A Causality-aware Paradigm for Evaluating Creativity of Multimodal Large Language Models

Zhongzhan Huang*, Shanshan Zhong*, Pan Zhou*, Shanghua Gao, Marinka Zitnik, Liang Lin, *Fellow, IEEE*

Abstract—Recently, numerous benchmarks have been developed to evaluate the logical reasoning abilities of large language models (LLMs). However, assessing the equally important creative capabilities of LLMs is challenging due to the subjective, diverse, and data-scarce nature of creativity, especially in multimodal scenarios. In this paper, we consider the comprehensive pipeline for evaluating the creativity of multimodal LLMs, with a focus on suitable evaluation platforms and methodologies. First, we find the Oogiri game—a creativity-driven task requiring humor, associative thinking, and the ability to produce unexpected responses to text, images, or both. This game aligns well with the input-output structure of modern multimodal LLMs and benefits from a rich repository of high-quality, human-annotated creative responses, making it an ideal platform for studying LLM creativity. Next, beyond using the Oogiri game for standard evaluations like ranking and selection, we propose LoTbench, an interactive, causality-aware evaluation framework, to further address some intrinsic risks in standard evaluations, such as information leakage and limited interpretability. The proposed LoTbench not only quantifies LLM creativity more effectively but also visualizes the underlying creative thought processes. Our results show that while most LLMs exhibit constrained creativity, the performance gap between LLMs and humans is not insurmountable. Furthermore, we observe a strong correlation between results from the multimodal cognition benchmark MMMU and LoTbench, but only a weak connection with traditional creativity metrics. This suggests that LoTbench better aligns with human cognitive theories, highlighting cognition as a critical foundation in the early stages of creativity and enabling the bridging of diverse concepts. Project Page.

Index Terms—Creativity, Multimodal Large Language Models, Benchmark, Causal Intervention.

1 Introduction

Large language models (LLMs) [1], [2], [3], [4], [5], [6], [7] have catalyzed in a transformative era in neural network reasoning, revolutionizing various domains within artificial intelligence. Recently, numerous benchmarks [8], [9], [10], [11], [12], [13], [14] have been proposed to evaluate LLMs' rigorous logical reasoning abilities, spurring the development of methods to enhance these capabilities, particularly the representative Chain-of-Thought (CoT) based methods [15], [16], [17], [18], [19]. These methods equip LLMs with human-like step-by-step reasoning capacity, enabling them to excel in complex reasoning tasks ranging from language comprehension to visual understanding. As illustrated in Fig. 1 (a), CoT instills LLMs with a sequential thinking process where each subsequent thought builds upon the previous one. This paradigm enhances precision and rigor in logical processing, making it highly effective for problems requiring closely linked logical reasoning. While CoT-based methods have proven effective for logical reasoning, they may fall short in capturing another equally important thinking mode: creative reasoning. This limitation stems primarily from their sequential nature. For instance, proving an algebraic inequality often follows a step-by-step CoT process, progressing from one inequality to the next. In contrast, a more creative solution might arise from an intuitive flash, such as a geometric interpretation. This type of insight, known as "Leap-of-Thought" (LoT) or mental leap [22], [23], [24], [25], represents the art of non-sequential thinking through association, drawing parallels between seemingly unrelated concepts, and facilitating a "leap" in knowledge transfer. Unlike CoT reasoning, LoT, as depicted in Fig. 1 (a), fosters associative reasoning and encourages thinking outside the box, bridging disparate ideas and facilitating conceptual leaps. Embracing LLMs with strong LoT abilities can unlock significant potential for creativity, contributing to advancements in creative applications.

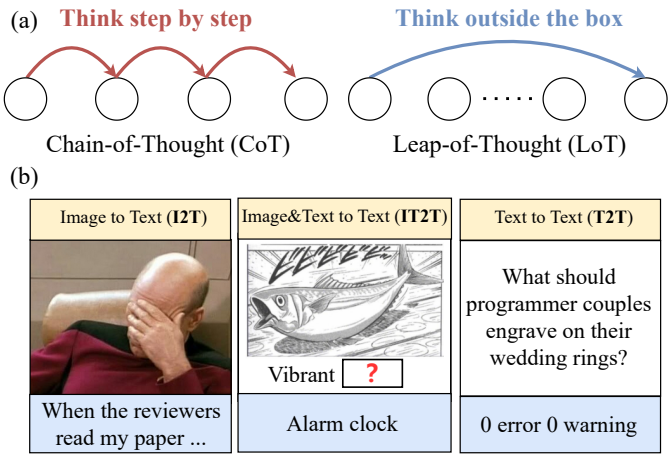

Fig. 1. Leap-of-Thought (LoT) for creativity. (a) Comparison of CoT and LoT. "◯" denotes the thought and "→" represents the connection between two thoughts. LoT is one of most important ability in creativity [20], [21]. (b) Examples of the three types of LoT-based Oogiri games. Players are required to make surprising and creative humorous responses (blue box) to the given multimodal information e.g., images, text, or both.

• *Z. Huang, S. Zhong and L. Lin are with the Sun Yat-sen University, China. P. Zhou is with the Singapore Management University. S. Gao and M. Zitnik are with the Harvard University, USA. L. Lin is also with Guangdong Key Laboratory of Big Data Analysis and Processing and Peng Cheng Laboratory. *Z. Huang, S. Zhong and P. Zhou have equal contribution for this paper. Corresponding author is L. Lin (e-mail: linliang@ieee.org).*

However, this crucial aspect of creativity presents unique challenges in benchmarking due to the subjective, diverse, and data-scarce nature of creative responses, especially for the multimodal scenarios [1]. This limitation hinders the development of methods to stimulate multimodal LLMs' creative abilities.

In this paper, we consider the comprehensive pipeline for evaluating multimodal LLMs' (MLLMs) creativity, with a focus on suitable evaluation platform and methodologies.

(1) **Suitable evaluation platform**. Thoroughly assessing LoT is challenging due to the complexity of measuring creative thinking [26], [27], [28] and the difficulty in gathering pertinent data, since generating novel ideas is challenging, even for humans [29]. Given these constraints, we propose studying LoT in MLLMs through the lens of Oogiri-style humor generation. Oogiri, a traditional Japanese creative game [30], requires participants to provide unexpected and humorous responses to prompts in the form of images, text, or a combination of both, as shown in Fig. 1 (b). This game challenges MLLMs to demonstrate a sudden burst of insight and strong associative thinking, presenting a unique challenge for CoT-based methods. Moreover, the Oogiri game aligns with the input-output paradigm of current MLLMs and, due to its popularity, offers a wealth of high-quality, human-annotated creative responses, making it an ideal platform for exploring LoT ability of MLLMs. Moreover, to investigate the LoT ability of LLMs in the Oogiri game, we initially present the multilingual and multimodal Oogiri-GO dataset which comprises more than 130,000 high-quality Oogiri samples in English, Chinese, and Japanese, and curated to prompt textual humor in response to inputs that can be images, text, or both.

(2) **Suitable evaluation methodologies**. First, following the popular standard LLM benchmark paradigm [8], [9], [10], [11], [12], [13], [14], we also establish a series of standard LLM evaluations by Oogiri-GO, such as ranking and selection [1], [31], [32], [33]. We find that even advanced LLMs and reasoning frameworks [3], [17], [34], [35], including GPT-4 and CoT, despite their exceptional reasoning capabilities and extensive prior knowledge of various forms of humor [17], still struggle to demonstrate adequate LoT ability for creative humor generation. Moreover, while standard evaluations offer simplicity and low assessment costs, we identify inherent risks associated with their use in assessing creativity, such as information leakage and limited interpretability. To address these issues, we first propose training LLMs to assist in generating specific high-quality human-level creative responses (HHCRs). Additionally, we introduce a multi-round interactive [36], [37] evaluation paradigm, LoTbench. With causal reasoning techniques, LoTbench measures creativity by analyzing the average number of rounds required for an LLM to reach HHCRs. Fewer required rounds indicate higher human-level creativity. LoTbench not only effectively evaluates LLM creativity but also provides interpretable visualizations of the LLM's innovative thought process during interactions.

The results of LoTbench demonstrate that while most MLLMs exhibit limited creativity, the gap between their creativity and human creativity is not substantial. Current MLLMs show the potential to surpass human creativity. Furthermore, we observe a strong positive correlation between the results of the well-known multimodal LLM cognition benchmark MMMU [11] and LoTbench, but a low correlation with standard creativity evaluation. This indicates that LoTbench's creativity measurements align more closely with human cognitive theories [38], [39], [40], [41], [42], suggesting that cognition serves as a critical foundation in the early stages of creativity, enabling leaps across diverse conceptual spaces. Unlike the conference version [1], we have improved the sampling efficiency of creativity data and proposed a more reasonable causality-aware paradigm for evaluating the creativity of multimodal LLMs.

In summary, our **contributions** are as follows:

(i) We discover the ideal platform for studying the LLMs' creativity, the Oogiri game, and develope a comprehensive standard evaluation pipeline to analyze and discuss how to stimulate LLM creativity.

(ii) Due the inherent risks in standard creativity evaluations, such as information leakage and limited interpretability, we further propose an interactive, causality-aware benchmark called LoTbench. We find that LoTbench align with human cognitive theories and reveal that while the current LLMs'creativity is not very high, it's close to human levels and has the potential to surpass human creativity.

2 Related Works

(1) Multimodal LLMs and their creativity. Recently, multimodal Language Models [34], [43], [44], [45] have garnered significant attention, particularly due to their impressive reasoning abilities [4], [5], [6]. Moreover, there is a growing focus on exploring the creativity [46], [47], [48] of LLMs for applications such as scientific discovery [49], [50], [51], creative writing [52], [53], [54], etc.

(2) Computational humor is a branch of computational linguistics and artificial intelligence that uses computers in humor research [55], [56], which encompasses various tasks, including humor detection [57], [58], [59] and humor generation [60], [61], [62], etc. With the advancement of generative LLMs [34], [35], [45], humor generation has become a popular focus while humor generation still faces challenges such as insufficient punchlines [63] and limited in multimodal contexts [64], [65].

(3) Chain-of-Thought based Methods provide the models with "chain of thoughts" [15], [16], [18], [19], [66], [67], [68], [69], i.e., reasoning exemplars [15], or a simple prompt "Let's think step by step" [17], to encourage LLMs to engage in reasoning rather than simply providing answers directly [70].

3 Evaluation Platform: Oogiri game

Unlike most logic reasoning benchmarks [8], [9], [10], [11], [12], [13], [14], creativity tasks suffer from a severe lack of data and high annotation costs, as it is challenging even for humans to generate a large number of creative responses. Recently, some works [31], [32], [33] have been proposed to study the lateral thinking capabilities of LLMs, but they primarily focus on information in the pure text modality. This makes exploring the creativity of multimodal LLMs a significant challenge, thereby hindering the development of methods to enhance their creativity. Fortunately, in this paper, we find that the Oogiri game serves as an ideal evaluation platform.

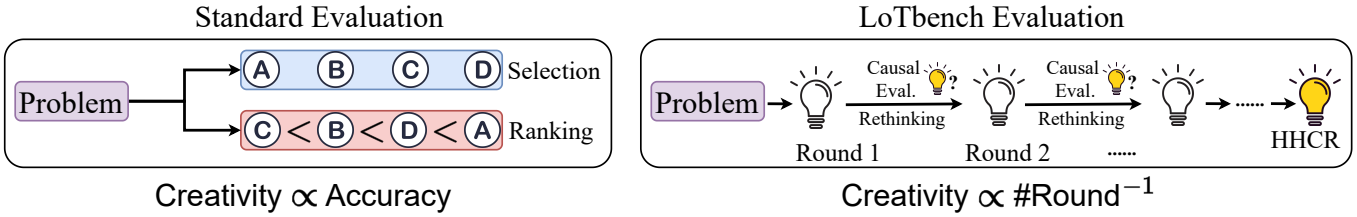

Fig. 2. The motivation of different paradigms to measure creativity. (Left) Standard Evaluation: Assess LLMs by designing selection and ranking tasks. Higher accuracy indicates greater creativity. (Right) LoTbench: LLMs generate multi-round responses, evaluated by a causal evaluator to determine whether they approach high-quality human-level creative responses (HHCRs). If not, the model enters a rethinking phase for the next round. Creativity is inversely proportional to the number of response rounds (# Round).

TABLE 1
Data distribution of the Oogiri-GO dataset. For the IT2T task, its English version is not available due to cultural preference.

Category	English	Chinese	Japanese	Total
I2T	17, 336	32, 130	40, 278	89, 744
T2T	6, 433	15, 797	11, 842	34, 072
IT2T	—	912	9, 420	10, 332

Oogiri game (大喜利) is a general term for a series of traditional Japanese comedy games. In ancient times, there were different types of Oogiri, such as actors performing sumo wrestling, telling ghost stories, etc. The modern Oogiri game mainly refers to one specific type known as Tonchi (頓智), typically presented in the format of game shows or intellectual quiz programs [30]. Players are provided with various multimodal contents, which can be simple questions, random images, etc., and are then prompted to come up with humorous, creative responses to achieve surprising comedic effects, as the examples are shown in Fig. 1 (b). It is worth noting that the character "頓" in both Japanese and Chinese denote "sudden", while "智" means "intelligence, insight or intuition". This highlights the connection between the Oogiri game and the requirement for strong associative abilities in LoT. Due to the fact that this creative game aligns with the input-output paradigm of current MLLMs and, because of its popularity, offers a wealth of high-quality, human-annotated creative responses, as well as rich scoring annotations for different responses, such as the number of likes, it makes an ideal platform for exploring the LoT ability of MLLMs.

