AssoCiAm: A Benchmark for Evaluating Association Thinking while Circumventing Ambiguity

Yifan Liu[1]*, Wenkuan Zhao[1]*, Shanshan Zhong[1], Jinghui Qin[2], Mingfu Liang[3], Zhongzhan Huang[1], Wushao Wen[1]†,

[1]Sun Yat-sen University, [2]Guangdong University of Technology, [3]Northwestern University

Abstract

Recent advancements in multimodal large language models (MLLMs) have garnered significant attention, offering a promising pathway toward artificial general intelligence (AGI). Among the essential capabilities required for AGI, creativity has emerged as a critical trait for MLLMs, with association serving as its foundation. Association reflects a model's ability to think creatively, making it vital to evaluate and understand. While several frameworks have been proposed to assess associative ability, they often overlook the inherent ambiguity in association tasks, which arises from the divergent nature of associations and undermines the reliability of evaluations. To address this issue, we decompose ambiguity into two types—internal ambiguity and external ambiguity—and introduce AssoCiAm, a benchmark designed to evaluate associative ability while circumventing the ambiguity through a hybrid computational method. We then conduct extensive experiments on MLLMs, revealing a strong positive correlation between cognition and association. Additionally, we observe that the presence of ambiguity in the evaluation process causes MLLMs' behavior to become more random-like. Finally, we validate the effectiveness of our method in ensuring more accurate and reliable evaluations. See Project Page for the data and codes.

1 Introduction

Recently, multimodal large language models (MLLMs) have made remarkable advancements, enabling them to better understand the human world and mimic human multimodal perception (Li et al., 2024e; Liang et al., 2023, 2024). This progress highlights a potential pathway toward achieving artificial general intelligence (AGI), a long-term goal in the field of artificial intelligence (Yin et al., 2024; Huang et al., 2025b). As a hallmark of human intelligence, creativity is a critical ability that AGI-level MLLMs should possess and be evaluated for (Bellemare-Pepin et al., 2024). A key aspect of evaluating creativity lies in assessing associative ability, which plays a central role in creativity by driving generative processes (Beaty and Kenett, 2023; Benedek et al., 2012; Marron et al., 2018; Benedek et al., 2020).

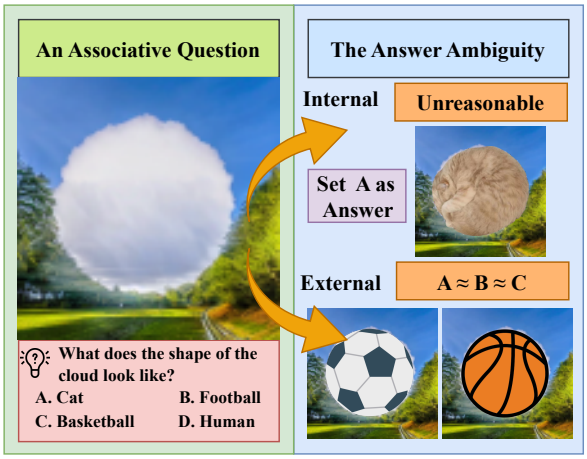

Figure 1: An example illustrating the inherent ambiguity in association evaluation. The left image presents an associative task where models associate the shape of the cloud with the most similar option from the given choices. The right image highlights two types of ambiguity: internal ambiguity (top), where the ground truth is set to option A with the unreasonable explanation that the cat's curled posture resembles a circle; and external ambiguity (bottom), where options B and C are equally correct due to their similar circular shapes but are not designated as the correct answers in this context.

Associative ability refers to the process of forming novel connections between seemingly unrelated concepts stored in memory (Beaty and Kenett, 2023). This process is divergent, originating from a single idea and expanding into a space of possible ideas through paradigms such as divergent thinking, lateral thinking or "thinking outside the box". To evaluate associative abilities, numerous researchers

*Equal contribution
†Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5203–5219
November 4-9, 2025

have proposed diverse evaluation frameworks and released various benchmarks, exploring different dimensions of this ability. Most of these benchmarks (Jiang et al., 2023; Kraaijveld et al., 2024; Lin et al., 2021; Zhang and Wan, 2022; Zhong et al., 2024b) adopt a multiple-option format, where a single answer is designated as correct.

However, these benchmarks overlook the answer ambiguity inherent in associative tasks, which stems from the divergent nature of associations. This ambiguity can be explicitly decomposed into two parts: **internal** and **external** ambiguity.

(1) **Internal Ambiguity.** Internal ambiguity refers to cases where the designated answer itself is unreasonable. For example, in the associative task shown in Fig. 1 (left), the task requires the models to associate the shape of a cloud with an object that resembles it (Zhong et al., 2024b). Here, the correct answer is set as Option A, "cat", because a cat can curl into a circular shape, as shown in Fig. 1 (right). However, this answer is unreasonable since the typical shape of a cat does not match the shape of the cloud. With internal ambiguity, models cannot identify the correct answer from options even if they possess strong associative ability, and therefore the evaluation results fail to reflect the true associative ability of models.

(2) **External Ambiguity.** External ambiguity arises when the correct options include multiple equally correct options. As illustrated in Fig. 1 (right), the correct answer, Option A, is not unique; Options B and C are also valid because their circular shapes closely resemble the shape of the cloud. In this associative task, external ambiguity occurs because multiple options share a resemblance to the natural object's shape. In such cases, directly judging a model's answer against a ground truth lead to inaccurate evaluations.

Neglecting internal or external ambiguity introduces risks that render evaluation results nonfaithful. Effectively circumventing answer ambiguity is critical for ensuring that benchmarks produce reliable evaluations. To address this issue, we construct a multimodal associative benchmark that circumvents ambiguity and provides more faithful quantification of associative ability.

In our task design, MLLMs are required to perform a associative task, such as that illustrated in Fig. 1 (left), where they identify the shape of a specific natural object in an image and then associate it with the most visually similar object. Ambiguity in this task arises from divergent visual associations. To mitigate ambiguity, we propose a novel hybrid computational method that mitigates both internal and external ambiguity.

Using our method, we construct the first multimodal associative benchmark circumventing ambiguity, AssoCiAm, to quantify models' associative abilities. Furthermore, we conduct large-scale evaluations on various models and identify a strong correlation between associative ability and cognitive capability. Through a series of experiments, we analyze the impact of ambiguity on association evaluation, which leads to random-like behavior in models, and validate the effectiveness of our method in mitigating ambiguity. Our contributions are summarized as follows:

- We first introduce the answer ambiguity inherent in association evaluation and discuss its impact on the reliability of association evaluation.
- We construct AssoCiAm, for evaluating associative ability and propose a hybrid computational method to circumvent the ambiguity.
- We conduct extensive experiments to assess associative abilities of MLLMs, show the impact of ambiguity on association evaluation, and validate the effectiveness of our proposed method.

