and quality control protocol. Three experts in educational creativity were recruited following the Consensual Assessment Technique (CAT), each with extensive experience. Before annotation, all experts underwent two training sessions to familiarize themselves with the assessment framework, clarify rubrics, and calibrate standards using example cases minimizing inter-rater variability. During annotation, experts evaluated both human and AI outputs across all three modalities. Annotation quality was ensured through ongoing agreement monitoring, regular calibration meetings, and both automated checks (e.g., completeness, consistency) and manual reviews for low-quality or ambiguous cases.

Dataset	Year	Modality			AIGC	H.F.	I.T.	O.W.
		Text	Proc.	Img				
OpenSketch	2021	✗	✗	✓	✗	✓	✗	✗
CID	2022	✓	✗	✓	✓	✓	✗	✓
Cambridge AUT	2023	✓	✗	✗	✗	✓	✗	✗
AesBench	2024	✓	✗	✓	✓	✓	✓	✓
APDD	2024	✗	✗	✓	✗	✓	✗	✗
IDEA	2024	✓	✗	✓	✗	✓	✗	✗
CreBench	2025	✓	✓	✓	✓	✓	✓	✓

Table 1: Feature comparison of creativity-related datasets. Proc.: process logs; H.F.: human feedback; I.T.: instruction tuning; O.W.: open world.

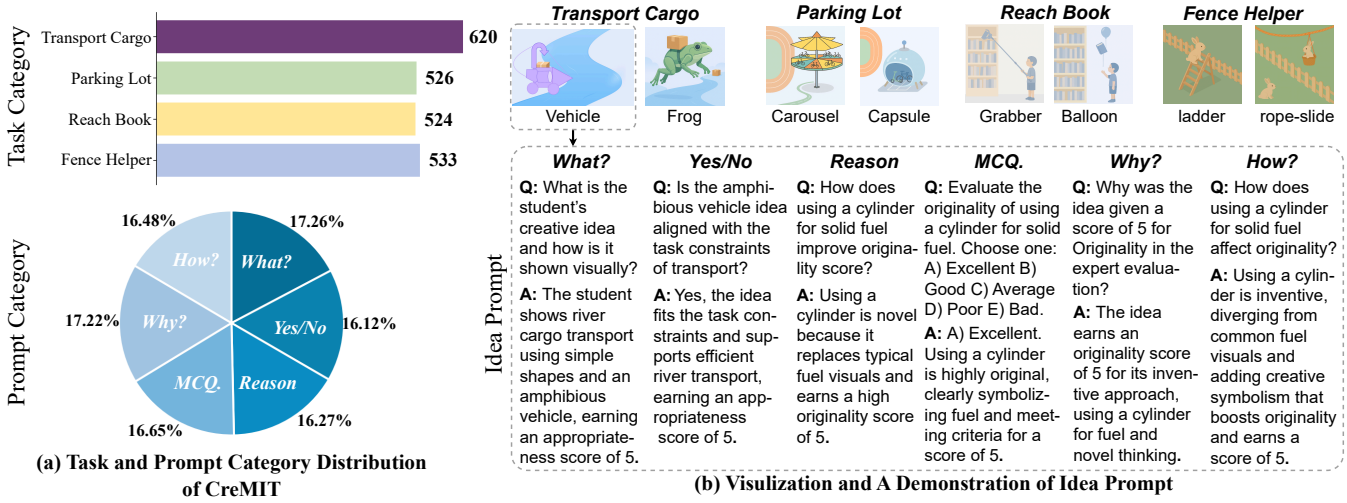

Figure 3: Data distribution and visualization. (a) We analyze the sample number for each task and the prompt number for each prompt category. (b) We only demonstrate the first sample of the idea prompt (among idea, process, and product prompts).

② ***Expert Feedback***: Based on this protocol, we constructed a multi-dimensional dataset through expert evaluations of 4 tasks, 3 creativity dimensions, and 12 sub-indicators. The dataset includes over 2.2K visual instances and 79.2K expert ratings. All annotations were conducted by qualified professionals to guarantee reliable creativity assessment. We have computed inter-rater reliability across all dimensions, the average Fleiss's κ = 0.71 and ICC (2,1) = 0.78, indicating substantial agreement.