3.1 Oogiri-GO Dataset

In this section, we collect Oogiri game data to build a large-scale Oogiri-GO dataset which serves as the sample of benchmarks to explore the LoT ability.

Specifically, Oogiri-GO is a multimodal and multilingual humor dataset, and contains more than 130,000 Oogiri samples in English, Chinese, and Japanese. Notably, in Oogiri-GO, 77.95% of samples are annotated with human preferences, namely the number of likes, indicating the popularity of a response. As illustrated in Fig. 1 (b), Oogiri-GO contains three types of Oogiri games according to the input that can be images, text, or both, and are respectively called "Text to Text" (T2T), "Image to Text" (I2T), and "Image & Text to Text " (IT2T) for brevity. Table 1 summarizes the distribution of these game types. For training purposes, 95% of the samples are randomly selected to construct the training dataset, while the remaining 5% form the test dataset for validation in standard evaluation and analysis.

To create the Oogiri-GO dataset, there are three main steps, including online data collection, machine filtering by LLM, and manual screening. Firstly, to collect sufficient data, we source Oogiri game data from the official Oogiri game platform, Bokete, and other popular platforms, such as Twitter and Weibo which also host some Oogiri-game-alike data. Then, to guard against the inclusion of bias, violence, explicit content, offensive language, etc., we have placed a strong emphasis on rigorous safety checks during both machine and manual screening. We first use the multimodal LLM Qwen-VL [45] to do the initial screening of the raw data by constructing safety-checking prompts. Then, manual checking is performed on the remaining data. See more details about the dataset creation in appendix of the conference version [1].

4 Standard Evaluation with Oogiri game

Inspired by the humor benchmarks in [71] and other standard LLM evaluations [1], [31], [32], [33], we first develop a standard evaluation, i.e., choice and ranking questions, as shown in Fig. 2 (Left), and then quantitatively evaluate the LoT ability of LLMs on the Oogiri-GO test dataset. For the *choice questions*, mTn for short, they need LLMs to choose n "leap-of-thought" humor responses from m options given the input. Here we build four types of mTn questions, including 2T1, 3T1, 4T1, and 5T2. 2T1 means two options, the ground-truth response (GTR) and an image caption generated by BLIP2 [72]. 3T1 adds unrelated answers, e.g., other image captions. 4T1 further adds the GTR rewrite by Qwen-14B [2]. 5T2 has an extra GTR. For these questions, their difficulty increases progressively, and is diverse to ensure comprehensive evaluation. For choice questions, we use accuracy as the evaluation metric. Additionally, for the questions in test set whose responses have ground-truth human preference, e.g., the number of likes, we develop the *ranking questions* that always rank five candidates. For evaluation, we adopt the top-1 accuracy and the widely used ranking metric,i.e., Normalized Discounted Cumulative Gain (NDCG) [73], [74]. See more experimental details in the Appendix of conference version [1].

5 LoTbench with Oogiri game

As mentioned in Section 1, most existing standard evaluations of LLMs [10], [12], [14], [31], [32], [33], including the evaluation presented in Section 4, are largely based on objective questions such as selection and ranking. These paradigms have significantly contributed to estimating LLM performance and have provided quantitative results, and

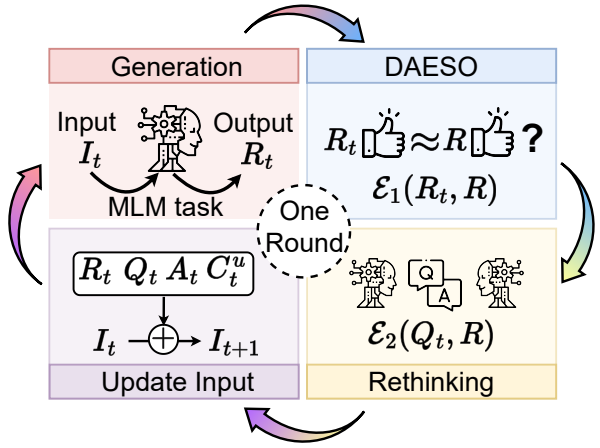

Fig. 3. The overview of proposed interactive creativity evaluation LoTbench for LLM. The main task in LoTbench is masked language modeling (MLM) task. DAESO denotes "different approach but equally satisfactory outcome", where $\mathcal{E}_1$ is the causal evaluator for check whether R_t and R are DAESO.

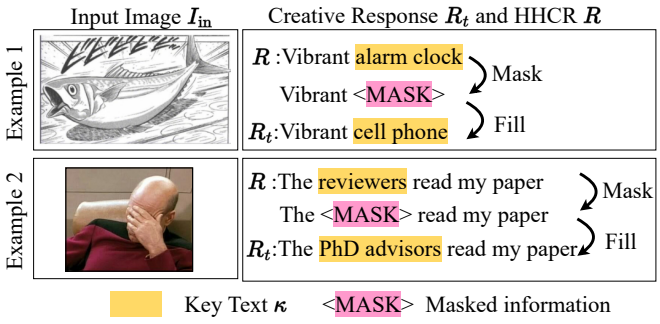

Fig. 4. The main task in LoTbench is masked language modeling (MLM). The LLMs are required to fill in the <MASK>in the sentence to make it a creative response relative to the provided image.

offer simplicity and low assessment costs. However, **for creativity**, these types of evaluations present certain risks, e.g., limited interpretability and information leakage.

(1) **Limited interpretability**. Standard evaluation typically adopts the selection and ranking problem format shown in Fig. 2 (Left). Since the LLM directly outputs answers to the questions, we have no insight into the reasoning process behind its "creative" decisions. In the experiments in Section 6, we also show that many advanced LLMs may exhibit lower performance in standard evaluation scenarios, indicating that there may be some biases when LLMs undergo such tests. However, due to limited interpretability, we cannot trace the reasons for this.

(2) **Information leakage**. First, A large amount of information available on the internet, including the Oogiri game, may already have been learned by existing LLMs.

Moreover, there is the issue of test prompt leakage. For example, in selection evaluations, the correct answer is often easily revealed among the options, which may lead to a less comprehensive assessment of creativity since the evaluation can inadvertently test recognition and logic abilities. For instance, consider the IT2T example in Fig. 1 (b). For the two options, "Alarm clock" and "Fish" , the tester can make a judgment **almost without creativity**, as the former is more unique and interesting. This type of evaluation paradigm does not usually present a problem for non-creativity evaluations, such as for a mathematical reasoning question like "5 + (6 * 4 + 3) = ?", where the options include "32" and "36". The tester must engage in rigorous reasoning to arrive at the correct answer. A truly reasonable creativity evaluation should assess the "measure the creativity level of LLM" rather than "recognize the creativity from LLM."

Algorithm 1 The details of LoTbench

Input: Given LLM $\mathcal{A}$ to be tested with a question prompt Q and a generation prompt G, along with an independent evaluator $\mathcal{E} = [\mathcal{E}_1, \mathcal{E}_2]$. A input $I_0 = [I_{\text{in}}, \mathcal{C}]$, where I_{in} and $\mathcal{C}$ are input image and its caption. A corresponding HHCR R. Maximum round N. The set of number of round $\mathbf{r}$ and the number of repeated times m. The Clue set $C_l = \{C_l^t\}_{t=1}^N$.
Output: Creativity score S_c.

While $m > 0$ **do**
  for t from 0 to N **do**
    Generate response $R_t \leftarrow \mathcal{A}(I_t|G)$ by Sec. 5.4
    Measure $\mathcal{E}_1(R_t, R)$ by Sec. 5.5
    if causal evaluator $\mathcal{E}_1(R_t, R)$ is True **do** break
    if causal evaluator $\mathcal{E}_1(R_t, R)$ is not True **do**
      Ask a question $Q_t \leftarrow \mathcal{A}(I_t, R_t|Q)$
      Get answer $A_t \leftarrow \mathcal{E}_2(Q_t, R)$ by Sec. 5.6
    Add R_t, Q_t, A_t and C_l^t into I_t by Sec. 5.4
  end for
  $m \leftarrow m - 1$ and add t into $\mathbf{r}$
end for
return Creativity score S_c with $\mathbf{r}$ by Sec. 5.7

Facing the issues mentioned above, in this section, we explore the creativity of LLMs from a novel perspective: **the cost required for LLMs to achieve high-quality human-level creative responses (HHCRs)**. As illustrated in Fig. 2 (Right), we frame this process as an interactive one. Under certain questions, the LLM generates creative responses over multiple rounds and evaluates, using causal inference techniques, whether these responses achieve a different approach but equally satisfactory outcome (DAESO) compared to HHCRs. The fewer rounds required to reach HHCRs, the more creative the LLM is deemed to be, and vice versa. It is worth noting that since these responses are generated by the LLM itself, ensuring the novelty of the test data can mitigate the problem of information leakage commonly encountered in standard evaluations. Furthermore, this interactive process can effectively visualize the LLM's innovative thinking process, offering a degree of interpretability. In Sec. 5.1, we formalize the LoTbench framework. Next, we introduce the CLoTv2 method in Sec. 5.2, which fine-tunes LLMs to help generate HHCRs in a specific format. In Sec. 5.3, we detail the construction of HHCRs. Finally, from Sec. 5.5 to Sec. 5.7, we present the other components of LoTbench.

5.1 The formulation of LoTbench

For brevity, we only consider constructing LoTbench with the Chinese and English data in Oogiri game. Inspired by situation puzzles [36], [75], given a input image I_{in} with its caption $\mathcal{C}$, we ask the LLM to provide an creative response R_t by a masked language modeling (MLM) task as shown in Fig. 4 . And in each round of interaction, we determine whether it reaches the creativity level of HHCR R by causal

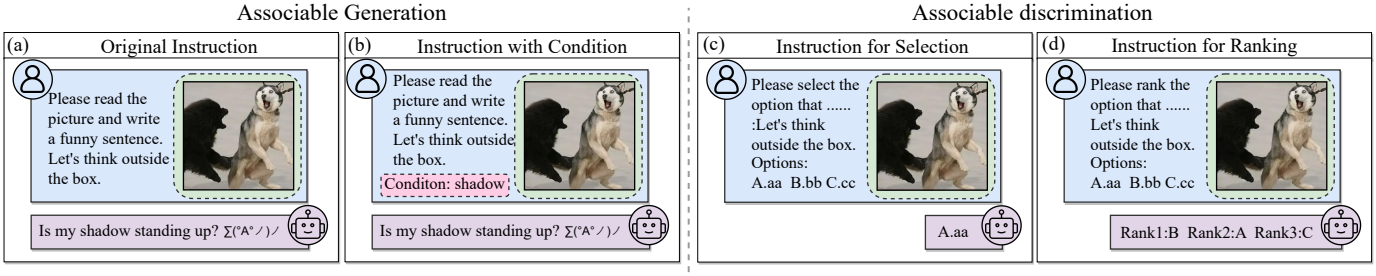

Fig. 5. The details of LoT-oriented instructions templates. We take "Image to Text" as an example, see the Appendix of the conference version [1] for the details of other categories' instructions. (a) and (b) are the instruction templates with/without conditions for associable generation. (c) and (d) are the two instructions about the selection and ranking of associable discrimination. All templates follow the formats in Fig. 6.

```
USER-INPUTs: Task-specific Prompt
   OPTIONs: <Image> <Condition>
ASSISTANT: Task-specific Responses
```

Fig. 6. The LoT-oriented instruction templates.

evaluator $\mathcal{E}_1$. Intuitively, fewer rounds imply statistically higher creativity for the LLM. Throughout this process, the LLM can continuously ask questions about R in each round, and the system $\mathcal{E}_2$ will respond with Yes/No. This rethinking of spontaneous questioning is also a manifestation of its own creativity [75], [76]. To ensure the test can end within a limited number of rounds, we provide the LLM with clues at regular intervals to control its thinking space. The specific algorithm process is shown in Fig. 3 and Alg. 1.

5.2 Tuning LLM for Data Synthesis by CLoTv2

In this section, we tune the LLM through two steps: associable instruction tuning and explorative self-refinement, as shown in Fig. 7, to acquire the ability to generate specific HHCRs in preparation for the test data synthesis of LoT-bench in Section 5.3.

5.2.1 Associable Instruction Tuning

LoT ability mainly includes associable generation and discrimination ability [77]. Given an input, associable generation draws its parallels with seemingly unrelated concepts via remote association and then generates innovative responses, e.g., the unexpected humor for the Oogiri input. Associable discrimination is to judge the matchiness among input and responses though they are seemingly unrelated, and then to select the most creative response.

Unfortunately, both associable generation and discrimination are not present in current LLMs, e.g., not good performance of GPT4o [78] in the Oogiri game observed in Sec. 6. Moreover, it is hard to improve these two LoT abilities via popular CoT-like prompt techniques. Indeed, as shown in Sec. 6, CoT even sometimes impairs the LoT performance of the LLMs like Qwen-VL [45] in the Oogiri game. To address this issue, we propose associable instruction tuning which trains LoRA [79] for LLMs on the Oogiri-GO dataset to achieve certain associable generation and discrimination abilities. It has two steps, including instruction generation and discrimination template design, and associable instruction learning.