2 Related Work

Multimodal Large Language Model. MLLMs are LLM-based systems capable of receiving, reasoning and producing outputs across multiple modalities (Yin et al., 2024). To endow models with such capabilities, researchers typically position pre-trained LLMs (Chung et al., 2024; Touvron et al., 2023a; Chiang et al., 2023; Bai et al., 2023a; Touvron et al., 2023b) as the brain of the system while employing multimodal encoders (Zhao et al., 2025; Cherti et al., 2023; Radford et al., 2021; Sun et al., 2023) as sensory components to process multimodal information. Recently, various techniques have been proposed (Li et al., 2023a; Zhong et al., 2023; Lu et al., 2022; Zhang et al., 2023b; Huang et al., 2025a; Zhong et al., 2024a; Yang et al., 2024; Wu and Xie, 2024) to enhance the emergent capabilities of MLLMs, bridging the gap between human and artificial intelligence. These advancements underscore the importance of evaluation in the research and development of MLLMs (Huang and Zhang, 2024).

Computational Creativity. The emergence of increasingly powerful MLLMs has reignited interest in computational creativity (Ismayilzada et al., 2024b). Computational creativity encompasses a

broad range of tasks, including linguistic creativity (Mittal et al., 2022; Arora et al., 2022; Xie et al., 2025; Chakrabarty et al., 2023; Ismayilzada et al., 2024a), creative problem solving (Jiayang et al., 2023; Lewis and Mitchell, 2024; Opiełka et al., 2024; Mitchell et al., 2023; Ahrabian et al., 2024), and artistic creativity (Yang et al., 2022; Popescu-Belis et al., 2023; Shi et al., 2023b; Brooks et al., 2024; Copet et al., 2024). A key underlying capability for all these forms of creativity is associative ability, as these creative tasks rely on associative thinking to form connections between diverse concepts. Therefore, gaining insight into creativity necessitates a deeper understanding of associative ability, which serves as the core of creativity.

Associative Thinking Evaluation. Various frameworks have been developed to evaluate associative ability from diverse perspectives. For instance, DAT (Chen and Ding, 2023) assesses models' associative abilities by measuring semantic distance, reporting that models outperform humans in this aspect. Additionally, the Alternate Uses Test (Guilford, 1967) and the Torrance Tests of Creative Thinking have been used to evaluate models, with studies indicating that GPT-3 (Brown, 2020) and GPT-4 (Achiam et al., 2023) achieve near-human performance (Góes et al., 2023; Guzik et al., 2023; Hubert et al., 2024; Koivisto and Grassini, 2023). Despite these advancements highlighting progress in associative tasks primarily dependent on divergent thinking, models continue to show gaps compared to humans on associative tasks requiring lateral thinking or "thinking outside the box" (Jiang et al., 2023; Kraaijveld et al., 2024; Lin et al., 2021; Zhang and Wan, 2022; Zhong et al., 2024b; Huang et al., 2025c). Our benchmark specifically focuses on this type of tasks, addressing the unique ambiguities inherent in associative evaluations.

3 The Need for New Benchmarks

Current benchmarks have provided valuable insights into the associative abilities of MLLMs while also highlighting their limitations. However, there is still room for improvement. Given the divergent nature of association, the ambiguity present in different datasets vary and often remains implicit. As a result, the ambiguity in an associative benchmark go unnoticed during its design. This makes it essential to explicitly account for ambiguity when constructing benchmarks to ensure more reliable and fair assessments. Furthermore, we analyze several existing benchmarks, including BiRdQA (Zhang and Wan, 2022), BrainTeaser (Jiang et al., 2023), and RiddleSense (Lin et al., 2021), by sampling 70 questions from each. Our analysis reveals that 15.7%, 24.3%, and 17.1% of the questions from these benchmarks contain ambiguity, indicating that ambiguity is both prevalent and non-trivial. These findings point to the need for greater fairness in evaluation and highlight the importance of developing new benchmarks that explicitly and systematically mitigate the ambiguity.

4 Construction of AssoCiAm

In this section, we present the pipeline for constructing AssoCiAm that mitigates ambiguity.

4.1 Overview of The Pipeline

AssoCiAm follows a multiple-option question-answering format. A test sample consists of two main components. (1) **An image.** As illustrated in Fig. 2 (d, e, f), an image is a mask, e.g. Fig. 2 (b), filled with natural objects such as clouds, beaches, and waterfalls, embedded in a natural background. (2) **A Question-Options pair.** Questions and options are presented in text format. For each question, models are given m options, only one of which is correct; we term this an mT1 question. Our benchmark includes three subtasks: 4T1, 7T1, and 10T1, each consisting of a series of mT1 questions with 4, 7, and 10 options, respectively.

To construct the benchmark, we design a two-stage pipeline. (1) **Avoiding Internal Ambiguity**, which collects images while avoiding internal ambiguity. We collect representative masks that eliminate internal ambiguity. Using these masks as guidance, we generate images and review them to ensure quality and consistency. (2) **Avoiding External Ambiguity**, which constructs Question-Options pairs while mitigating external ambiguity. We introduce a structured way to model the options to systematically mitigate external ambiguity. Based on this model, we apply an optimization method to select appropriate distractors and construct reliable test samples.

4.2 Avoiding Internal Ambiguity

We propose a hybrid computational approach to eliminate the internal ambiguity during the process of collecting masks and images that we need. In this context, eliminating internal ambiguity means ensuring that shapes of masks are representative.

First, we collect masks extracted from images. The images come from the publicly available ILSVRC12 dataset (Russakovsky et al., 2015). The dataset is specifically designed for image classification, focusing on typical features of each class, making it well-suited for extracting representative masks. We sample 25 images per class and use the SAM model (Kirillov et al., 2023) to extract their corresponding masks. For instance, Fig. 2 (b) is the mask extracted from Fig. 2 (a).

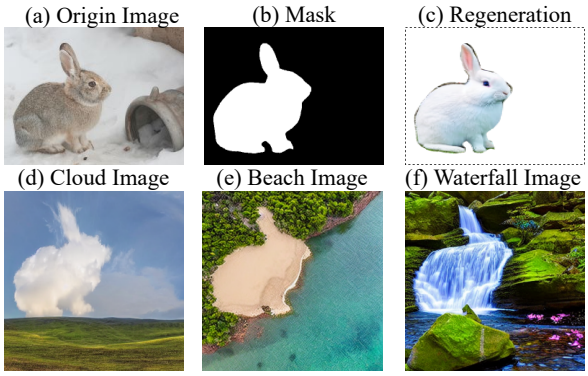

Figure 2: Examples for images used in our method. (a) is a image collected. (b) is the mask extracted from (a). (c) are images regenerated guided by (b) and overlaid with (b). (d, e, f) are final images used in our benchmark.