Instruction-following Data Figure 2 (Stage 3) illustrates the construction of large-scale instruction-following data from expert feedback. Building on the expert-annotated dataset, we prompt GPT-4 to convert expert feedback into instruction–response pairs following prior work (Wu et al. 2024a), yielding 4.7M samples across creativity dimensions and question types. As shown in Figure 3(b), to support diverse real-world queries and enhance MLLMs' creativity understanding, we design six question types:

① ***Reasoning-style***: focus on analyzing the rationale behind expert-assigned scores, encouraging the model to justify the evaluation results based on rubric-defined criteria.

② ***What-style***: investigate the key features, intentions, or expressive elements of the creative idea that contribute to its originality and appropriateness;

③ ***How-style***: explore implementation strategies or mechanisms, guiding the model to explain how the creative idea was realized or executed;

④ ***Why-style***: seek to reveal the causes behind evaluation outcomes, encouraging causal interpretation;

⑤ ***Yes/No-style***: questions provide binary judgments on aspects such as novelty, relevance, or feasibility, followed by brief justifications;

⑥ ***MCQ–style***: convert evaluation scenarios into structured rating selections (Excellent to Bad), aligned with rubric-based judgment frameworks (Liu et al. 2024a; Lu et al. 2022; Wu et al. 2024b).

Following this pipeline, the final CreMIT dataset comprises 4.7M multi-typed instructions explicitly crafted to enhance creativity evaluation in MLLMs. As summarized in Table 1, CreMIT significantly advances in scale and diversity, offering a valuable benchmark for MLLM-based creativity understanding and assessment.

CreExpert

Model Architecture

The proposed CreExpert is developed following the architectural paradigm of LLaVA-1.5 (Liu et al. 2023), and consists of three core modules: a vision encoder built upon CLIP-ViT-L14 (Radford et al. 2021) with a 336×336 input resolution, which encodes input images into 576 visual tokens; a modality bridging component implemented as a two-layer MLP that aligns visual and linguistic representations; and a language decoder initialized from the open-source Vicuna-v1.5 (Zheng et al. 2023), which is responsible for instruction understanding and response generation. In this work, we instantiate the CreExpert model based on the LLaVA-1.5-7B variant, aiming to evaluate its capability in creativity assessment tasks using the CreMIT dataset.

Supervised Fine-Tuning

The development of open-source Multimodal Large Language Models (MLLMs) (Liu et al. 2023) generally consists of two key phases: (1) aligning the visual encoder with the language model using large-scale image-text datasets collected from the web (Lu et al. 2022), and (2) enhancing multimodal reasoning through instruction tuning with curated vision-language datasets (Liu et al. 2023). In contrast to previous work that emphasizes general-purpose understanding (Liu et al. 2023; Su et al. 2023; Ye et al. 2023), the objective of this study is to improve the ability of MLLMs to perform creativity evaluation. To achieve this, we apply supervised instruction tuning to models that have been pretrained on generic visual tasks, using the proposed CreMIT