(1) Instruction Generation & Discrimination Templates. We design LoT-oriented instruction templates to transform the Oogiri-GO dataset into instruction tuning data, and then train LLM to achieve associable generation and discrimination abilities. Our templates primarily comprise two components in Fig. 6: task-specific prompt and response. For different abilities, the templates need some special design.

For associable generation, "USER-INPUTs" contains "Task-specific Prompt" along with two optional conditions, "Image" and "Condition". For "Task-specific Prompt", we elaborately design several templates for different types of Oogiri game. See the Appendix of the conference version [1] for details and there is an image-2-text (I2T) Oogiri example in Fig. 5. For "Image" condition, it relies on the type of Oogiri game, e.g., being the image embeddings in I2T game and empty in T2T type. For the "condition" option, it's set to empty with a probability of ρ_c, and otherwise is randomly set as one word (including noun, verb, adjective or adverb) in "task-specific responses". This design gives the LLM a clue to connect the game input and the correct responses while also encouraging LLM to explore and unleash its creative thinking with probability ρ_c. Finally, "Task-specific Responses" are the ground truth responses of an Oogiri-GO data, and need to be predicted by LLM during training. This task enforces the LLM to draw parallels between seemingly unrelated concepts in inputs and responses for giving innovative responses, e.g., the humor for the Oogiri input. This associable generation ability can assist the LLM to think outside the box and learn remote association thinking.

Regarding associable discrimination, we aim to develop fundamental LoT discrimination skills for LLM. Based on the Oogiri-GO data, we design choice questions to enhance LLM's LoT discrimination ability, i.e., **selection** skill. Besides, as 77.95% of the Oogiri-GO data have human preference annotations, i.e., the number of likes of several responses (see Sec. 3.1), we design ranking questions to improve another discrimination skill, i,e., **ranking** ability.

For a choice question, as shown in Fig. 5 (c), the options in "Task-specific Prompt" contain the random permutations of ground truth response (GTR), image captions generated by BLIP2 [72], GTR from other images, rewrites of GTR by Qwen-14B [2]. See details in Appendix of conference version [1]. For "task-specific responses", it is the GTR. This design is to train LLM to improve its LoT selection ability. For a ranking question, as shown in Fig. 5 (d), it is to enforce LLM to rank multiple distinct responses of a given input to match

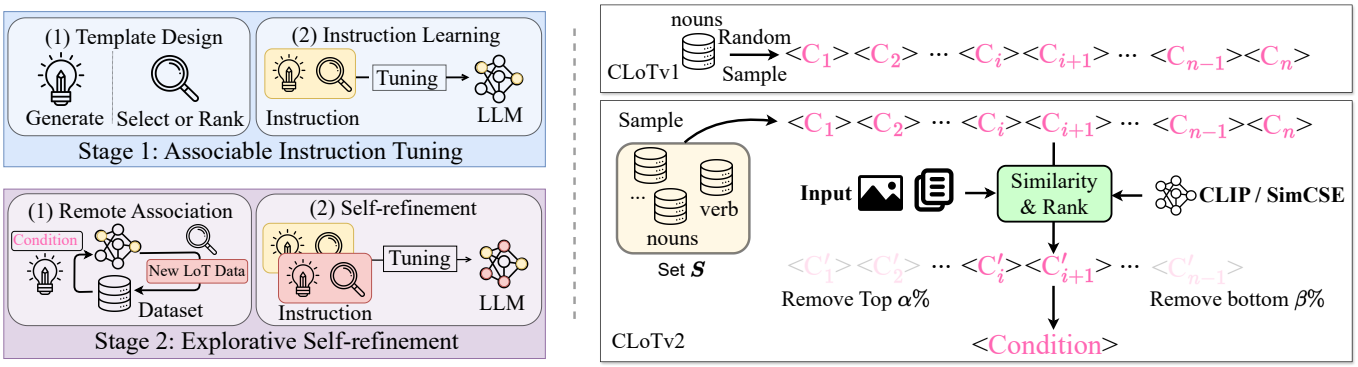

Fig. 7. The overview of CLoTv2 to tune LLM for data synthesis. Left: CLoTv2 relies on two LoT-boosting stages, including associable instruction tuning and explorative self-refinement. Right: the condition sampling method in conference version CLoTv1 [1] and CLoTv2.

their human preferences. By training on the choice and ranking questions, LLM is encouraged to distinguish LoT responses and align human creative preferences, improving its LoT discriminative selection and ranking abilities.

(2) Associable Instruction Learning. By using the above instruction templates, we augment the 130,000 samples in the Oogiri-GO dataset to more than 600k instructions whose formulation is in Fig. 6. During training, LLM is required to predict the "task-specific responses" according to the "USER-INPUTs" which include "Task-specific Prompt" and two additional optional conditions like image and text condition. To avoid over-fitting, we only train standard LoRA [79] for the LLM with the associable instruction data. See more details in Appendix of the conference version [1].

5.2.2 *Explorative Self-Refinement*

After associable instruction tuning, we aim to generate more HHCRs which are then used to train LLM for self-refinement. To this end, we introduce an innovative stage called explorative self-refinement, inspired by human LoT exercise process of "remote association & self-refinement", also known as mental leap [22], [25], [77]. The remote association process refers to generating new ideas by associating remote concepts or thoughts, and self-refinement uses the generated data to enhance one's own LoT ability. In the following, we design two similar LoT exercise processes for LLM to improve its LoT ability.

(1) Explorative Remote Association. The core here is to prompt the LLM to generate a diverse array of creative responses under weakly-associated conditions.

To implement this, as shown in Fig. 7 (Right), we first extract a set of keywords, including nouns, verbs, adjectives, and adverbs, denoted as $\boldsymbol{\mathcal{S}}$, from the text in the Oogiri-GO training data. Then, for given image and text in each sample, we construct a series of effective weakly-associated conditions as follows. We sample n candidate conditions $\{\boldsymbol{C}_i\}_{i=1}^{n}$ from $\boldsymbol{\mathcal{S}}$ with equal probability, then use CLIP or SimCSE [80], [81] to batch compute the similarity between these candidates and the image or text in samples. The similarities of I2T and IT2T are determined by their CLIP scores based on their images and keywords, while SimCSE is used to calculate the similarity for T2T. The candidates are ranked by similarity in descending order to get $\{\boldsymbol{C}'_i\}_{i=1}^{n}$. We remove the top $\alpha\%$ and bottom $\beta\%$ of elements from $\{\boldsymbol{C}'_i\}_{i=1}^{n}$, and add an empty condition ϕ to obtain the final weakly-associated conditions:

$$\boldsymbol{C}_W = \{\boldsymbol{C}'_i\}_{i=\lfloor \alpha\%n \rfloor}^{\lfloor (1-\beta\%)n \rfloor+1} \cup \{\phi\}. \tag{1}$$

Next, we add each condition from $\boldsymbol{C}_W$ into the instruction shown in Fig. 6 and feed them into the LLM to generate a humor candidate. We mix each generated humor candidate with its corresponding ground truth responses (GTR), and select the top-1 as the final response using the selection ability learned in Sec. 5.2.1. Finally, if the selected top-1 response is the GTR, we discard this generated humor candidate. By repeating this process, we progressively gather sufficient new HHCRs.

The core of this approach is the selection of weakly-associated conditions, which can encourage the LLM to engage in remote associations. This is because the empty conditions allow LLM to operate freely, while the other conditions compel the LLM to draw connections between seemingly unrelated concepts. This mechanism facilitates the establishment of links between seemingly-unrelated and weakly-related concepts, encouraging the LLM to explore knowledge outside of traditional cognitive limitations. The exploration ability distinguishes our CLoTv2 from CoT which primarily guides the LLM to exploit its inherent reasoning ability without emphasizing knowledge exploration.

Unlike the conference version CLoTv1 [1], which approximates weakly-associated conditions by relying solely on random sampling, as shown in Fig. 7 (Right), CLoTv2 explicitly models the conditions sampling based on the similarity between the content of each sample and the candidate conditions. By using parameters α and β, it ensures that the conditions are sufficiently different from the content but not completely unrelated—achieving a truly "weakly-associated" state. This approach mitigates the issue in CLoTv1, where random sampling often leads to the generation of many irrelevant conditions, which typically fail to produce effective responses, resulting in considerable computational waste. In this paper, we set $n = 100$, $\alpha = 25$ and $\beta = 70$.

(2) Self-refinement. We combine the above generated instructions with vanilla instruction tuning samples in Sec. 5.2.1 to form a dataset with more than 660k samples to train our LLM again. Since the above generated data is

of high diversity because of its exploration strategy, they prevent performance collapse [82], [83] during this phase.

After the two LoT-boosting phases above, the LLM gains sufficient LoT ability and can assist us in synthesizing new HHCRs to construct the test data for LoTbench in Section 5.3, which can mitigate the issue of imformation leakage.

5.3 The Data Construction in LoTbench

Task type. The primary task in LoTbench is a masked language modeling (MLM) task, as illustrated in Fig.4. Unlike I2T tasks that directly generate a complete response, MLM is a variation of IT2T. This design choice is guided by two key considerations: (1) To ensure the task leverages LLM's core strengths. Otherwise, limitations in some specific capabilities might lead to mediocre performance, interfering with the assessment of LLM creativity. MLM is precisely the type of task where LLMs excel [84]; (2) To simplify evaluation complexity. Since creativity is inherently diverse, allowing LLMs to freely generate responses R_t as in I2T tasks would make it difficult to assess whether they match the creativity level of given HHCR R due to high variability. Therefore, some constraints on R_t are necessary, and MLM tasks naturally provide this by fixing certain textual content, making it a suitable choice.

Specifically, as shown in Fig. 4, we manually annotate the key text κ in each HHCR R, mask it, and ask the LLM to complete the response, aiming for creative and high-quality responses. The "key text" κ refers to some textual contents that most crucially link the image and response, making the responses creative. Removing these contents would strip the text-image combination of its creativity. For example, in Fig. 4 Example 1, "alarm clock" is the key text in R. Moreover, identifying such key text accurately is challenging for different automated tools, including LLMs, so in this paper, we manually annotate them one by one during the data construction process.

Data structure. Each sample in LoTbench consists of six parts: the input image I_{in} with its corresponding HHCR R, image caption $\mathcal{C}$, and key text κ. It also includes a detailed explanation E_{xp} of why each R is innovative, and a Clue set $C_l = \{C_l^t\}_{t=1}^N$ designed to help LoTbench complete the evaluation within a limited number of rounds. For instance, in the sample shown in Fig. 4 Example 1, the input image I_{in} is the one on the left, with R being "Vibrant alarm clock", $\mathcal{C}$ being "A freshly caught fish, still flopping on the table, made a loud noise", and κ being "alarm clock". The explanation E_{xp} is "The lively fish rapidly flopping on the table and making a lot of noise closely resembles the moment when an alarm clock goes off. In R, the visual association of imagining the fish's flopping as an alarm clock ringing is both surprising and intriguingly interesting." Additionally, we have structured the Clue set $C_l = \{C_l^t\}_{t=1}^N$ to include both substantive clues and empty clues. At regular intervals, such as every several rounds (set as 5 in our paper) as indicated in line eight of Alg. 1, a substantive clue is added to the user-input. An example of a substantive clue is "It is a noun; It is a commonly used object at home; Pay attention to rapid jumping; Related to sound; Related to time."

Data volume.

Due to some unavoidable reasons, the data volume of LoTbench's test samples is limited: (1) There is a shortage of data suitable for constructing MLM tasks. On one hand, as mentioned in Section 1, creativity data itself is rare, and HHCR data is even scarcer due to the high-quality requirements. Additionally, even when some HHCR data is available, responses in the Oogiri game are typically very short, with entire or major portions often consisting of key text, making it challenging to create MLM tasks. (2) Fairness of the benchmark also needs consideration. Given the nature of the Oogiri game, many HHCR key texts are often culturally or knowledge-specific, so we must filter out these examples to ensure fair evaluation across most LLMs.

For these intrinsic limitations above in creativity data, we carefully and manually curated 106 HHCR samples suited for LoTbench, with Oogiri-GO and the help of CLoTv2 trained in Section 5.2 to generate brand-new HHCRs that meet MLM requirements. While LoTbench contains fewer samples than typical LLM evaluations [8], [9], [10], [11], [12], [13], [14], it boasts extremely high quality. Prior work [85], [86] has emphasized that test data quality is far more important than quantity, noting that the sample in many well-known benchmarks contain severe redundancy and a small number of test cases—less than 1%—can also yield evaluation results comparable to a full dataset. The consistency with human cognitive theories [38], [39], [40], [41], [42] demonstrated in LoTbench's evaluation results, shown in Section 6.2, also indicates that the current data construction is sufficient for assessing LLM creativity to a certain extent.