Second, we filter and augment masks that represent image classes, following a **key principle**: a representative mask should be inherently recognizable and align with general human perception.

To facilitate this, we first apply a computational filtering method to simplify the selection process and help human experts work more efficiently and precisely when refining and selecting representative masks. The method is based on the premise that a truly representative mask should enable a **regenerated image**, guided by the mask, to be **recognized** as belonging to its original class.

Specifically, we use a control diffusion model (Rombach et al., 2022; Zhang et al., 2023a; Huang et al., 2023; Shi et al., 2023a) to regenerate images guided by masks, with irrelevant regions overlaid to reduce visual distraction. For each mask, eight images are regenerated to mitigate randomness. As shown in Fig. 2 (c), a rabbit image is regenerated and overlaid with its corresponding mask from Fig. 2 (b).

Next, we use CLIP (Radford et al., 2021) to classify the regenerated images. A mask's representativeness is measured by the average classification probability across its regenerated images. Masks scoring above 97% are retained for further review.

Human experts then are invited to select the **most representative** masks from this **filtered set**. Additionally, human experts include a small number of masks sourced from the public website flaticon, which provides free-to-use images for project purposes. These masks, though excluded from the selected classes, are commonly associated with the context of our task. This final collection of masks thus constitutes the **original mask set**.

With the original mask set, we use the control diffusion model (Rombach et al., 2022; Zhang et al., 2023a; Huang et al., 2023; Shi et al., 2023a) to generate images of clouds, beaches, and waterfalls, each guided by its corresponding mask, as shown in Fig. 2 (d, e, f). A subsequent manual filtering step ensures consistency, clarity, and completeness of the generated images, thereby eliminating internal ambiguity in AssoCiAm.

4.3 Avoiding External Ambiguity

To construct Question-Options pairs for AssoCiAm, we use the selected classes as candidate distractors. To avoid **external ambiguity**, it is essential to prevent distractor classes with shapes similar to the correct answer from being selected. Since the most representative mask for each class has been identified, the similarity between shapes of two classes corresponds to the similarity between their corresponding masks.

Given this, we represent the relationships among options using a graph. Specifically, we define an undirected complete graph $G = \langle V, E \rangle$, where $V = \{v_1, v_2, \ldots, v_n\}$ represents the set of classes, and $E = \{e_{ij}\}$ is the set of edges. Each edge e_{ij} quantifies the similarity between masks v_i and v_j, serving as a measure of their resemblance.

To determine the optimal set of distractors and answers, we define a subgraph $G' = \langle V', E' \rangle$, where $V' \subseteq V$ and $E' \subseteq E$, with $v_0 \in V'$ representing the correct answer. Ensuring the uniqueness of the correct answer corresponds to minimizing the following function:

$$S(G') = \sum_{v_i \in V', i \neq 0} \frac{1}{|V'| - 1} e_{0i}, \qquad (1)$$

where e_{ij} represents the similarity between the classes v_i and v_j. This ensures that distractors are as dissimilar as possible from the correct answer.

However, minimizing $S(G')$ alone can lead to a scenario where the distractors are highly similar to

each other. This allows the model to identify the correct answer by eliminating the distractors based on their mutual resemblance rather than truly recognizing the correct answer. To address this issue, we include the **variance** $\sigma^2(G')$ of all edge weights in G' as a regularization term, which measures the spread of similarities among options relative to their mean and does not directly indicate the magnitude of similarity. The combined objective function is expressed as:

$$F(G') = S(G') + \lambda\sigma^2(G'), \tag{2}$$

where $\sigma^2(G')$ represents the variance of all the edge weights, and λ is a regularization parameter that controls the trade-off between minimizing similarity and maintaining variance.

To construct the graph, we calculated the similarity between masks via DINO-v2 (Oquab et al., 2023; Darcet et al., 2023), which focuses exclusively on their geometric shapes since the masks are binary (black and white) and differ primarily in shape. Then, Genetic Algorithm (Mirjalili and Mirjalili, 2019; Holland, 1992; Mitchell, 1998; Goldberg, 2013) is applied for the optimization. The implementation of the Genetic Algorithm is provided in Appendix A.3.

Finally, we construct Question-Options pairs for the benchmark within the generated images and the algorithm. For each mT1 subtask, we select three questions per image. These questions express the same meaning but are phrased differently to query the model. This design helps minimize randomness in the evaluation process. Then, for each question, our algorithm selects distractors based on the correct answer to generate m options. In this way, we construct the final test samples and assemble AssoCiAm, which circumvents both internal and external ambiguity.

4.4 AssoCiAm Analysis

AssoCiAm consists of 2,025 test samples, each comprising an image with a corresponding Question-Options pair. The images are of high quality, with a resolution of 512, ensuring sufficient detail to effectively convey the necessary information. The benchmark includes three mT1 tasks with varying difficulty levels. AssoCiAm covers 25 classes, encompassing diverse aspects of daily life, ensuring its comprehensiveness. These key attributes make AssoCiAm a lightweight yet comprehensive benchmark for evaluating the associative ability of models. Table 1 provides an overview of the key statistics of AssoCiAm.

Statistics	Value
Class	25
Image	225
Image Resolution	512×512
4T1 Question	675
7T1 Question	675
10T1 Question	675

Table 1: Key statistics of AssCiAm. mTn question represents question with m options and n correct answers.

5 Experiment Setup

5.1 Model Seclection

MLLMs. To comprehensively assess model performance, we consider a range of both open- and closed-source multimodal large language models. For each model family, we include a series of available models, ranging from good to better performance. (i) InternVL2 (Chen et al., 2024c,b; Gao et al., 2024; Wang et al., 2024; Chen et al., 2024a), (ii) LLaVA-Next (Liu et al., 2023b,a, 2024; Zhang et al., 2024; Li et al., 2024b,a,d), (iii) LLaVA-OneVision (Li et al., 2024c), (iv) MiniCPM-V (Yao et al., 2024), (v) Yi (AI et al., 2024), (vi) CogVLM (Wang et al., 2023; Hong et al., 2023, 2024), (vii) mPLUG-Owl (Ye et al., 2023a,b, 2024), (viii) Otter (Li et al., 2023c,b), (ix) Qwen-VL (Bai et al., 2023b), (x) Visual-GLM (Ding et al., 2021; Du et al., 2022), (xi) MiniGPT-4 (Zhu et al., 2023; Chen et al., 2023), (xii) ChatGPT-4o mini (Hurst et al., 2024), (xiii) Gemini-1.5-pro (Team et al., 2024).

Human experts. To investigate human performance on this task, we invite three human experts to complete a quiz consisting of 10% of our benchmark, randomly sampled from the benchmark.