Model	Creative Idea			Creative Process						Creative Product						Overall	Rank
	Ori.	*App.*	*Avg.*	*Imm./Pre.*	*Div.*	*Str.*	*Eva.*	*Ela.*	*Avg.*	*Eff.*	*Aes.*	*Nov.*	*Man.*	*Sys. C.*	*Avg.*		
CreExpert (Ours)	**85.35%**	**82.92%**	**84.14%**	**92.24%**	**82.97%**	**61.64%**	**58.97%**	**65.13%**	**72.19%**	**39.32%**	**31.22%**	**43.30%**	**42.05%**	**45.00%**	**40.18%**	**65.50%**	**1**
GPT-4V	16.20%	14.12%	15.16%	61.64%	39.48%	39.21%	36.70%	48.01%	45.01%	26.10%	20.28%	27.64%	32.56%	31.63%	27.64%	29.27%	2
Gemini-Pro-Vision	14.00%	8.93%	11.47%	85.32%	43.72%	32.49%	53.07%	57.33%	54.39%	17.75%	11.87%	19.84%	21.62%	16.41%	17.50%	27.78%	3
mPLUG-Owl2	15.16%	13.51%	14.34%	50.29%	30.19%	23.05%	22.09%	20.91%	29.31%	20.89%	16.46%	23.31%	30.46%	27.69%	23.76%	22.47%	4
LLaVA-1.5-7B	13.23%	12.89%	13.06%	44.11%	32.17%	20.02%	26.19%	21.41%	28.78%	15.67%	12.63%	18.98%	28.35%	23.72%	19.87%	20.57%	5
InstructBLIP	9.26%	14.89%	12.08%	46.92%	35.01%	19.42%	24.88%	22.77%	29.80%	16.94%	9.83%	17.19%	29.34%	25.27%	19.71%	20.53%	6
Seed1.5-VL	19.97%	9.82%	14.90%	57.74%	17.97%	11.65%	18.30%	18.26%	24.78%	14.74%	26.27%	19.75%	21.03%	17.49%	19.86%	19.85%	7
Qwen2.5-VL	10.48%	14.23%	12.36%	22.97%	32.70%	11.25%	25.62%	24.15%	23.34%	19.84%	10.54%	20.12%	32.06%	30.74%	22.66%	19.45%	8
MiniGPT-v2	10.28%	11.89%	11.09%	20.18%	30.83%	13.09%	22.10%	19.72%	21.18%	18.00%	8.24%	14.90%	25.09%	21.27%	17.50%	16.59%	9
GLM	7.39%	9.19%	8.29%	40.98%	28.31%	20.14%	20.58%	19.01%	25.80%	12.09%	6.90%	10.03%	19.08%	17.93%	13.21%	15.77%	10
MiniGPT-4	5.89%	4.87%	5.38%	10.23%	10.83%	9.42%	14.80%	10.08%	11.07%	7.63%	10.23%	13.28%	6.72%	12.06%	9.98%	8.81%	11
TinyGPT	3.98%	2.59%	3.29%	8.72%	4.80%	8.92%	10.17%	8.15%	8.15%	7.22%	8.43%	9.82%	9.02%	4.95%	7.89%	6.44%	12

Table 2: Performance comparisons of the proposed CreExpert with existing MLLMs. Ori.: Originality, App.: Appropriateness, Imm./Pre.: Immersion/Preparation, Div.: Divergence, Str.: Structuring, Eva.: Evaluation, Ela.: Elaboration, Eff.: Effectiveness, Aes.: Aesthetic, Nov.: Novelty, Man.: Manufacturability, Sys. C.: Systemic Complexity. The Overall score is computed as the average across the three major dimensions: *Creative Idea*, *Creative Process*, and *Creative Product*.

dataset. Drawing on insights from recent research (Wu et al. 2024a), our training method preserves the model's general knowledge while equipping it with task-specific reasoning abilities tailored to creativity assessment. For the sake of training efficiency and fair comparison, the visual encoder is kept frozen, and fine-tuning is applied only to the projection module and the language model. As a result, we obtain a unified multimodal expert model that is capable of evaluating creativity across the full process, including idea generation, process execution, and final product realization.

Experiments

Dataset and Metrics

CreBench contains 2.2K samples from four tasks, each comprising a creative idea, process, and product. The dataset is split evenly into fine-tuning and evaluation sets. CreBench evaluates MLLMs across 12 fine-grained dimensions, using the Pearson correlation coefficient to measure consistency between model predictions and human feedback, reflecting alignment with human judgment.

Implementation Details

We fine-tuned pre-trained multimodal language models on the constructed CreMIT dataset using LoRA within the LLaMA-Factory framework. To ensure fair comparison, we adopted the default hyperparameter settings provided by the original models during the fine-tuning process. All training and evaluation were conducted on a server equipped with eight NVIDIA Tesla A40 48GB GPUs.