5.4 The Details of I_t and R_t

In Alg. 1, I_t initially includes only the input image I_{in} and its corresponding image caption $\mathcal{C}$. The generation prompt G contains the complete instruction, including the system prompt, example prompt for in-context learning, and task-specific prompts. Additionally, G also includes R with the key text κ masked. In each round of evaluation, the LLM under test is required to creatively fill in the masked key text in R through the given I_t and prompt G. As the LoTbench interactive evaluation progresses, I_t will continuously incorporate the current round's generated response R_t, the obtained clue C_l^t, and the question and answer Q_t and A_t. In the next round, this historical information will help the LLM to further produce a creative response. See supplementary for all prompts and other details.

5.5 To Measure DAESO with Causal Evaluator $\mathcal{E}_1$?

5.5.1 Criteria

In Alg. 1, we need an evaluator $\mathcal{E}_1$ to determine whether R_t and a given HHCR R exhibit a similar level of creativity. On one hand, since creativity is diverse, R_t and R are unlikely to be identical at the character level, so $\mathcal{E}_1$ cannot assess them through string matching. On the other hand, we also cannot rely solely on semantic similarity, as is common in natural language processing [81]. For example, as illustrated in Fig. 4 Example 1, if R is "vibrant alarm clock" and an LLM outputs R_t as "vibrant cell phone," we may still consider R_t to have a similar level of creativity to R despite the semantic distance between "cell phone" and "alarm clock". Through the analysis above, in this paper, the $\mathcal{E}_1(R_t, R)$ is set to assess whether R_t and R are a "different approach but

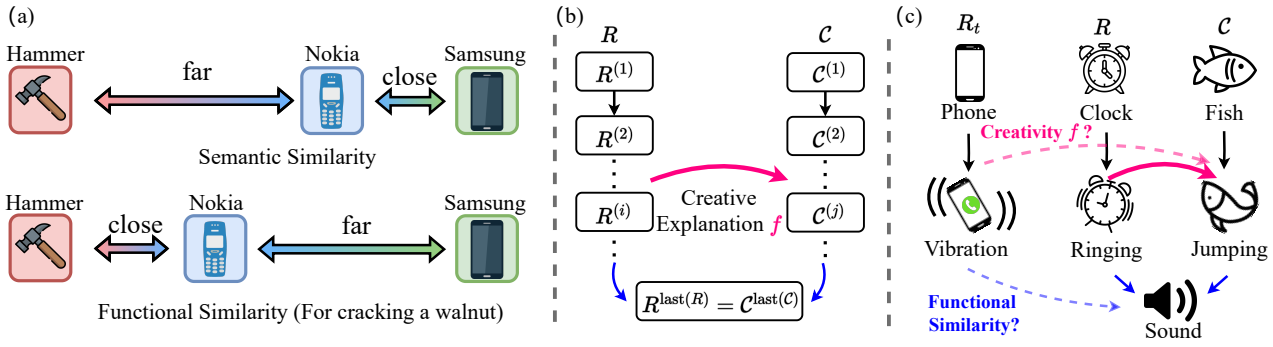

Fig. 8. The overview of DAESO. (a) The difference between semantic similarity and functional similarity. (b) The mathematical modeling for DAESO by causal chains of given HHCR R and corresponding image caption $\mathcal{C}$. (c) Two criteria of DAESO.

equally satisfactory outcome" (DAESO). To achieve this, we propose two criteria for DAESO: (1) R_t and R share the same creative point; (2) R_t and R are functionally similar rather than semantically similar.

For criterion (1), if R_t is "vibrant drum," even though a drum can also make sound, it lacks the vivid, jump-out image of an "alarm clock" in the given context, thus differing in innovation point. For criterion (2), The DAESO between "cell phone" and "alarm clock" should be judged functionally rather than semantically. As shown in Fig. 8 (a), if we compare "Nokia", "Samsung" and "Hammer", "Nokia" and "Samsung" would seem closer semantically, as both are well-known electronics companies with popular phone products. However, in a given scenario like "cracking a walnut", "Nokia" and "Hammer" are more similar functionally, as "Nokia" products are famously durable and can serve to crack walnuts, while "Samsung" devices are more fragile and thus less suitable for the task.

Based on these criteria, we propose a novel evaluation mechanism for $\mathcal{E}_1$ in following Section 5.5.2, which assesses DAESO by analyzing the causal chain in the LLM responses to judge whether R_t and R align in terms of DAESO.

5.5.2 Modeling

Causal Construction. For a given sample, we can first leverage the image caption $\mathcal{C}$ and carefully annotated explanation E_{xp} of "Why R is creative," as mentioned in Section 5.3, to model the causal chain for R and $\mathcal{C}$. R and $\mathcal{C}$ are expanded in the form of Eq. (2) as follows:

$$\begin{aligned} \mathcal{C} &\simeq \left(\mathcal{C}^{(1)}, \mathcal{C}^{(2)} \cdots, \mathcal{C}^{(\text{last}(\mathcal{C}))}\right) \\ R &\simeq \left(R^{(1)}, R^{(2)}, \cdots, R^{(\text{last}(R))}\right), \end{aligned} \tag{2}$$

where $\mathcal{C}^{(i)}$ and $R^{(j)}$ represent individual nodes within $\mathcal{C}$ and R, respectively, with Fig. 8 (b) illustrating a diagram. $\text{last}(\mathcal{C})$ and $\text{last}(R)$ denote the number of nodes in chain of $\mathcal{C}$ and R, respectively, and they may not be equal. First, for criterion (1), there should be a creative explanation f and $i \leq \text{last}(R)$, $j \leq \text{last}(\mathcal{C})$, ensuring that:

$$f(R^{(i)}) \to \mathcal{C}^{(j)}, \tag{3}$$

i.e., there exists a function f such that one node in R can be mapped to another node in R, and this mapping f is the reason why R is considered creative. Fig. 8 (c) provides a specific analysis for Fig. 4 Example1 where the "alarm clock" appears creative and visually engaging because its bouncing motion when ringing corresponds to the bouncing of a "fish", thus creating a vivid and creative image. Furthermore, for criterion (2), we can consider the final nodes in $\mathcal{C}$ and R to be the same, that is,

$$\mathcal{C}^{\text{last}(\mathcal{C})} = R^{\text{last}(R)}, \tag{4}$$

to represent that the functions of the two causal chains are similar. For example, in the case shown in Fig. 8 (c), whether it is "clock ringing" or "fish jumping", they all functionally serve to "produce sound."

Causal Intervention. After completing the aforementioned modeling, next, given a response R_t generated by an LLM, we begin to analyze whether R_t and R are DAESO. Since R_t does not have finely annotated explanations E_{xp} like R, it's not easy to establish a causal chain as in Eq. (2). However, note that according to Section 5.3 and Fig. 4 illustrating the MLM task type, R_t and R differ only in the key text κ part. Therefore, we consider intervening in the causal chain of R in Eq. (2) with intervention $do(\kappa(R) \to \kappa(R_t))$ to approximate the chain that produces R_t, denoted as

$$\left(R_t^{(1)}, R_t^{(2)}, \cdots, R_t^{(\text{last}(R_t))}\right), \tag{5}$$

where $\text{last}(R_t)$ is the number of its nodes. According to criterion (1), if R_t and R are DAESO, for Eq. (3), there also exists an $i' \leq \text{last}(R_t)$ such that

$$f(R_t^{(i')}) \to \mathcal{C}^{(j)}. \tag{6}$$

This implies that $R_t^{(i')}$ and $R^{(i)}$ both map to the same node $\mathcal{C}^{(j)}$ through the creative interpretation f, meaning that R_t and R share the same point of creativity. Next, from criterion (2), if R_t and R are DAESO, we have

$$\Delta P\{R^{(\text{last}(R))} | do(\kappa(R) \to \kappa(R_t))\} \to 0, \tag{7}$$

Where ΔP denotes the change in probability. That is, after the causal chain of R is modified by replacing $\kappa(R)$ with $\kappa(R_t)$ and restructured, the last node $R^{\text{last}(R)}$ that determines the functionality of these chains changes very little, i.e., $R^{\text{last}(R)} = R_t^{\text{last}(R_t)}$.

5.5.3 Measuring DAESO by $\mathcal{E}_1$

In this section, we provide the details of evaluator $\mathcal{E}_1$ by the modeling in Section 5.5.2. According to Section 5.5.2, there are three steps to determine whether R_t and R are DAESO:

TABLE 2
The accuracy (%) of choice questions and the NDCG (%) of ranking questions on **mutilmodal multilingual models**. mTn choice question selects n correct answers from m options. "Avg." is the average of all metrics. "AIT" and "AITv2" denotes the the LLM with only associable instruction tuning of CLoTv1 and CLoTv2, respectively. The best results for each backbone are highlighted in bold and the second-best results are emphasized with some underlines.

Model	Size	Image&Text to Text (IT2T)					Image to Text (I2T)					Text to Text (T2T)				
		3T1	4T1	5T2	Rank	Avg.	3T1	4T1	5T2	Rank	Avg.	3T1	4T1	5T2	Rank	Avg.
GPT4v [78]	-	19.3	14.9	3.2	56.7	23.5	29.1	15.1	3.9	60.4	27.1	27.1	16.8	6.8	53.5	26.1
LLaVA-1.5 [43]	13B	13.2	13.7	13.9	68.1	27.2	29.3	22.7	3.9	60.9	29.2	33.8	25.2	4.0	62.6	31.4
MiniGPT-v2 [44]	7B	6.1	3.4	4.0	60.7	18.6	5.3	4.0	3.8	60.5	18.4	10.8	7.3	3.5	59.4	20.3
mPLUG-Owl$_{\text{Multilingual}}$ [87]	7B	28.1	26.0	10.5	64.4	32.2	19.2	18.6	6.0	60.5	26.1	24.4	22.2	10.7	60.1	29.4
VisualGLM-6B [88]	6B	24.1	22.5	9.7	67.4	30.9	14.3	20.4	8.8	61.9	26.4	13.1	20.2	7.1	61.3	25.4
GPT-4o [78]	-	26.5	20.1	8.9	60.7	29.1	22.6	18.6	11.9	60.4	28.4	30.6	24.2	11.5	59.4	31.4
GPT-4o mini [78]	-	19.8	24.6	14.4	67.4	31.6	25.6	22.1	8.8	61.2	29.4	32.6	25.8	13.9	61.8	33.5
Claude 3.5 Sonnet	-	20.8	15.9	9.7	64.4	27.7	19.2	18.6	10.6	60.5	27.2	20.6	25.2	8.6	60.6	28.8
Gemini 1.5 Pro [89]	-	18.6	20.4	10.6	66.1	28.9	20.5	16.6	6.6	60.4	26.0	26.4	15.6	5.6	62.6	27.6
Intern-VL2 [90]	40B	8.2	11.6	3.2	60.7	20.9	14.3	8.8	3.8	61.4	22.1	20.8	16.8	7.1	59.4	26.0
miniCPM-V [91]	8B	20.6	18.6	10.6	64.4	28.6	19.2	18.6	8.6	59.8	26.6	30.6	28.6	10.7	61.3	32.8
Yi-VL [92]	34B	19.3	16.6	6.8	59.6	25.6	20.4	14.8	6.6	59.4	25.3	16.5	10.5	6.8	55.6	22.4
Qwen2-VL [93]	72B	28.6	20.4	8.8	66.6	31.1	24.6	20.5	10.2	60.6	29.0	16.8	25.2	6.8	61.3	27.5
Qwen-VL [45]	7B	30.2	26.0	10.4	67.7	33.6	23.2	23.1	11.9	62.2	30.1	23.4	25.0	13.3	59.6	30.3
Qwen-VL$_{+\text{AITv1}}$ [1]	7B	39.7	38.9	15.7	67.3	40.4$_{+ 6.8}$	38.8	30.5	15.7	62.3	36.8$_{+ 6.7}$	30.6	28.7	16.7	62.6	34.6$_{+ 4.3}$
Qwen-VL$_{+\text{CLoTv1}}$ [1]	7B	41.8	38.7	21.6	68.5	42.7$_{+ 9.1}$	39.8	35.1	22.7	64.4	40.5$_{+10.4}$	38.8	29.4	21.0	64.7	38.5$_{+ 8.2}$
Qwen-VL$_{+\text{AITv2}}$ (ours)	7B	40.1	39.1	17.0	67.9	41.0$_{+ 7.4}$	39.2	32.5	17.2	61.8	37.7$_{+ 7.6}$	32.6	28.5	18.3	63.1	35.6$_{+ 5.3}$
Qwen-VL$_{+\text{CLoTv2}}$ (ours)	7B	**42.2**	**39.2**	**22.3**	**69.0**	**43.2**$_{+ 9.6}$	**40.4**	**37.1**	**24.6**	**65.2**	**41.8**$_{+11.7}$	**39.2**	**29.8**	**23.1**	**65.3**	**39.4**$_{+ 9.1}$

(1) **Causal construction**. Establish the causal chain of R and caption $\mathcal{C}$ based on E_{xp} and $\mathcal{C}$; (2) **Causal Intervention**. Intervene in the causal chain of R using $do(\kappa(R) \rightarrow \kappa(R_t))$ to build the causal chain of R_t; (3) **Judgment DAESO**. Determine whether Eq. (6) and Eq. (7) hold.