5.2 Evaluation Settings

In our experiments, we use Top-1 accuracy as the evaluation metric, as it provides a fair measure for the multiple-option question-answering format and is widely adopted in prominent commonsense reasoning tasks (Mihaylov et al., 2018; Talmor et al., 2018; Bisk et al., 2020). We evaluate model performance across three subtasks, reporting individual accuracy scores for each. To obtain a comprehensive understanding of models' associative abilities, we compute the weighted average accuracy across

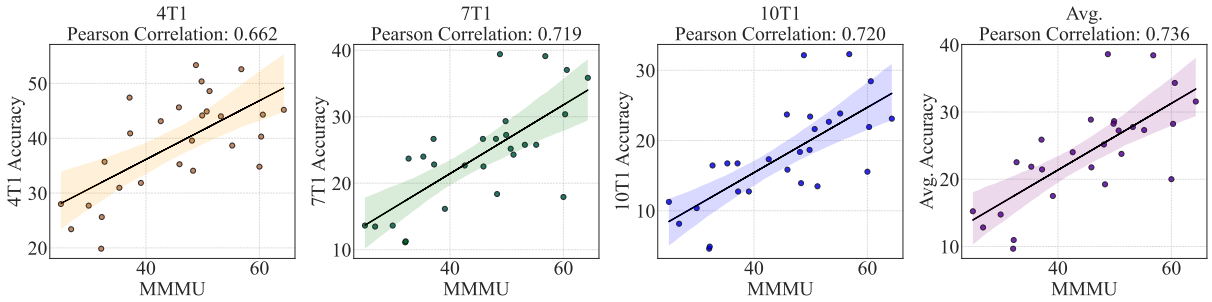

Figure 3: Linear regression lines illustrate the relationship between scores of association and cognition. mT1 represents association performance on mT1 tasks, while Avg. indicates the overall performance. The X-axis represents the cognition scores, and the Y-axis represents the association scores. The Pearson Correlation quantifies the correlation between association and cognition. The shaded area indicates the data points within a 95% confidence interval around the regression line.

Model	Size	4T1	7T1	10T1	Avg.
GPT-4o-mini	-	34.81	17.93	15.56	20.01
InternVL2-40B	40B	38.67	25.78	23.85	27.32
InternVL2-26B	26B	44.89	25.19	21.63	27.25
InternVL2-8B	8B	48.59	24.30	13.48	23.77
InternVL2-4B	4B	34.07	18.37	13.93	19.25
Qwen-VL-Max (0809)	-	45.19	35.85	23.11	31.56
Qwen-VL-Plus (0809)	-	44.00	25.78	22.67	27.77
LLaVA-OneVision-72B	72B	52.59	39.11	**32.30**	38.43
LLaVA-OneVision-7B	7B	**53.33**	**39.41**	32.15	**38.60**
LLaVA-Next-72B	72B	44.15	27.26	23.41	28.64
LLaVA-Next-34B	34B	39.56	26.67	18.37	25.17
LLaVA-Next-mistral-7B	7B	30.96	24.00	16.74	21.87
MiniCPM-V 2.6	8B	50.37	29.33	18.67	28.26
MiniCPM-Llama3-V 2.5	8.5B	45.63	26.67	23.70	28.87
MiniCPM-V 2	2.8B	47.41	26.67	16.74	25.89
MiniCPM-V	3B	40.89	22.81	12.74	21.46
Yi-vision	-	40.30	30.37	21.93	28.24
Yi-VL-34B	34B	35.26	22.52	15.85	21.77
Yi-VL-6B	6B	31.85	16.15	12.74	17.52
CogVLM-17B-chat	17B	19.85	11.11	4.59	9.67
CogVLM2-19B-chat	19B	43.11	22.67	17.33	24.02
VisualGLM-6B	8B	27.70	13.63	10.37	14.76
MiniGPT-4-v2	8B	28.00	13.63	11.26	15.24
MiniGPT-4 (VicunaV0-13B)	13B	23.41	13.48	8.15	12.83
mPLUG-Owl3	8B	34.22	23.70	19.26	23.59
mPLUG-Owl2	8.2B	35.70	23.70	16.44	22.53
OTTER-Image-MPT7B	7B	25.63	11.26	4.89	10.96
Gemini-1.5-pro	-	44.30	37.04	28.44	34.33
human	-	100.00	100.00	100.00	100.00
random	-	22.67	13.63	10.67	13.94

Table 2: The results of the evaluation. mT1 represents the score for a task with m options and one correct answer. The Avg. column represents the weighted average score, calculated by the formula: $Avg. = \frac{(4\times 4T1+7\times 7T1+10\times 10T1)}{4+7+10}$.

the three subtasks due to their different difficulty levels. The weight assigned to an mT1 task is m.

To ensure a fair evaluation, all models are tested using a unified prompt template within a two-shot demonstration setup. This setup is designed to stimulate associative reasoning across models.

6 Results

Overall Performance. The main results are presented in Table 2. LLaVA-OneVision 7B and LLaVA-OneVision 72B achieve the top-2 performances across the three tasks and on the overall average score, with only minor differences between them. However, all models do not achieve high scores and their scores decrease as the number of distractors increases. These results underscore the challenges posed by the benchmark and indicate that mitigating ambiguity does not necessarily make the benchmark too easy for models. Moreover, models still exhibit a significant gap compared to human performance. This behavior highlights a discrepancy between models and humans in associative tasks. Notably, human participants consistently demonstrate robust associative abilities even as task difficulty increases, whereas model performance declines under similar conditions. The inconsistency observed across model responses indicates that models lack a strong ability to form meaningful connections between seemingly unrelated concepts—the core of association. This limitation further underscores the gap between current models and human-level associative capabilities.

Correlation between cognition and association. Previous studies (Martinsen, 1994, 1993; Kaufmann, 1979; Runco and Chand, 1995; Huang et al., 2025d; Mednick, 1962) have highlighted that human cognitive abilities (knowledge, perception, and reasoning) form the foundation of associative ability (e.g., association requires adequate knowledge, the ability to perceive each concept clearly, and the capacity to reason from one concept to another.). These theories suggest that for both humans and models, association and cognition should exhibit at least some degree of correlation.

To analyze the correlation between cognition

and association, we use scores from the MMMU benchmark (Yue et al., 2024) as an indicator of models' cognitive ability, as MMMU is specifically designed to evaluate comprehensive cognitive skills of MLLMs. The Pearson Correlation Coefficient (Sedgwick, 2012; Schober et al., 2018; Cohen et al., 2009) is computed to quantify the linear correlation between cognition and association. As shown in Fig. 3, all Pearson Correlation Coefficient values exceed 0.66, with three exceed 0.71, demonstrating a strong positive correlation between cognition and association. Most data points fall within the confidence interval, further supporting the linear relationship. Consistent with the perspective, the results show a strong correlation between association and cognition, aligning well with human cognitive theories. This further validates that our benchmark is well designed in accordance with human cognitive principles.