Main Result

We compare the performance of CreExpert with 11 top MLLMs. The comparison includes two widely used proprietary models, GPT-4V and Gemini Pro Vision, as well as 9 advanced open-source variants, i.e., LLaVA-1.5-7B, mPLUG-Owl2, InstructBLIP, Seed1.5-VL, Qwen2.5-VL, MiniGPT-v2, GLM, MiniGPT-4, and TinyGPT. All of these open-source models represent recent variants built upon foundational vision-language architectures. As shown in Table 2, CreExpert achieves the best performance, surpassing the state-of-the-art LLaVA-1.5-7B baseline by nearly 45%. Among existing closed-source models, GPT-4V (Yang et al. 2023) performs the best, yet still lags behind our CreExpert by more than 35% in terms of overall score. These results demonstrate the superior creativity evaluation capability of CreExpert and highlight the effectiveness of the constructed CreMIT dataset in enhancing the alignment of multimodal foundation models.

Task	Model	Creative Idea		Overall
		Ori.	App.	
Transport	Baseline	12.80%	11.72%	12.26%
	CreExpert	**72.42%**	**69.40%**	**70.91%**
	Improvement	+59.62%	+57.68%	+58.65%
Parking	Baseline	14.98%	14.96%	14.97%
	CreExpert	69.08%	65.17%	67.13%
	Improvement	+54.10%	+50.21%	+52.16%
Reach	Baseline	14.12%	15.06%	14.59%
	CreExpert	**83.91%**	**78.07%**	**80.99%**
	Improvement	+69.79%	+63.01%	+66.40%
Fence	Baseline	11.90%	14.12%	13.01%
	CreExpert	**80.28%**	**83.55%**	**81.92%**
	Improvement	+68.38%	+69.43%	+68.91%

Table 3: Comparison of baseline MLLMs and the proposed CreExpert model on the Creative Idea dimension across four creative tasks.

Ablation Study

Creative Idea Evaluation Ability. From Table 3, we observe that fine-tuning baseline MLLMs with CreMIT significantly enhances their ability to generate creative ideas across various tasks. Among the four tasks evaluated, Transport, Parking, Reach, and Fence, the largest overall improvement is observed in the Fence task, where performance increased by almost 69%. Notably, across all tasks, the Originality dimension consistently shows the most substantial gains, with improvements ranging from +54.10% to +69.79%. This suggests that MLLMs benefit particularly from CreMIT in gen-

Task	Model	Creative Process					Overall
		Imm./Pre.	Div.	Str.	Eva.	Ela.	
Transport	Baseline	38.16%	25.87%	16.17%	22.31%	18.48%	24.20%
	CreExpert	**89.23%**	**80.64%**	**58.72%**	**57.76%**	**62.45%**	**69.76%**
	Improvement	+51.07%	+54.77%	+42.55%	+35.45%	+43.97%	+45.56%
Parking	Baseline	41.53%	30.92%	20.25%	23.72%	21.56%	27.60%
	CreExpert	**87.43%**	**75.63%**	**56.41%**	**53.62%**	**60.42%**	**66.70%**
	Improvement	+45.90%	+44.71%	+36.16%	+29.90%	+38.86%	+39.10%
Reach	Baseline	39.07%	27.16%	28.32%	22.12%	19.53%	27.24%
	CreExpert	**90.52%**	**81.83%**	**59.73%**	**56.49%**	**61.27%**	**69.97%**
	Improvement	+51.45%	+54.67%	+31.41%	+34.37%	+41.74%	+42.73%
Fence	Baseline	36.91%	25.47%	24.82%	19.16%	18.64%	25.00%
	CreExpert	**85.17%**	**72.28%**	**51.34%**	**49.96%**	**58.72%**	**63.4%**
	Improvement	+48.26%	+46.81%	+26.52%	+30.8%	+40.08%	+38.49%

Table 4: Comparison of baseline MLLMs and the proposed CreExpert model on the Creative Process dimension across four creative tasks.