Considering the complexity of E_{xp} and $\mathcal{C}$ across different samples, explicitly constructing the causal chain is highly challenging. If the chain is explicitly established, modeling the changes in chain nodes during key text interventions becomes difficult, making it hard to develop an effective algorithm to judge Eq. (6) and Eq. (7). To address this, in this paper, we propose using a powerful text-based LLM to map the causal chain between R and caption $\mathcal{C}$ into the text space based on E_{xp}, $\mathcal{C}$, and a specially constructed prompt. This generates a long text to describe these causal chains, effectively establishing them. Leveraging the fault-tolerance and reorganization capabilities of LLMs, we semantically replace the key text $\kappa(R)$ with $\kappa(R_t)$ to reorganize the chain, then describe and judge Eq. (6) and Eq. (7) in language form, ultimately determining whether R_t and R are DAESO. In Section 8, we will validate the effectiveness of the proposed method for judging DAESO through a series of experiments. For details about the specific prompts, please refer to the supplementary materials.

5.6 How to ask a Question Q_t and answer it by $\mathcal{E}_2$?

As mentioned in section 5.1, the rethinking about spontaneous questioning is also a manifestation of one's own creativity [75], [76], which can visualize the process of achieving creativity in LLMs. Given the current input I_t and an incorrect response R_t, LLMs utilize a question prompt Q, which includes a series of instructions related to questioning, such as system prompts, example prompts for in-context learning, and task-specific prompts. We require the LLM to propose a speculative question Q_t about R, such as "Is it related to daily life?" or "Is it a type of appliance?" and so on, to help itself generate more human-like creative responses in the next round. Subsequently, for Q_t, we consider using a textual independent LLM, denoted as $\mathcal{E}_2$, to directly output a binary judgment A_t containing Yes or No based on R. In Section 8, we find that selecting GPT-4o mini is suitable for this task. More details about the prompts are shown in the supplementary.

5.7 The Creativity Score S_c

As mentioned in section 5, LoTbench aims to explore how many rounds of creative thinking are required for LLMs to achieve HHCRs, with fewer rounds indicating statistically higher creativity. Therefore, we believe the creativity score S_c should meet at least two requirements for the set of the number of rounds $\mathbf{r} = [t_r^{(1)}, t_r^{(2)}..., t_r^{(m)}]$ with m times repeated evaluation: (1) As $\min(\mathbf{r}) \rightarrow \infty$, $S_c \rightarrow 0$, meaning that if an LLM has not reached the creativity level of HHCR after a sufficiently large number of rounds, its contribution to creativity in the current sample tends to zero; (2) S_c should be inversely proportional to the round, meaning that the faster the LLM reaches HHCR, the more creative it is considered to be. Thus, we propose the following formula to define S_c.

$$S_c = \frac{1}{mn} \sum_{j=1}^{m} \sum_{r=1}^{n} \beta_c \exp[-\alpha_c \cdot t_r^{(j)}], \tag{8}$$

where n represents the number of test samples in LoTbench, while β_c and α_c are hyperparameters, and set to 1.0 and 0.2 respectively in this paper. The m denotes the number of rounds for repeating independent tests on a single sample. The rationale behind conducting multiple experiments is to reduce the errors in evaluator judgments as shown in Section 8 and provide more opportunities for LLMs to engage in creative thinking, as creative responses are not always produced [1]. Additionally, considering the cost of inference, we set $m = 3$.

6 Experiments under Standard Evaluation

In this section, we explore the creativity of different LLMs through standard evaluation shown in Section 4, while testing the creative response generation capability of CLoTv2 proposed in Section 5.2. Noticing that we considered setting

TABLE 3
The accuracy (%) of choice questions and the NDCG (%) of ranking questions on various **mutilmodal non-multilingual models** (English). See notations in Table 2. We only consider I2T and T2T since English IT2T is not available due to cultural preference.

Model	Size	Image to Text (I2T)						Text to Text (T2T)					
		2T1	3T1	4T1	5T2	Rank	Avg.	2T1	3T1	4T1	5T2	Rank	Avg.
InstructionBLIP [94]	13B	19.8	13.7	15.5	1.1	65.5	23.1	22.3	16.0	17.0	0.7	59.5	23.1
mPLUG-Owl$_{\text{LLaMA2}}$ [87]	7B	22.3	12.7	15.0	4.2	59.9	22.8	24.2	13.7	12.6	3.1	59.2	22.6
Otter [95]	7B	15.8	9.9	8.5	7.1	61.3	20.5	3.8	3.3	4.8	5.4	58.5	15.1
CogVLM-17B [34]	7B	37.6	26.4	18.3	2.5	64.6	29.9	35.1	27.8	24.8	7.5	64.1	31.9
CogVLM-17B$_{+\text{AITv1}}$ [1]	7B	57.4	37.4	33.5	21.8	64.6	42.9$_{+13.1}$	55.4	46.5	26.4	18.2	64.4	42.2$_{+10.3}$
CogVLM-17B$_{+\text{CLoTv1}}$ [1]	7B	66.9	47.6	43.4	30.7	**69.4**	51.6$_{+21.7}$	64.8	52.9	33.6	21.8	68.6	48.3$_{+16.4}$
CogVLM-17B$_{+\text{AITv2}}$ (Ours)	7B	59.2	39.1	35.5	23.8	65.2	44.6$_{+14.7}$	57.1	47.1	27.1	19.3	65.1	43.1$_{+11.2}$
CogVLM-17B$_{+\text{CLoTv2}}$ (Ours)	7B	**68.4**	**49.8**	**46.4**	**32.5**	69.3	**53.3**$_{+23.4}$	**66.3**	**54.3**	**35.6**	**23.8**	**68.8**	**49.8**$_{+17.9}$

TABLE 4
The accuracy (%) of choice questions and the NDCG (%) of ranking questions on various **large language models**. Here we use English T2T task for test. See notations in Table 2.

Model	Size	3T1	4T1	5T2	Rank	Avg.
GPT-3.5 [78]	-	45.3	30.4	6.7	61.6	36.0
GPT-4 [78]	-	49.2	20.4	3.6	54.7	32.0
LLAMA2 [35]	7B	18.9	13.5	1.1	60.4	23.5
	13B	15.6	20.0	1.8	60.5	24.5
	70B	27.8	16.1	3.8	62.0	27.4
Baichuan2 [96]	7B	28.3	22.6	11.6	64.6	31.8
	13B	21.7	18.3	8.9	61.5	27.6
Qwen [2]	7B	23.1	20.4	8.0	61.4	28.2
	14B	27.4	22.2	12.3	59.5	30.3
ChatGLM3 [88]	6B	15.6	17.0	5.4	59.4	24.3
Vicuna-v1.5 [3]	7B	32.6	23.5	0.0	63.0	29.8
	13B	30.2	23.0	2.7	62.2	29.5
Qwen-VL$_{+\text{CLoTv1}}$ [1]	7B	51.7	32.3	24.8	65.0	43.4
CogVLM-17B$_{+\text{CLoTv1}}$ [1]	7B	52.9	33.6	21.8	68.6	44.2
Qwen-VL$_{+\text{CLoTv2}}$ (ours)	7B	**53.9**	33.2	**26.2**	65.1	44.6
CogVLM-17B$_{+\text{CLoTv2}}$ (ours)	7B	53.6	**34.8**	23.1	**68.8**	**45.1**

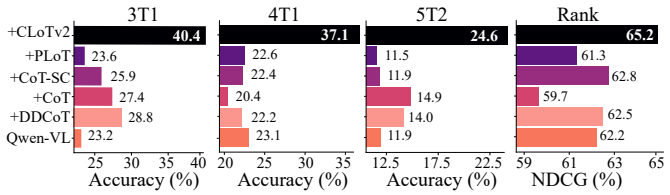

Fig. 9. The accuracy (%) of choice questions and the NDCG (%) of ranking questions on our CLoT and various reasoning frameworks. The baseline is Qwen-VL on multilingual I2T task. For mTn choice questions, one needs to select n correct answers from m options.

the condition in the instruction to empty ϕ in both the associable instruction tuning and Explorative Self-Refinement stages, this ensures that the LLM can generate creative responses without specific conditions, facilitating practical use of the model without the need to set conditions. Therefore, under the settings of CLoTv2, after training, the LLM can directly perform model inference through the instruction shown in Fig. 6.

6.1 Evaluation by Choice and Ranking Questions

Evaluation on Multimodal Multilingual LLMs. We plug our associable instruction tuning (AITv2) and our CLoTv2 into the advanced open-source multimodal multilingual model Qwen-VL [45] to obtain Qwen-VL$_{+\text{AITv2}}$ and Qwen-VL$_{+\text{CLoTv2}}$, respectively. Table 2 shows that, on three tasks (IT2T, I2T and T2T) which include English, Chinese and Japanese questions, Qwen-VL achieves the best LoT performance among all open-source baselines in most cases. In comparison, Qwen-VL$_{+\text{AITv2}}$ achieves a noticeable improvement on the advanced Qwen with average accuracy enhancements of 7.4%, 7.6%, and 5.3% on the three tasks, respectively. Importantly, Qwen-VL$_{+\text{CLoTv2}}$ further enhances Qwen-VL, showing improvements of 9.6%, 11.7%, and 9.1% in accuracy across these tasks. These results demonstrate the efficacy of the two stages in CLoTv2, i.e., associable instruction tuning and explorative self-refinement.

Evaluation on Multimodal Non-multilingual LLMs. Here we integrate our CLoTv2 with the advanced multimodal non-multilingual model, CogVLM-17B [34], and evaluate it on the English I2T and T2T tasks. Table 3 shows that CogVLM-17B$_{+\text{AITv2}}$ achieves remarkable improvements over the standard CogVLM-17B, and CogVLM-17B$_{+\text{CLoTv2}}$ consistently demonstrates significantly superior performance compared to CogVLM-17B.

Evaluation on Single-Modal LLMs. Now we test LLMs that can handle only pure texts, using the English T2T task for evaluation. Table 4 also indicates the insufficient LoT ability within existing LLMs, ranging from small to large models. Fortunately, our CLoTv2 significantly improves the LoT ability of these LLMs, as demonstrated by the notable improvement in accuracy.

Comparison with CoT-alike Reasoning Frameworks. We also find that existing reasoning frameworks are not as effective as CLoTv2 in enhancing LoT ability. Fig. 9 compares CLoTv2 with CoT [15], [17], CoT-SC [97], DDCoT [69], and prompted-based LoT (PLoT) with the prompt "let's think outside the box". The results reveal that CoT-alike frameworks do not enhance LoT performance of LLMs, while CLoT framework demonstrates the ability to consistently enhance LLMs.

Our experiments and analysis reveal that, unlike CoT-based methods, LoT cannot be directly achieved by prompting alone. This is because the inherent reasoning capabilities and extensive knowledge of LLMs are not sufficient to enable LoT ability. However, when trained with our proposed CLoTv2 method, LLMs can effectively engage in a range of creative tasks. Additionally, the use of specific prompting techniques can enhance the LoT ability of CLoTv2-trained LLMs. These findings suggest that LoT could potentially be considered an additional general reasoning ability for LLMs that is not contained in current LLMs or we may need to use more advanced methods to stimulate their LoT.

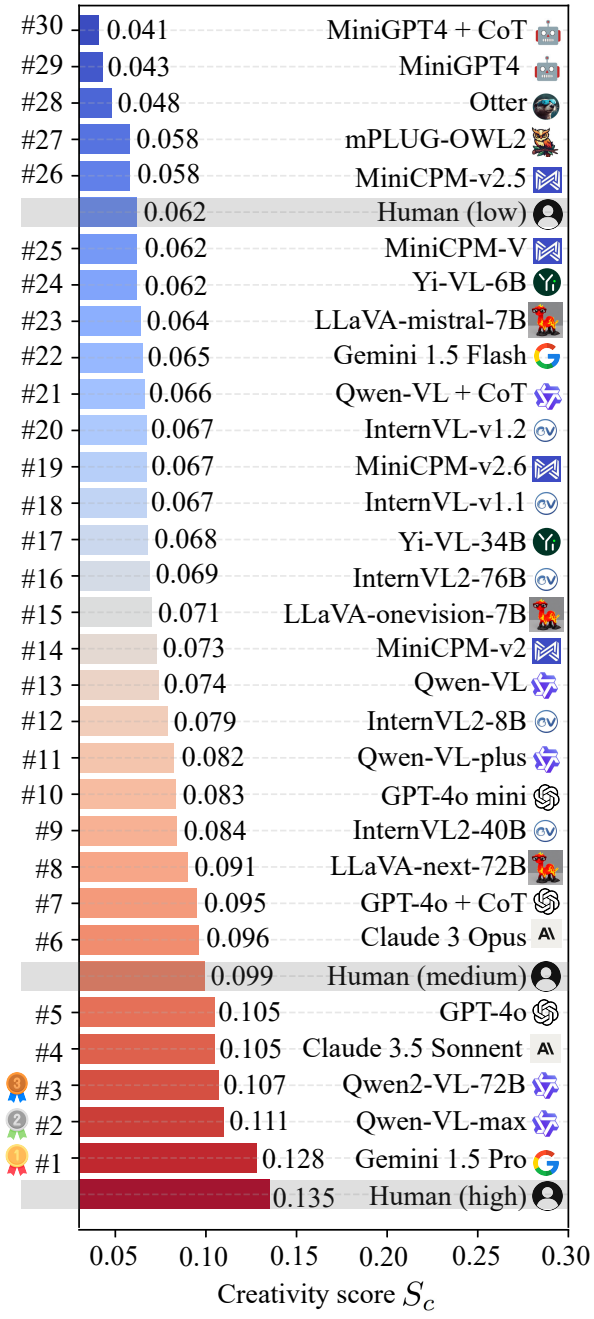

Fig. 10. The ranking results of LLM's creativity by LoTbench.