7 Analysis

To analyze the impact of ambiguity and the effectiveness of our method, we focus on the following four research questions (RQs):

- RQ1: How does the presence of internal ambiguity affect the evaluation outcomes?
- RQ2: How does the presence of external ambiguity affect the evaluation outcomes?
- RQ3: Does DINO-v2 (Oquab et al., 2023; Darcet et al., 2023) focus on geometric shapes of masks?
- RQ4: Does our algorithm effectively mitigate external ambiguity in distractor selection process?

The ambiguity affects association evaluation (RQ1&RQ2). To investigate the impact of ambiguity, we first sample classes and their associative questions with four options to construct an original validation set (Ori), a subset of AssoCiAm. Subsequently, we derive two validation sets from Ori: one with **internal ambiguity** (Int) and one with **external ambiguity** (Ext).

For Int, we invite human experts to replace the original answers in Ori with unreasonable answers and ensure that options remain dissimilar, to explicitly introduce internal ambiguity. For instance, the original answer, 'A. Goblet,' is replaced with 'A. Blackboard,' and it still serves as the correct answer. Due to external ambiguity avoidance in the Ori, this process creates Int, where internal ambiguity is the primary source of ambiguity.

To study the effects of internal ambiguity comprehensively, we assess performance of models in

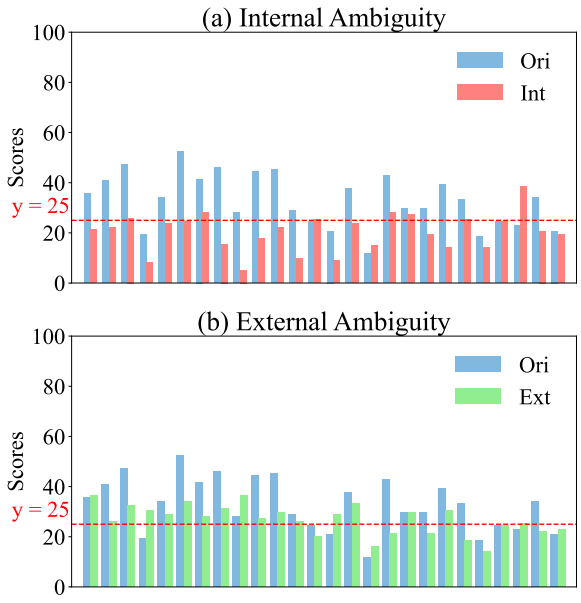

Figure 4: The evaluation results of models. Ori indicates the scores evaluated on Ori, Int indicates the scores evaluated on Int and Ext indicates the scores evaluated on Ext. The red line $y = 25$ indicates the expected score of answering questions randomly.

Section 6 on Ori and Int. As shown in Fig. 4 (a), most of the scores in Int tend to approximate random-choice performance levels, compared to those in Ori. This finding suggests that internal ambiguity renders models incapable of distinguishing answers based on questions. Additionally, the result of scores nearing random performance highlights the lack of faithfulness in these evaluations, as models fail to answer questions through meaningful associations and answer questions randomly.

Figure 5: Examples of a series of scale masks.

For Ext, we invite human experts to add new classes to explicitly construct external ambiguity. To achieve this, the human experts follow this procedure: observe the masks in the **original mask set** (Section 4.2), use association to infer what the masks resemble, and generate ambiguous distractors based on these associations. By replacing the distractors in Ori with these ambiguous distractors, we construct Ext. In this set, all distractors are ambiguous, while all answers remain reasonable,

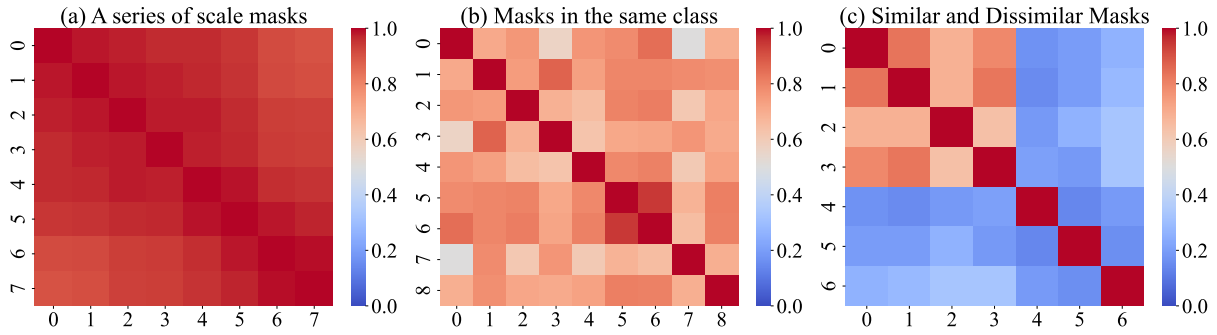

Figure 6: Heatmaps of similarity among masks. A similarity score of 1 indicates a high degree of similarity, while a score of 0 indicates a high degree of dissimilarity. The number in axis denotes the id of masks. (a) illustrates the similarity among a series of scale rabbit masks. (b) illustrates similarity among different chrysanthemum masks. In (c), id 3 is the shark mask; id 0 to 2 are masks of ambiguous distractors; id 4 to 6 are masks of unambiguous distractors. See Appendix A.5 for more results.

making external ambiguity the primary risk.

As shown in Fig. 4 (b), we use the same models as in RQ1 and observe that most of the scores in Ext also approach random-choice levels compared to Ori. This result indicates that models are confused by the choices due to all options being plausible answers and fail to answer questions correctly, even though they make meaningful associations.

DINO-v2 calculate the similarity among masks (RQ3). To assess whether DINO-v2 primarily relies on geometric shapes, we scale masks to generate a series of versions, as shown in Fig. 5, and calculate their pairwise similarity. The result, presented in Fig. 6 (a), shows high similarity values, indicating that DINO-v2 primarily focuses on geometric shapes.

Next, we evaluate whether DINO-v2 produces consistent similarity representations for masks within the same semantic class, despite minor geometric variations. Human experts are invited to select representative masks from filtered set (Section 4.2), following the **key principle** outlined in Section 4.2. DINO-v2 is then employed to calculate the similarity among these masks belonging to the same class. As shown in Fig. 6 (b), the similarity among these representative masks is high and exceeds threshold 50%, indicating that DINO-v2 exhibits robustness to small shape differences within the same class.