Task	Model	Creative Product					Overall
		Eff.	Aes.	Nov.	Man.	Sys. C.	
Transport	Baseline	13.35%	11.24%	17.54%	28.12%	22.69%	18.59%
	CreExpert	**36.84%**	**28.49%**	**41.28%**	**40.73%**	**43.72%**	**38.21%**
	Improvement	+23.49%	+17.25%	+23.74%	+12.61%	+21.03%	+19.62%
Parking	Baseline	39.92%	24.81%	10.94%	20.67%	27.71%	24.81%
	CreExpert	**53.45%**	**30.17%**	**20.26%**	**28.89%**	**39.84%**	**34.52%**
	Improvement	+13.53%	+5.36%	+9.32%	+8.22%	+12.13%	+9.71%
Reach	Baseline	22.94%	19.97%	17.28%	27.90%	26.13%	22.84%
	CreExpert	**28.93%**	**24.06%**	**20.29%**	**25.09%**	**31.87%**	**26.05%**
	Improvement	+5.99%	+4.09%	+3.01%	-2.81%	+5.74%	+3.21%
Fence	Baseline	25.10%	11.09%	19.79%	19.18%	39.57%	22.95%
	CreExpert	**30.35%**	**28.62%**	**35.86%**	**32.16%**	**48.76%**	**35.15%**
	Improvement	+5.25%	+17.53%	+16.07%	+12.98%	+9.19%	+12.2%

Table 5: Comparison of baseline MLLMs and the proposed CreExpert model on the Creative Product dimension across four creative tasks.

erating novel ideas. A plausible explanation is that creative idea generation primarily involves textual expression, which enables MLLMs to better align with the key elements emphasized in human creativity evaluation, such as novelty and divergence.

Creative Process Evaluation Ability. Table 4 compares the baseline MLLMs and the CreExpert on the Creative Process dimension across four creative tasks. The results demonstrate that CreExpert consistently outperforms the baseline across all sub-dimensions—including Immersion/Preparation, Divergence, Structuring, Evaluation, and Elaboration. The most substantial improvements are observed in Immersion/Preparation and Divergence, exceeding 50% across most tasks. For example, in the Reach task, CreExpert improves Immersion/Preparation by +51.45% and Divergence by +54.67%. These results indicate that CreMIT enables MLLMs to evaluate novel ideas and simulate the multi-stage reasoning and exploration process underlying human creativity. Consistent gains in structuring and evaluation also suggest enhanced planning and critical thinking capabilities in the generated responses.

Creative Product Evaluation Ability. Table 5 compares the performance of baseline MLLMs and the CreExpert model on the Creative Product dimension across four tasks. CreExpert consistently improves over the baseline across all sub-dimensions, including Effectiveness, Aesthetics, Novelty, Manufacturability, and System Complexity. Notably, the Transport task exhibits the most prominent overall gain (+19.62%), with improvements above 20% in several key sub-dimensions such as Novelty (+23.74%) and System Complexity (+21.03%). In contrast, the Reach task shows only modest gains, with a slight drop in Manufacturability. These results suggest that while CreMIT strengthens the product-level creative expression of MLLMs, the degree of improvement varies across tasks depending on domain-specific constraints. Importantly, improvements in Novelty and Aesthetics—dimensions closely tied to user-perceived creativity—indicate that the model generates not only functional but also appealing and original solutions.

Conclusion

In this work, we attempt to leverage the innovation perception capability of multi-modality foundation models. Specifically, we first construct a corpus-rich, multi-dimensional innovation process evaluation database through human creative solutions, based on which we further establish a comprehensively annotated multimodal instruction tuning dataset for innovation processes (CreMIT). In addition, we propose multimodal innovation expert models based on innovation instruction fine-tuning, achieving significantly improved innovation perception performance. Building on these multimodal innovation expert models, we design a CreExpert to align model-based innovation evaluation with human judgments of innovation. We believe this work represents a solid step toward enhancing the innovation perception ability of MLLMs, and we hope that our contribution will inspire the research community to develop multi-modality foundation models capable of understanding highly abstract human creative processes.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (Grant 2024YFC3308500), the Natural Science Foundation Joint Fund for Innovation and Development of Chongqing Municipal Education Commission (Grant CSTB2024NSCQ-LZX0132), the National Natural Science Foundation of China (Grant 62406036), the Beijing Municipal Natural Science Foundation (Grant L251042), the State Key Laboratory of Networking and Switching Technology (Grant NST20250110), and the SMP-Zhipu.AI Large Model Cross-Disciplinary Fund (Grant ZPCG20241029322). Ethical approval for this study was obtained from the Ethics Committee of the University of Southampton, and written informed consent was obtained from all participants and their guardians.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Brophy, D. R. 1998. Understanding, measuring, and enhancing individual creative problem-solving efforts. *Creativity Research Journal*, 11(2): 123–150.