6.2 Experiments under LoTbench

In this section, we assess the creativity of various multimodal LLMs using LoTbench. To better understand the creativity score S_c, we introduced 21 human subjects aged between 13 and 44, testing only the samples in LoTbench for languages they are proficient in. Ultimately, we divided their S_c into three equal groups based on their S_c rankings, with each group containing 9 individuals, and calculated the average S_c for reference, naming them human (high), human (medium), and human (low).

From the results in Fig. 10, most LLMs do not exhibit high creativity in the LoTbench scenario, but the gap between their creativity and the average level of human participants is not particularly large. On one hand, generating creativity is inherently difficult for both humans and LLMs, which may be a primary reason for the current scarcity of creativity data. On the other hand, the testing process in LoTbench requires multiple rounds of interaction. We found that, after several interactions, human subjects of all levels often experience pauses, such as hesitations about how to respond. This represents a disadvantage for humans in long-term interactive evaluations. However, LLMs, which usually are based on next token prediction [1], [2], [3], [7], do not face this issue and can continuously generate responses. Therefore, LoTbench, designed specifically for LLMs, may not fully capture the average creativity level of humans during testing and should be considered as a reference. In this sense, the creativity levels of currently strong LLMs and human subjects are quite similar. Furthermore, since most LLMs generate responses based on next token prediction, if they can be sufficiently stimulated for creativity, they have the potential to produce a vast number of responses continuously, some of which may contain valuable and highly creative ideas. This could be an important direction for future scientific advancement. For all LLM in Fig. 10, the

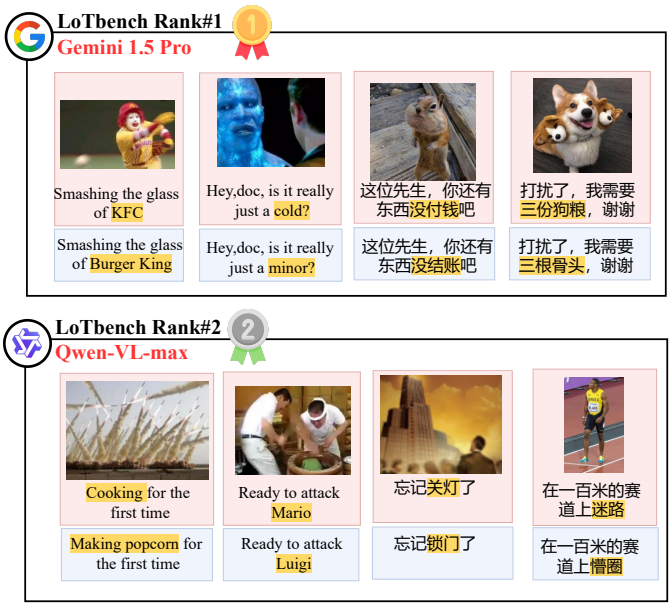

Fig. 11. Specific creative responses. We visualize the outputs of the two best-performing LLMs shown in Fig. 10. The red boxes indicate the original HHCRs, while the blue boxes are these LLMs' final outputs.

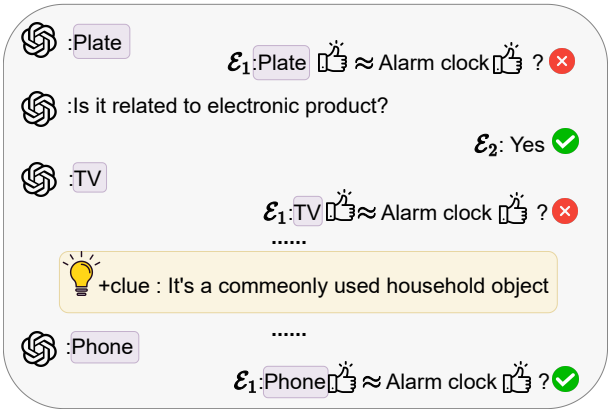

Fig. 12. Example of visualization for creative thinking in LoTbench.

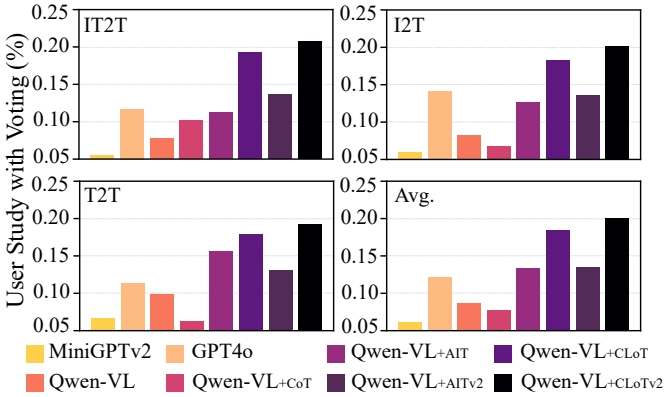

Fig. 13. User study with voting (%) for Oogiri-style creative responses by different models and improved methods. "Avg." denotes the avervage voting of three types of Oogiri game results.

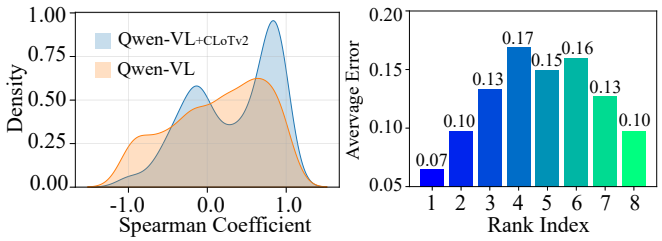

Fig. 14. The comparison between human preference and LLM-as-a-Judge on Oogiri game. (Left) Distribution of Spearman's rank correlation between different LLM and user study results. (Right) The ranking error of Qwen-VL$_{+\text{CLoTv2}}$ under different rank indices.

average number of rounds to complete the test is 13.65, with the best-performing Gemini 1.5 Pro achieving an average of 12.44 rounds. According to our setup in Section 5.3, the upper limit for the number of rounds per sample is 15, which intuitively reflects the relatively low creativity level of current LLMs. This limited capability results in an average forward inference cost that is over more than 10 times higher in the current LoTbench testing paradigm.. Next, in Fig. 11, we illustrate specific response examples for the first and second place winners in Fig. 10.

Moreover, in Fig. 12, we take Fig. 4 example 1 as a case study to demonstrate the entire process of GPT4o mini undergoing the LoTBench assessment. This process visualizes the entire thought process of GPT4o mini, providing an interpretable observation for humans to understand the creative generation process of LLMs.

7 Analysis

7.1 Ohter Types of Evaluation

In Section 5, we shown that the truly reasonable creativity evaluation should assess the "measure the creativity level of LLM" rather than "recognize the creativity from LLM." Actually, to evaluate the "measure the creativity level of LLM", there are also some previous methods like human evaluation and LLM-as-a-Judge [98], [99], [100], [101]. In this section, we consider these two kinds of evaluations for LLM's creativity, and show the necessity of LoTbench.

7.1.1 Human Evaluation for LLMs' Creativity

We conduct a user preference study to test creativity of LLMs. Here we select eight LLMs to generate responses for a total of twenty-one questions across three tasks (IT2T, I2T and T2T). We use choice questions, and ask users to choose the creative and humorous responses they think. Fig. 13 summarizes the statistical analysis of 56 valid surveys. The results indicate a strong user preference for the enhanced outputs from both the associable instruction tuning and explorative self-refinement stages across all three tasks, highlighting the effectiveness of our proposed method for synthesizing HHCRs. See more details in Appendix of the conference version [1]. The human evaluation provides the reasonable evaluation since it assesses whether the generated responses directly align with human creativity. However, it has clear drawbacks: it is unsustainable and requires additional manpower, leading to high costs when evaluating the creativity of new models. Additionally, for the new LLM under test, there might be fairness issues if the participants voting each time are different.

7.1.2 LLM-as-a-Judge for LLMs' Creativity

Moreover, following analysis through the Oogiri game, we find that directly using LLM-as-a-Judge might also struggle to accurately assess creativity of LLM.

Specifically, we consider the following settings to explore the relationship between LLM-as-a-Judge and human preferences in Fig. 13. Using the setup shown in Fig. 13, we randomly select 5 responses out of 8 generated for each sample and construct a ranking based on human preferences as the ground truth. Then, we have Qwen-VL and Qwen-VL$_{+\text{CLoTv2}}$ rank the 5 responses as well, comparing their rankings to the ground truth using the Spearman (SP) rank correlation coefficient [102]. A higher SP score indicates a ranking closer to the human preference; a lower score indicates less similarity. As shown in Fig. 14 (Left), we observe that while the original Qwen-VL has some ability to align with human-recognized creativity, it is nearly unusable in practice since a large number of SP values are concentrated around zero, and even in the negative region. In contrast, the enhanced Qwen-VL$_{+\text{CLoTv2}}$ significantly improves the LLM's judgment of responses. These observations are consistent with the standard evaluation results in Section 6.1. However, since SP scores for Qwen-VL$_{+\text{CLoTv2}}$ occasionally fall below 0.0, it indicates that this model is not fully reliable for judging or precisely scoring arbitrary responses.

Next, we further analyze Qwen-VL$_{+\text{CLoTv2}}$'s ranking accuracy by counting the ranking errors across different rank indices. Fig. 14 (Right) shows the error distribution across these indices. We observe that Qwen-VL$_{+\text{CLoTv2}}$ has fewer errors at the highest and lowest ranks, performing best at identifying the highest- and lowest-quality responses, while its performance is more ambiguous with mid-range quality responses. This suggests that although Qwen-VL$_{+\text{CLoTv2}}$ may struggle to give fine-grained scores or make fully accurate judgments for any sample, it remains viable as a data filtering tool as discussed in Section 5.2.

In summary, due to the limitations of different evaluation methods mentioned above, we propose LoTbench in this paper is necessary. Instead of directly scoring LLM responses, LoTbench estimates creativity by measuring the

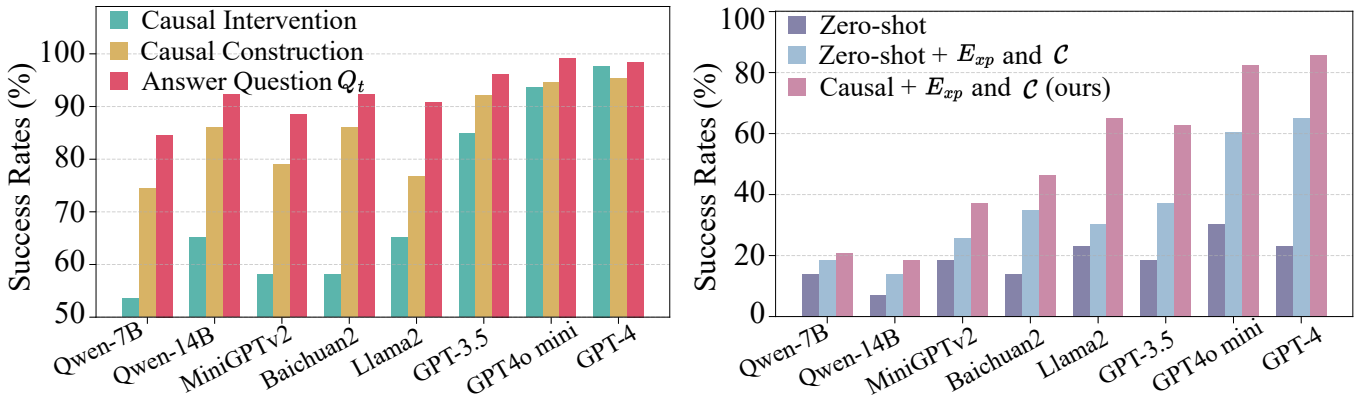

Fig. 15. Analysis of the discriminative abilities of $\mathcal{E}_1$ and $\mathcal{E}_2$. (Left) We examined the success rates (%) of various advanced LLMs in performing causal construction and causal intervention as described in Section 5.5.2. We also tested whether these LLMs could effectively engage the target LLM in interactive responses as outlined in Section 5.6. (Right) We analyzed the accuracy (%) in determining whether R_t and R are DAESO under different settings.