Finally, we examine whether DINO-v2 can capture shape-level similarity across different semantic classes. To this end, human experts gather representative masks of ambiguous distractors from RQ2 to create an **extension set**. We sample classes and calculate similarity among their own masks, the masks of their ambiguous distractors, and the masks of their unambiguous distractors. The results, illustrated in Fig. 6 (c), show that similarity among shark and its ambiguous distractors is high, whereas the similarity scores between shark and its unambiguous distractors are much lower. Furthermore, the similarity between shark and its ambiguous distractors is significantly higher than that between shark and its unambiguous distractors. These results indicate that DINO-v2 effectively measures geometric similarity between masks from both similar and dissimilar classes.

Overall, our findings demonstrate that DINO-v2 predominantly encodes geometric shape information, enabling robust assessment of visual similarity across a wide range of mask classes.

The algorithm circumvents external ambiguity (RQ4). In Section 4.2, we invited human experts to refine the masks and images to ensure the avoidance of internal ambiguity. So, we only analyze the effectiveness of the algorithm in Section 4.3. In this experiment, we use the extension set (RQ3) and manually label ambiguous distractors for each class in the original mask set (Section 4.2). For each class in the original set, we perform two processes: randomly selecting distractors from the extension set, repeated ten times, and applying the algorithm to select distractors, also repeated ten times. An answer and its corresponding distractors form a multiple-option set. We then count the number of these multiple-option sets that contain ambiguous distractors. As shown in Table 3, 15% of the randomly selected multiple-option sets overall include ambiguous distractors. For each option quantity, the proportion of randomly selected multiple-option sets containing ambiguous distractors increases as the number of op-

tions grows. In contrast, none of the algorithm-selected sets contain any ambiguous distractors. These results clearly highlight the effectiveness of the algorithm in avoiding external ambiguity.

Quantity	Random	Algorithm
4	9.60%	0.00%
7	17.20%	0.00%
10	18.00%	0.00%
overall	15.00%	0.00%

Table 3: Comparison of distractor selection methods: random selection and algorithmic selection. Quantity refers to the number of options in a multiple-option set. Random represents the overall proportion of multiple-option sets with ambiguous distractors when selected randomly, while Algorithm represents the proportion when selected via the algorithm.

8 Conclusion

This paper identifies the inherent answer ambiguity in association evaluation and validates its significant and unavoidable impact on evaluation reliability. To address this issue, we propose AssoCiAm, a benchmark constructed through a hybrid computational method designed to mitigate ambiguity. Our evaluation of MLLMs on AssoCiAm reveals that these models still exhibit a noticeable gap compared to human performance in associative tasks. Furthermore, we find that associative ability is closely related to cognitive capability, highlighting the importance of advancing both aspects in future MLLM development.

Limitations

While our tasks focus on association based on object shapes to reflect this ability from a specific perspective, we aim to expand association evaluation into a more comprehensive and broader framework in the future. Additionally, during the evaluation process, differences in training methods and data among MLLMs may result in varied interpretations of images and options. This could lead to models selecting the correct answer by excluding distractors through understanding the similar semantics among the distractors rather than directly solving the task as intended. Finally, our benchmark assigns a single correct answer to each question. However, considering that ambiguity arises from answer diversity, future work should explore more evaluation frameworks, such as multiple-correct-option or open-ended questions, which can not only mitigate the answer ambiguity but also preserve and leverage answer diversity serving as an intrinsic property of association evaluation.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No. 623B2099, No. 62206314 and Science and Technology Projects in Guangzhou under Grant No. 2024A04J4388.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kian Ahrabian, Zhivar Sourati, Kexuan Sun, Jiarui Zhang, Yifan Jiang, Fred Morstatter, and Jay Pujara. 2024. The curious case of nonverbal abstract reasoning with multi-modal large language models. *arXiv preprint arXiv:2401.12117*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Aseem Arora, Gaël Dias, Adam Jatowt, and Asif Ekbal. 2022. Transfer learning for humor detection by twin masked yellow muppets. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–7.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Roger E Beaty and Yoed N Kenett. 2023. Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7):671–683.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.

Mathias Benedek, Julian Jurisch, Karl Koschutnig, Andreas Fink, and Roger E Beaty. 2020. Elements of creative thought: Investigating the cognitive and neural correlates of association and bi-association processes. *NeuroImage*, 210:116586.

Mathias Benedek, Tanja Könen, and Aljoscha C Neubauer. 2012. Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):273.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.

Honghua Chen and Nai Ding. 2023. Probing the creativity of large language models: Can models produce divergent semantic association? *arXiv preprint arXiv:2310.11158*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. 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*, pages 24185–24198.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. 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)*, pages 320–335.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. 2024. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*.

Luis Fabricio Góes, Marco Volpe, Piotr Sawicki, Marek Grses, and Jacob Watson. 2023. Pushing gpt's creativity to its limits: Alternative uses and torrance tests.

David E Goldberg. 2013. *Genetic algorithms*. pearson education India.

Joy Paul Guilford. 1967. The nature of human intelligence. *New York: Macgraw Hill*.

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

John H Holland. 1992. Genetic algorithms. *Scientific american*, 267(1):66–73.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for gui agents. *Preprint*, arXiv:2312.08914.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. 2025a. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. *arXiv preprint arXiv:2503.10657*.

Zhongzhan Huang, Guoming Ling, Shanshan Zhong, Hefeng Wu, and Liang Lin. 2025b. Minilongbench: The low-cost long context understanding benchmark for large language models. *arXiv preprint arXiv:2505.19959*.

Zhongzhan Huang, Shanshan Zhong, Pan Zhou, Shanghua Gao, Marinka Zitnik, and Liang Lin. 2025c. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhongzhan Huang, Shanshan Zhong, Pan Zhou, Shanghua Gao, Marinka Zitnik, and Liang Lin. 2025d. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhongzhan Huang, Pan Zhou, Shuicheng Yan, and Liang Lin. 2023. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. *Advances in Neural Information Processing Systems*, 36:70376–70401.

Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14(1):3440.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Lonneke van der Plas, and Duygu Ataman. 2024a. Evaluating morphological compositional generalization in large language models. *arXiv preprint arXiv:2410.12656*.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024b. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023. Brainteaser: Lateral thinking puzzles for large language model. *arXiv preprint arXiv:2310.05057*.

Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv preprint arXiv:2310.12874*.

Geir Kaufmann. 1979. The explorer and the assimilator: A cognitive style distinction and its potential implications for innovative problem solving. *Scandinavian Journal of Educational Research*, 23(3):101–108.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601.

Koen Kraaijveld, Yifan Jiang, Kaixin Ma, and Filip Ilievski. 2024. Columbus: Evaluating cognitive lateral understanding through multiple-choice rebuses. *arXiv preprint arXiv:2409.04053*.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: What else influences visual instruction tuning beyond data?

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023b. Mimic-it: Multi-modal in-context instruction tuning.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023c. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024c. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024d. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. 2024e. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, and Manmohan Chandraker. 2024. Aide: An automatic data engine for object detection in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14695–14706.