Charyton, C.; Jagacinski, R. J.; Merrill, J. A.; Clifton, W.; and DeDios, S. 2011. Assessing creativity specific to engineering with the revised creative engineering design assessment. *Journal of Engineering Education*, 100(4): 778–799.

Christensen, B. T.; and Ball, L. J. 2016. Dimensions of creative evaluation: Distinct design and reasoning strategies for aesthetic, functional and originality judgments. *Design studies*, 45: 116–136.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Howard, T. J.; Culley, S. J.; and Dekoninck, E. 2008. Describing the creative design process by the integration of engineering design and cognitive psychology literature. *Design studies*, 29(2): 160–180.

Huang, Y.; Yuan, Q.; Sheng, X.; Yang, Z.; Wu, H.; Chen, P.; Yang, Y.; Li, L.; and Lin, W. 2024. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

Joy, P. 1950. Guilford. 1950. Creativity. *American Psychologist*, 5(9): 444–454.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36: 36652–36663.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, B.; Ge, Y.; Ge, Y.; Wang, G.; Wang, R.; Zhang, R.; and Shan, Y. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13299–13308.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Cahyono, J. A.; Yang, J.; Li, C.; and Liu, Z. 2025. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, 423–439. Springer.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Maerten, A.-S.; Chen, L.-W.; De Winter, S.; Bossens, C.; and Wagemans, J. 2025. LAPIS: A novel dataset for personalized image aesthetic assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6302–6311.

Mumford, M. D.; Giorgini, V.; Gibson, C.; and Mecca, J. 2013. Creative thinking: Processes, strategies and knowledge. In *Handbook of research on creativity*, 249–264. Edward Elgar Publishing.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, 2408–2415. IEEE.

Okada, T.; and Ishibashi, K. 2017. Imitation, inspiration, and creation: Cognitive process of creative drawing by copying others' artworks. *Cognitive science*, 41(7): 1804–1837.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2085–2094.

Pi, R.; Gao, J.; Diao, S.; Pan, R.; Dong, H.; Zhang, J.; Yao, L.; Han, J.; Xu, H.; Kong, L.; et al. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Runco, M. A.; and Acar, S. 2012. Divergent thinking as an indicator of creative potential. *Creativity research journal*, 24(1): 66–75.

Runco, M. A.; and Charles, R. E. 1993. Judgments of originality and appropriateness as predictors of creativity. *Personality and individual differences*, 15(5): 537–546.

Runco, M. A.; and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal*, 24(1): 92–96.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; and Cai, D. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Urban, K. K. 2005. Assessing creativity: The Test for Creative Thinking-Drawing Production (TCT-DP). *International Education Journal*, 6(2): 272–280.

Wallas, G. 1926. *The art of thought*. 24. Harcourt, Brace.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024a. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25490–25500.

Wu, H.; Zhu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Li, C.; Wang, A.; Sun, W.; Yan, Q.; et al. 2024b. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, 360–377. Springer.

Yang, Y.; Xu, L.; Li, L.; Qie, N.; Li, Y.; Zhang, P.; and Guo, Y. 2022. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19861–19869.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800.

Zhang, S.; Sun, P.; Chen, S.; Xiao, M.; Shao, W.; Zhang, W.; Liu, Y.; Chen, K.; and Luo, P. 2024. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European conference on computer vision*, 52–70. Springer.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.