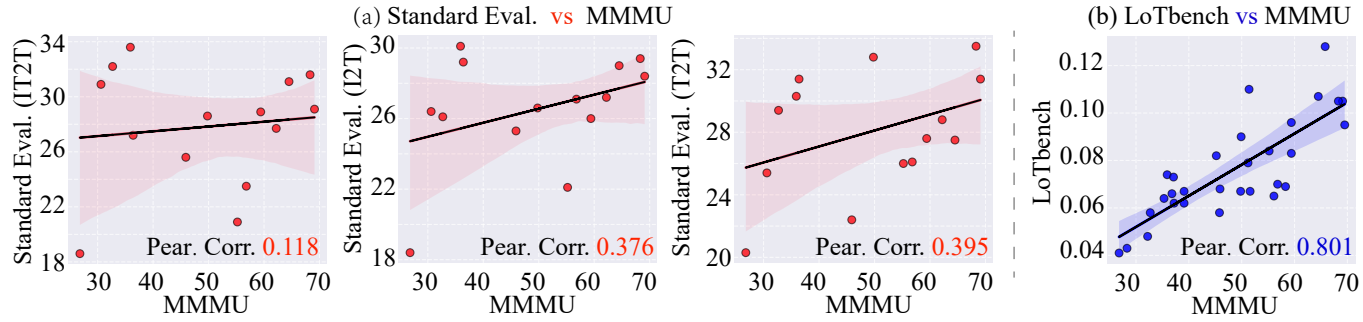

Fig. 16. The correlation analysis between LLM cognition benchmark MMMU and the proposed creatvity benchmark. (a) The correlation between MMMU and average results from different types of standard evaluation in Table 2. (b) The correlation between MMMU and LoTbench. blue dot ● and red dot ● denote the multimodal LLM to be tested in LoTbench and standard evaluation, respectively.

average cost required for the target LLM to achieve carefully designed HHCRs through interaction with some specific LLMs. Lower costs indicate higher creativity. This approach maintains the automation advantages of LLM-as-a-Judge based methods through the involvement of a specific LLM, while considering the average distance to HHCRs ensures alignment with human preferences. Additionally, the carefully designed HHCRs and interactive approach help mitigate the risk of information leakage in standard evaluation and improve the interpretability.

7.2 The Effectiveness of $\mathcal{E}_1$

To validate this LLM-based DAESO judgment method proposed in Section 5.5, we used a validation set of 43 DAESO samples collected during the construction of LoTbench to perform tests from multiple perspectives. First, we conducted experiments on causal chain construction and intervention using Qwen-7B, Qwen-14B, MiniGPTv2, Baichuan2, Llama2, GPT-3.5, GPT4o mini and GPT-4 with this validation set, manually inspecting each result for accuracy. As shown in Fig. 15 (Left), we can find significant performance differences in causal chain construction and intervention among these LLMs, with GPT series outperforming the others. Moreover, in Fig. 15 (Right), we considered three settings: (1) Using only a prompt to let the LLM directly zero-shot determine whether R_t and R are DAESO; (2) Zero-shot DAESO judgment with detailed E_{xp} and $\mathcal{C}$ provided; (3) Our proposed method of causal chain modeling in text space based on E_{xp} and $\mathcal{C}$. From the results, we can see that E_{xp} and $\mathcal{C}$ are crucial—without them, all LLMs performed poorly, almost guessing randomly. When this information is provided, modeling and intervening in the causal chain resulted in better judgment outcomes. Therefore, based on these experimental results, to ensure the accuracy of LoTbench and minimize reasoning costs, like API fee, we adopt GPT-4o mini as $\mathcal{E}_1$ in the actual evaluation of LoTbench and provided detailed E_{xp} for each sample.

7.3 The Effectiveness of $\mathcal{E}_2$

In Section 5.6, we need a powerful text-based LLM as evaluator $\mathcal{E}_2$ to accurately respond to the tester's spontaneous questioning Q_t with a precise answer A_t. In this section, we find that selecting GPT-4o mini is suitable for this task as shown in Fig. 15 (Left). Specifically, during the construction of the LoTbench test set, we also collect a simple validation set of 130 examples to test whether various LLMs have the ability to provide judgments for Q_t. Fig. 15 (Left) shows the judgment results of different LLMs, where most LLMs can achieve an accuracy rate of over 80%, and GPT-4o mini not only has a relatively low reasoning cost but also an accuracy rate of up to 98%. Therefore, we choose it as the judge $\mathcal{E}_2$.

7.4 The Correlation with Cognition Benchmark

(2) The creativity assessment results from LoTbench align more closely with current human cognitive theories [38], [39], [40], [41], [42]. In this Section, we compare the evaluation results of the well-known and comprehensive cognitive ability assessment MMMU with those of our proposed

LoTbench in Fig. 16 (b). We found that, although MMMU focuses on cognitive abilities while LoTbench emphasizes the creativity of LLMs, there is a significant strong correlation between their results. This indicates that the evaluation of LoTbench aligns with current human cognitive theories and reflects human-like creativity, suggesting that cognition forms the basis of creativity. The components of perception, knowledge, and reasoning enable the recognition and connection of different concepts through creative thinking, which is key to early creativity. However, it is important to note that in Fig. 10, we also observe that methods like CoT, which enhance LLM logical reasoning abilities, do not always improve S_c. This is consistent with the observations in Section 6. It suggests that creativity requires higher demands on cognitive abilities, and merely enhancing logical reasoning is insufficient to consistently boost creativity. Furthermore, we also analyze the standard evaluation shown in Section 6 in the same way, and the results, shown in Fig. 16 (a), indicate that the relationship between "creativity" by standard evaluation and LLM cognitive ability is not very significant. Of course, this does not imply that standard evaluation is an incorrect benchmark, and it indeed helps humans understand LLM creativity from different perspectives and provides preliminary quantitative results.

7.5 Limitation

While the proposed LoTbench offers intuitive and user-friendly features, it does have certain limitations. For instance, as a multi-turn interactive benchmark, it may not be suitable for evaluating all types of subjects. For example, as mentioned in Section 6.2, humans are not particularly adept at long-term interactions. Similarly, CLoTv2 faces this issue as well. Due to the multi-turn tuning described in Section 5, while its creativity shows some improvement in standard evaluations, other abilities may be slightly diminished [103], especially the ability to follow context. This limitation prevents it from being assessed using LoTbench. Currently, LoTbench is primarily designed for evaluating the creativity of general LLMs. In the future, we need to explore better methods to stimulate LLM creativity using LoTbench while minimizing the catastrophic forgetting [103] of general capabilities. On the other hand, LoTbench constructs creativity scores by estimating the average cost for the target LLM to achieve certain HHCRs through interactive methods. While this definition of LLM creativity facilitates benchmark design, it is not the only definition, and the concept of creativity still lacks a clear consensus [104], [105]. In the future, the community should explore more advanced definitions of creativity to address potential risks associated with LoTbench.

8 Conclusion

This paper investigates creativity in LLMs and provides an in-depth analysis of their Leap-of-Thought (LoT) abilities through the Oogiri game. In particular, given some inherent issues, like information leakage, in current assessments of LLM creativity, we introduce a novel interactive benchmark, LoTbench, to effectively evaluate LLM creativity. Our findings reveal that while LLMs exhibit limited creativity, the gap between LLM and human creativity is not significant. Additionally, we find a strong correlation between LoTbench results and MMMU, a comprehensive benchmark for multimodal LLM cognition. This suggests that LoTbench aligns with human cognitive theories, capturing human-like creativity and emphasizing cognition as a foundational element in the early stages of creativity, enabling the integration of diverse concepts.

Acknowledgments

This work was supported by National Science and Technology Major Project (No.2021ZD0111601), National Natural Science Foundation of China under Grants No. 623B2099 and 62325605, Guangdong Basic and Applied Basic Research Foundation (No.2023A1515011374), and Guangzhou Science and Technology Program (No.2024A04J6365). Pan Zhou was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grants (project ID: 23-SIS-SMU-028 and 23-SIS-SMU-070).

References

[1] S. Zhong, Z. Huang, S. Gao, W. Wen, L. Lin, M. Zitnik, and P. Zhou, "Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 246–13 257.

[2] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[3] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[4] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[5] A. Saparov and H. He, "Language models are greedy reasoners: A systematic formal analysis of chain-of-thought," *arXiv preprint arXiv:2210.01240*, 2022.

[6] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[7] S. Zhong, Z. Huang, W. Wen, J. Qin, and L. Lin, "Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 567–578.

[8] T. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, X. Zhang *et al.*, "What can large language models do in chemistry? a comprehensive benchmark on eight tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 59 662–59 688, 2023.

[9] Z. Liang, K. Guo, G. Liu, T. Guo, Y. Zhou, T. Yang, J. Jiao, R. Pi, J. Zhang, and X. Zhang, "Scemqa: A scientific college entrance level multimodal question answering benchmark," *arXiv preprint arXiv:2402.05138*, 2024.

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

[11] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. 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*, 2024, pp. 9556–9567.

[12] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[13] J. Li and W. Lu, "A survey on benchmarks of multimodal large language models," *arXiv preprint arXiv:2408.08632*, 2024.

[14] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[16] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," *arXiv preprint arXiv:2210.03493*, 2022.

[17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[18] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models, may 2023," *arXiv preprint arXiv:2305.10601*, 2023.

[19] J. Long, "Large language model guided tree-of-thought," *arXiv preprint arXiv:2305.08291*, 2023.

[20] A. Talmor, O. Tafjord, P. Clark, Y. Goldberg, and J. Berant, "Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 227–20 237, 2020.

[21] E. Callaway, "Cognitive science: Leap of thought," 2013.

[22] K. J. Holyoak, P. Thagard, and S. Sutherland, "Mental leaps: analogy in creative thought," *Nature*, vol. 373, no. 6515, pp. 572–572, 1995.

[23] C. Olson, "The leap of thinking: A comparison of heidegger and the zen master dogen," *Philosophy Today*, vol. 25, no. 1, p. 55, 1981.

[24] D. Hofstadter, "A review of mental leaps: analogy in creative thought," *AI Magazine*, vol. 16, no. 3, pp. 75–75, 1995.

[25] K. J. Holyoak and P. Thagard, *Mental leaps: Analogy in creative thought*. MIT press, 1996.

[26] J. Kitto, D. Lok, and E. Rudowicz, "Measuring creative thinking: An activity-based approach," *Creativity Research Journal*, vol. 7, no. 1, pp. 59–69, 1994.

[27] M. Mölle, L. Marshall, B. Wolf, H. L. Fehm, and J. Born, "Eeg complexity and performance measures of creative thinking," *Psychophysiology*, vol. 36, no. 1, pp. 95–104, 1999.

[28] H. Jiang and Q.-p. Zhang, "Development and validation of team creativity measures: A complex systems perspective," *Creativity and Innovation Management*, vol. 23, no. 3, pp. 264–275, 2014.

[29] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[30] Wikimedia, "Glossary of owarai terms," *https://en.wikipedia.org/wiki/Glossary_of_owarai_terms*, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Glossary_of_owarai_terms

[31] B. Y. Lin, Z. Wu, Y. Yang, D.-H. Lee, and X. Ren, "Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge," *arXiv preprint arXiv:2101.00376*, 2021.

[32] Y. Jiang, F. Ilievski, and K. Ma, "Brainteaser: Lateral thinking puzzles for large language model," *arXiv preprint arXiv:2310.05057*, 2023.

[33] Y. Zhang and X. Wan, "Birdqa: A bilingual dataset for question answering on tricky riddles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 748–11 756.

[34] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv: 2311.03079*, 2023.

[35] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[36] Q. Chen, B. Zhang, G. Wang, and Q. Wu, "Weak-eval-strong: Evaluating and eliciting lateral thinking of llms with situation puzzles," 2024. [Online]. Available: https://arxiv.org/abs/2410.06733

[37] S. Huang, S. Ma, Y. Li, M. Huang, W. Zou, W. Zhang, and H.-T. Zheng, "Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles," *arXiv preprint arXiv:2308.10855*, 2023.

[38] O. Martinsen, "The effect of individual differences in cognitive style and motives in solving insight problems," *Scandinavian Journal of Educational Research*, vol. 38, no. 2, pp. 83–96, 1994.

[39] ——, "Insight problems revisited: The influence of cognitive styles and experience on creative problem solving," *Creativity Research Journal*, vol. 6, no. 4, pp. 435–447, 1993.

[40] G. Kaufmann, "The explorer and the assimilator: A cognitive style distinction and its potential implications for innovative problem solving," *Scandinavian Journal of Educational Research*, vol. 23, no. 3, pp. 101–108, 1979.

[41] M. A. Runco and I. Chand, "Cognition and creativity," *Educational psychology review*, vol. 7, pp. 243–267, 1995.

[42] S. Mednick, "The associative basis of the creative process." *Psychological review*, vol. 69, no. 3, p. 220, 1962.

[43] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[44] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[45] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[46] H. Chen and N. Ding, "Probing the creativity of large language models: Can models produce divergent semantic association?" *arXiv preprint arXiv:2310.11158*, 2023.

[47] Z. Ling, Y. Fang, X. Li, T. Mu, M. Lee, R. Pourreza, R. Memisevic, and H. Su, "Unleashing the creative mind: Language model as hierarchical policy for improved exploration on challenging problem solving," *arXiv preprint arXiv:2311.00694*, 2023.

[48] D. Summers-Stay, C. R. Voss, and S. M. Lukin, "Brainstorm, then select: a generative language model improves its creativity score," in *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023.

[49] M. Park, E. Leahey, and R. J. Funk, "Papers and patents are becoming less disruptive over time," *Nature*, vol. 613, no. 7942, pp. 138–144, 2023.

[50] S. Liang, Z. Huang, and H. Zhang, "Stiffness-aware neural network for learning hamiltonian systems," in *International Conference on Learning Representations*, 2021.

[51] Z. Huang, S. Liang, H. Zhang, H. Yang, and L. Lin, "On fast simulation of dynamical system with neural vector enhanced numerical solver," *Scientific Reports*, vol. 13, no. 1, p. 15254, 2023.