Mingfu Liang, Ying Wu, et al. 2023. Toa: Task-oriented active vqa. *Advances in Neural Information Processing Systems*, 36:54061–54074.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Tali R Marron, Yulia Lerner, Ety Berant, Sivan Kinreich, Irit Shapira-Lichter, Talma Hendler, and Miriam Faust. 2018. Chain free association, creativity, and the default mode network. *Neuropsychologia*, 118:40–58.

Ø yvind Martinsen. 1993. Insight problems revisited: The influence of cognitive styles and experience on creative problem solving. *Creativity Research Journal*, 6(4):435–447.

Øyvind Martinsen. 1994. The effect of individual differences in cognitive style and motives in solving insight problems. *Scandinavian Journal of Educational Research*, 38(2):83–96.

Sarnoff Mednick. 1962. The associative basis of the creative process. *Psychological review*, 69(3):220.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Seyedali Mirjalili and Seyedali Mirjalili. 2019. Genetic algorithm. *Evolutionary algorithms and neural networks: theory and applications*, pages 43–55.

Melanie Mitchell. 1998. *An introduction to genetic algorithms*. MIT press.

Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. 2023. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. Ambipun: Generating humorous puns with ambiguous context. *arXiv preprint arXiv:2205.01825*.

Gustaw Opiełka, Hannes Rosenbusch, Veerle Vijverberg, and Claire E Stevenson. 2024. Do large language models solve arc visual analogies like people do? *arXiv preprint arXiv:2403.09734*.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision.

Andrei Popescu-Belis, Alex R Atrio, Bastien Bernath, Étienne Boisson, Teo Ferrari, Xavier Theimer-Lienhardt, and Giorgos Vernikos. 2023. Gpoet: a

language model trained for rhyme generation on synthetic data. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Mark A Runco and Ivonne Chand. 1995. Cognition and creativity. *Educational psychology review*, 7:243–267.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Philip Sedgwick. 2012. Pearson's correlation coefficient. *Bmj*, 345.

Changhao Shi, Haomiao Ni, Kai Li, Shaobo Han, Mingfu Liang, and Martin Renqiang Min. 2023a. Exploring compositional visual generation with latent classifier guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023b. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2025. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *Preprint*, arXiv:2408.04840.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *Preprint*, arXiv:2311.04257.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403.

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

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11748–11756.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. 2025. Accelerating multimodal large language models by searching optimal vision token reduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29869–29879.

Shanshan Zhong, Shanghua Gao, Zhongzhan Huang, Wushao Wen, Marinka Zitnik, and Pan Zhou. 2024a. Moextend: Tuning new experts for modality and task extension. *arXiv preprint arXiv:2408.03511*.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024b. 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*, pages 13246–13257.

Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. 2023. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 567–578.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Details for collecting images

To collect images, whose classes are applicable to the associative task, we first merge similar classes into a single category to address subtle differences between certain classes, such as cocks and hens. Additionally, classes that are uncommon for human recognition or lack distinct shapes are excluded.

A.2 Details for extracting masks

As the segmentation process operates in batches, SAM (Kirillov et al., 2023) generates multiple masks per image, which may include both class-relevant masks and redundant masks. For instance, Fig. 7 (b) illustrates a representative mask from Fig. 7 (a), while Fig. 7 (c, d) shows redundant masks.

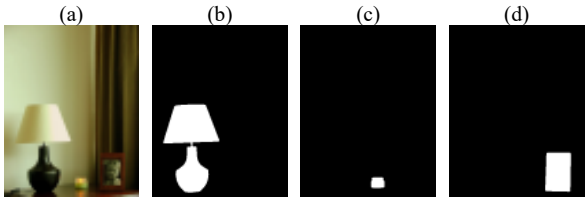

Figure 7: An example of masks extracted by SAM (Kirillov et al., 2023). (a) represents the original image of a table lamp. (b) shows the mask corresponding to the table lamp, which is the accurately segmented mask. (c) and (d) illustrate redundant masks extracted from other objects in the image (a), namely, a hand cream and a photo frame.

To eliminate redundant masks, the following approach is adopted: for each image, model CLIP (Radford et al., 2021) is utilized to rank a series of images from overlaying segmented masks onto the original image, as illustrated in Fig. 8. The mask corresponding to the top-ranked image is then selected as the class-relevant one. This approach ensures the accurate extraction of masks per image.

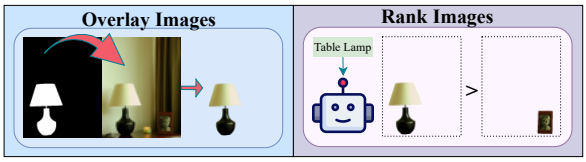

Figure 8: An overview of eliminating redundant masks

A.3 Details for the Genetic Algorithm

The optimization of the function 2 is an NP-hard problem. To address this, we adopt a genetic algorithm, a classic randomized search method, that is widely used for solving NP-hard problems. Below are the main details of our implementation:

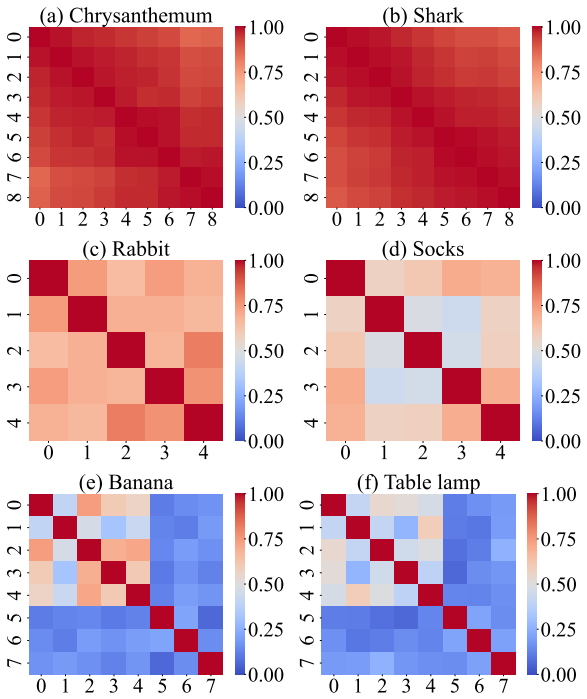

Figure 9: (a, b) are similarities among a series of scale masks. The masks belong to chrysanthemum or shark. (c, d) are similarities among masks belonging to the same class. In (e, f), id 0,1,2,3 indicates masks of ambiguous distractors, id 5,6,7 are masks unambiguous distractors. id 4 is the masks of correct answer.

Encoding and Initialization. Each solution is represented as a binary string, where 1 indicates a question is selected and 0 indicates it is not. We initialize the population with 50 randomly generated selection combinations.