[52] B. Swanson, K. Mathewson, B. Pietrzak, S. Chen, and M. Dinalescu, "Story centaur: Large language model few shot learning as a creative writing tool," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021, pp. 244–256.

[53] T. Chakrabarty, V. Padmakumar, and H. He, "Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing," *arXiv preprint arXiv:2210.13669*, 2022.

[54] T. Wu, E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry, and C. J. Cai, "Promptchainer: Chaining large language model prompts through visual programming," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–10.

[55] K. Binsted, A. Nijholt, O. Stock, C. Strapparava, G. Ritchie, R. Manurung, H. Pain, A. Waller, and D. O'Mara, "Computational humor," *IEEE intelligent systems*, vol. 21, no. 2, pp. 59–69, 2006.

[56] Z. Xu, S. Yuan, L. Chen, and D. Yang, "" a good pun is its own reword": Can large language models understand puns?" *arXiv preprint arXiv:2404.13599*, 2024.

[57] D. Shahaf, E. Horvitz, and R. Mankoff, "Inside jokes: Identifying humorous cartoon captions," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1065–1074.

[58] K. Tanaka, H. Yamane, Y. Mori, Y. Mukuta, and T. Harada, "Learning to evaluate humor in memes based on the incongruity theory," in *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, 2022, pp. 81–93.

[59] H. Xu, W. Liu, J. Liu, M. Li, Y. Feng, Y. Peng, Y. Shi, X. Sun, and M. Wang, "Hybrid multimodal fusion for humor detection," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022, pp. 15–21.

[60] M. Amin and M. Burghardt, "A survey on approaches to computational humor generation," in *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2020, pp. 29–41.

[61] H. Zhang, D. Liu, J. Lv, and C. Luo, "Let's be humorous: Knowledge enhanced humor generation," *arXiv preprint arXiv:2004.13317*, 2020.

[62] N. Hossain, J. Krumm, T. Sajed, and H. Kautz, "Stimulating creativity with funlines: A case study of humor generation in headlines," *arXiv preprint arXiv:2002.02031*, 2020.

[63] O. Popova and P. Dadić, "Does ai have a sense of humor? clef 2023 joker tasks 1, 2 and 3: using bloom, gpt, simplet5, and more for pun detection, location, interpretation and translation," *Proceedings of the Working Notes of CLEF*, 2023.

[64] D. S. Chauhan, G. V. Singh, A. Ekbal, and P. Bhattacharyya, "Mhadig: A multilingual humor-aided multiparty dialogue generation in multimodal conversational setting," *Knowledge-Based Systems*, vol. 278, p. 110840, 2023.

[65] S. Zhong, Z. Huang, D. Li, W. Wen, J. Qin, and L. Lin, "Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima," *arXiv preprint arXiv:2402.11262*, 2024.

[66] Z. Yuan, J. Ren, C.-M. Feng, H. Zhao, S. Cui, and Z. Li, "Visual programming for zero-shot open-vocabulary 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 623–20 633.

[67] G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, and L. Wang, "Towards revealing the mystery behind chain of thought: a theoretical perspective," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[68] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch, "Faithful chain-of-thought reasoning," *arXiv preprint arXiv:2301.13379*, 2023.

[69] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang, "Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5168–5191, 2023.

[70] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

[71] J. Hessel, A. Marasović, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi, "Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker Caption Contest," in *Proceedings of the ACL*, 2023.

[72] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[73] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[74] F. Radlinski and N. Craswell, "Comparing the sensitivity of information retrieval metrics," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 667–674.

[75] P. Sloane, *The leader's guide to lateral thinking skills: Powerful problem-solving techniques to ignite your team's potential*. Kogan Page Publishers, 2003.

[76] L. Thinking, "Creativity step by step," *By Edward de Bono*, 1970.

[77] J. Lee, "Mental leap," in *Encyclopedia of the Sciences of Learning*, N. M. Seel, Ed. Boston, MA: Springer US, 2012, pp. 2194–2194. [Online]. Available: https://doi.org/10.1007/978-1-4419-1428-6_1557

[78] OpenAI, "Gpt-4 technical report," 2023.

[79] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[80] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[81] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[82] R. Hataya, H. Bao, and H. Arai, "Will large-scale generative models corrupt future datasets?" in *ICCV*, 2023.

[83] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "Model dementia: Generated data makes models forget," *arXiv preprint arXiv:2305.17493*, 2023.

[84] N. Micheletti, S. Belkadi, L. Han, and G. Nenadic, "Exploration of masked and causal language modelling for text generation," *arXiv preprint arXiv:2405.12630*, 2024.

[85] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin, "tinybenchmarks: evaluating llms with fewer examples," *arXiv preprint arXiv:2402.14992*, 2024.

[86] A. Kipnis, K. Voudouris, L. M. S. Buschoff, and E. Schulz, "metabench–a sparse benchmark to measure general ability in large language models," *arXiv preprint arXiv:2407.12844*, 2024.

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

[88] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335.

[89] G. Team, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[90] Z. e. a. Chen, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

[91] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[92] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang *et al.*, "Yi: Open foundation models by 01. ai," *arXiv preprint arXiv:2403.04652*, 2024.

[93] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[94] W. Dai, J. Li, and et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[95] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[96] Baichuan, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023. [Online]. Available: https://arxiv.org/abs/2309.10305

[97] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[98] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[99] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "Gpteval: A survey on assessments of chatgpt and gpt-4," *arXiv preprint arXiv:2308.12488*, 2023.

[100] W. Ge, S. Chen, G. Chen, J. Chen, Z. Chen, S. Yan, C. Zhu, Z. Lin, W. Xie, X. Wang *et al.*, "Mllm-bench, evaluating multi-modal llms using gpt-4v," *arXiv preprint arXiv:2311.13951*, 2023.

[101] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.

[102] N. Arsov, M. Dukovski, B. Evkoski, and S. Cvetkovski, "A measure of similarity in textual data using spearman's rank correlation coefficient," *arXiv preprint arXiv:1911.11750*, 2019.

[103] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma, "Investigating the catastrophic forgetting in multimodal large language models," *arXiv preprint arXiv:2309.10313*, 2023.

[104] A. Elgammal and B. Saleh, "Quantifying creativity in art networks," *arXiv preprint arXiv:1506.00711*, 2015.

[105] A. Egan, R. Maguire, L. Christophers, and B. Rooney, "Developing creativity in higher education for 21st century learners: A protocol for a scoping review," *International Journal of educational research*, vol. 82, pp. 21–27, 2017.

APPENDIX A: THE DETAILS OF I_t AND GENERATION OF R_t

Now, please help me complete a task involving filling in a blank with a humorous and creative . All the input information is provided in the INPUT section, which includes:
1. IMAGE: A given image.
2. IMAGE CAPTION: A detailed description of the given image.
3. RESPONSE: A sentence with a blank to be filled. You need to complete the part based on all the INPUT information.
4. TIPS: Some guidelines for the task, which include:
• Q&A: Questions and answers related to . Carefully analyze and follow these hints to generate a creative and humorous .
• CLUE: Descriptive hints about Understand and adhere to these clues to generate a creative and humorous .
• WRONG-ANS: Examples of that are neither innovative nor humorous. Avoid completing the blank with similar content.
Based on the provided INPUT information and image, use divergent thinking to complete the part. Ensure that the final RESPONSE pairs well with the IMAGE, making it witty, humorous, and creative. Output the result in the specified OUTPUT format.

Here are some examples:

Example1:

INPUT: "IMAGE": ,
"IMAGE CAPTION": "A soldier holding a big knife, staring angrily ahead, seems very angry",
"RESPONSE": "⟨WORD⟩ After reading my paper...",
"TIPS":
"WRONG-ANS (⟨WORD⟩ is not the following content)":
1: "Programmer",
2: "Mountain climber",
"SYSTEM CLUE":
"CLUE1": "⟨WORD⟩ is a kind of person",
"CLUE2": "This kind of person has high knowledge",
"Q&A (OUTPUT should not be repeated with Q&A)":
1: "Q1": "Is it related to the soldier?",
"A1": "No",
2: "Q2": "Is it related to the school?",
"A2": "Yes"

OUTPUT:
"⟨WORD⟩": "Tutor",
"RESPONSE": "After the tutor read my paper..."

.....
Example2: .....
Example3: .....

Referring to the example above, please use the latest INPUT information provided below and the accompanying image to creatively and humorously complete the ⟨WORD⟩. Ensure that the supplemented RESPONSE matches the provided IMAGE, making it witty, imaginative, and engaging. Format the result strictly according to the example shown in the OUTPUT.

INPUT:
"IMAGE": "⟨image⟩",
"IMAGE CAPTION": "⟨caption⟩",
"RESPONSE": "⟨response⟩",
"TIPS": ⟨tips⟩
OUTPUT:

APPENDIX B: THE DETAILS OF RETHINKING AND GENERATING Q_t

Given an image IMAGE and its detailed description IMAGE CAPTION. It is known that this IMAGE has a very humorous and creative caption RESPONSE.
Now there is a task of looking at the image to complete the caption, that is, to generate a humorous and creative ⟨WORD⟩. All the input information is in INPUT, they are

1. IMAGE: a given image
2. IMAGE CAPTION: a detailed description of the given image IMAGE
3. RESPONSE: a text of an IMAGE with the content ⟨WORD⟩ to be completed, you need to complete the ⟨WORD⟩ part according to IMAGE and IMAGE CAPTION
4. Q&A: some known queries and corresponding answers about ⟨WORD⟩
5. CLUE: some descriptive hints related to ⟨WORD⟩
6. WRONG-ANS: some innovative and humorous ⟨WORD⟩, you should not complete similar content

In order to better complete the ⟨WORD⟩ in RESPONSE, so that the combination of IMAGE and RESPONSE is very humorous and creative, you can first use divergent thinking to ask a general question (that is, a question with the answer of Yes or No) for possible ⟨WORD⟩ to help the generation of ⟨WORD⟩.

Here are some examples:

Example1:

INPUT:
"IMAGE": ,
"IMAGE CAPTION": "A soldier holding a big knife, staring angrily ahead, seems very angry",
"RESPONSE": "⟨WORD⟩ After reading my paper...",
"TIPS":
"WRONG-ANS (⟨WORD⟩ is not the following content)":
1: "Programmer",
2: "Mountain climber",
"SYSTEM CLUE":
"CLUE1": "⟨WORD⟩ is a kind of person",
"CLUE2": "This kind of person has high knowledge",
"Q&A (OUTPUT should not be repeated with Q&A)":
1:
"Q1": "Is it related to the soldier?",
"A1": "No",
2: "Q2": "Is it related to the school?",
"A2": "Yes"
OUTPUT: ⟨WORD⟩ Is it something edible?

.....
Example2: .....
Example3: .....

Referring to the above example, please use the latest INPUT information and the pictures provided below to think divergently and ask a general question (i.e., a question with a yes or no answer) for possible ⟨WORD⟩ to help generate ⟨WORD⟩. Please note that only general questions are output.
INPUT:
"IMAGE": "⟨image⟩",
"IMAGE CAPTION": "⟨caption⟩",
"RESPONSE": "⟨response⟩",
"TIPS": ⟨tips⟩

OUTPUT:

APPENDIX C: ANSWER THE QUESTION Q_t AND PROVIDE A_t

Given a ⟨WORD⟩ and a question QUESTION about ⟨WORD⟩, you need to give the answer to this question QUESTION through common sense and reasoning. If it is correct, output Yes in the format, if it is wrong, output No in the format. Here are some examples,

Example1:

INPUT:
"⟨WORD⟩": "Mentor",
"QUESTION": "Is ⟨WORD⟩ related to soldiers?"
OUTPUT: No

Example2:

INPUT:
"⟨WORD⟩": "Cat",
"QUESTION": "Is ⟨WORD⟩ an animal?"
OUTPUT: Yes

Please read the above examples carefully, and output OUTPUT strictly in the format given the new INPUT shown below

INPUT:
"⟨WORD⟩": "⟨word⟩",
"QUESTION": "⟨question⟩"

OUTPUT:

APPENDIX D: THE PROMPT DETAILS OF DAESO

Given a detailed text description of an image ⟨IMAGE CAPTION⟩, it has a very humorous and creative caption ⟨GTR⟩ and its detailed explanation ⟨EXP⟩, please help me parse the entities, relationships and causal chains of ⟨EXP⟩.
And analyze, if ⟨GTW⟩ in ⟨GTR⟩ is replaced with ⟨RESPONSE⟩, does ⟨RESPONSE⟩ still meet the analysis, that is, does it still have similar humor, creativity and function?
.....
Please answer strictly in the following format:

.....
Example1: .....
Example2: .....

"SUMMARY": "Yes/No",
"EXPLANATION": "..."
The content of SUMMARY is Yes or No, indicating whether ⟨RESPONSE⟩ still has similar sense of humor and creativity after ⟨GTW⟩ in ⟨GTR⟩ is replaced with ⟨RESPONSE⟩;
The content of EXPLANATION is the analysis of the entities, relationships and causal chains of the paragraph ⟨EXP⟩, as well as the analysis of whether ⟨GTR⟩ still has similar sense of humor and creativity.
If ⟨RESPONSE⟩ is not a simple phrase or sentence, but a complex format, then SUMMARY is No.