Parent Selection. We use a tournament selection strategy to choose parent pairs based on fitness scores, which is calculated by function 2.

Searching. For each parent pair, we apply single-point crossover with a probability of 0.8, i.e.,

$$\text{offspring} = \text{parent1}[:\text{point}] + \text{parent2}[\text{point}:].$$

With a probability of 0.2, the offspring is directly copied from one of the parents without crossover. Each offspring undergoes bit-flip mutation with a probability of 0.05. To satisfy the selection constraints, we randomly flip excess 0s to 1s or vice versa in the offspring if needed. This process is repeated for a fixed number of generations to optimize the objective function 2.

A.4 Algorithm in Section 4.3

We summarize our algorithm and the process of our algorithm is shown in Alg. 1

Algorithm 1 External Ambiguity Avoidance

Input: The masks set V and answer mask v_0. The regularization parameter λ. The number of masks n and choices m.

Output: The optimal set of distractors and answers V'^*

```
Initialize graph G: e_ij = DINO-v2 (v_i, v_j)
Construct objective function F(G') in Eq. (2)
Optimal F(G') via GA(n, m, G, v_0, λ)
                              ▷ GA: Genetic Algorithm
Select the G'* = argmin_{G'} F(G')
Decode V'* from G'* =< V'*, E'* >
return V'*
```

A.5 DINO-v2 calculate the similarity

As shown in Fig. 9 (a, b), the similarities among a series of scale masks are consistently high. Additionally, Fig. 9 (c, d) demonstrates that similarities among masks within the same class are predominantly above 70%, with the lowest values still close to 50%. Fig. 9 (e, f) highlights a clear distinction between similar and dissimilar masks: similar masks exhibit higher similarity scores, while dissimilar masks have notably lower scores. Although in Fig. 9 (f), the similarities between the correct answer masks and ambiguous distractor masks are not as high as in other cases, they remain significantly higher than those between the correct answer masks and unambiguous distractor masks. These results underscore that DINO-v2 primarily focuses on the geometric shapes of masks.

A.6 λ balances $S(G')$ and $\sigma^2(G')$

We analyze the effect of λ using the original mask set, as selecting an appropriate λ for constructing AssoCiAm depends on the properties of the original set. A series of λ values are tested by sampling some masks as v_0 and applying our algorithm to calculate the corresponding $S(G')$ and $\sigma^2(G')$. As shown in Table 4, increasing λ results in an upward trend in $S(G')$ and a downward trend in $\sigma^2(G')$. With a well-chosen λ, $S(G')$ increases moderately, while $\sigma^2(G')$ decreases significantly, achieving a balanced trade-off between the two metrics.

A.7 Prompt

Prompt template used in evaluation is shown below:

You are a test subject participating in an associative ability test. Your task is as follows: Given an image and a question, you need to fully utilize your associative abilities and choose the best option from the given choices to answer the question. The given image and multiple-choice question are in the INPUT, specifically:
1.IMAGE: A provided image
2.QUESTION: A given question related to the image, requiring association to answer
3.OPTION: The given options, from which you need to select the best option based on your associative thinking to answer the question
Your OUTPUT should include ANSWER, formatted as follows:
A/B/C/D
One of the letters A, B, C, or D, representing the best option for answering the question.
Please strictly follow this format to output the answer for the question posed in the INPUT

Associative thinking helps people understand things from different perspectives, enabling them to quickly adapt to new situations and improve problem-solving abilities. In the field of LLM, associative ability is also a very important metric.

V_0	λ	$S(G')$	$\sigma^2(G')$	$\Delta S(G')$	$\Delta\sigma^2(G')$
12	0	**0.1278**	0.0062	0.00%	0.00%
	0.5	**0.1278**	0.0062	0.00%	0.00%
	1	0.1309	0.0029	2.39%↑	-52.72%↓
	2	0.1309	0.0029	2.39%↑	-52.72%↓
	5	0.1350	**0.0012**	5.62%↑	-79.83%↓
17	0	**0.1423**	0.0087	0.00%	0.00%
	0.5	**0.1423**	0.0087	0.00%	0.00%
	1	0.1468	0.0034	3.10%↑	-60.31%↓
	2	0.1468	0.0034	3.10%↑	-60.31%↓
	5	0.1557	**0.0010**	9.42%↑	-89.02%↓
18	0	0.**1503**	0.0081	0.00%	0.00%
	0.5	0.1510	0.0032	0.45%↑	-60.63%↓
	1	0.1510	0.0032	0.45%↑	-60.63%↓
	2	0.1535	**0.0011**	2.17%↑	-86.44%↓
	5	0.1535	**0.0011**	2.17%↑	-86.44%↓
25	0	**0.1601**	0.0233	0.00%	0.00%
	0.5	0.1610	0.0036	0.53%↑	-84.56%↓
	1	0.1610	0.0036	0.53%↑	-84.56%↓
	2	0.1610	0.0036	0.53%↑	-84.56%↓
	5	0.1712	**0.0009**	6.93%↑	-96.27%↓

Table 4: The values with different λ. V_0 indicates the id of the answer. $\Delta S(G')$ and $\Delta\sigma^2(G')$ are the the change of $S(G')$ or $\sigma^2(G')$ relative to that calculated with $\lambda = 0$. $\Delta S(G') = \frac{(S(G') - S(G')_{\lambda=0})}{S(G')_{\lambda=0})}$; $\Delta\sigma^2(G') = \frac{(\sigma^2(G') - \sigma^2(G')_{\lambda=0})}{\sigma^2(G')_{\lambda=0})}$

Now, we will test your associative thinking. Based on the given image, fully utilize your associative abilities to think about the given question, and select the best option from the given choices as your answer. Then, output your answer in the specified format.

Below are some examples. Please strictly follow the format.

Example 1:
INPUT: {
"IMAGE": ,
"QUESTION": "Observe the mountain in the image and use your associative thinking to consider what it resembles. Please choose the best option from the following choices to answer the question.",
"OPTION": "(A) Fish tank
(B) Tissue
(C) Eye
(D) Water bottle"
}
OUTPUT: C

Example 2: INPUT: {
"IMAGE": ,
"QUESTION": Carefully observe the shape of the lake in the image. What does it remind you of using your associative thinking? Please choose the best option from the following choices to answer the question.",
"OPTION": "(A) Cat
(B) Phone
(C) Rock
(D) Alarm clock"
}
OUTPUT: A

Referencing the above examples, based on the following latest INPUT information, use associative thinking to analyze the given image, and select the best option from the provided choices to answer the question. Strictly output the result in the same format as shown in the examples.

INPUT: {
"IMAGE": ,
"QUESTION": "<question>",
"OPTION": "<option>"
}
OUTPUT: