Unveiling the Tapestry of Consistency in Large Vision-Language Models

Yuan Zhang[1,2]**, Fei Xiao**[2]**, Tao Huang**[3]**, Chun-Kai Fan**[1]**, Hongyuan Dong**[2]**,**
Jiawen Li[2]**, Jiacong Wang**[2,4]**, Kuan Cheng**[1]**, Shanghang Zhang**[1]*__**, Haoyuan Guo**[2]*

[1] School of Computer Science, Peking University [†]   [2] ByteDance Inc
[3] The University of Sydney   [4] School of Artificial Intelligence, UCAS
https://github.com/foundation-multimodal-models/ConBench

Abstract

Large vision-language models (LVLMs) have recently achieved rapid progress, exhibiting great perception and reasoning abilities concerning visual information. However, when faced with prompts in different sizes of solution spaces, LVLMs fail to always give consistent answers regarding the same knowledge point. This inconsistency of answers between different solution spaces is prevalent in LVLMs and erodes trust. To this end, we provide a multi-modal benchmark ConBench, to intuitively analyze how LVLMs perform when the solution space of a prompt revolves around a knowledge point. Based on the ConBench tool, we are the first to reveal the tapestry and get the following findings: (1) In the discriminate realm, the larger the solution space of the prompt, the lower the accuracy of the answers. (2) Establish the relationship between the discriminative and generative realms: the accuracy of the discriminative question type exhibits a strong positive correlation with its Consistency with the caption. (3) Compared to open-source models, closed-source models exhibit a pronounced bias advantage in terms of Consistency. Eventually, we ameliorate the consistency of LVLMs by trigger-based diagnostic refinement, indirectly improving the performance of their caption. We hope this paper will accelerate the research community in better evaluating their models and encourage future advancements in the consistency domain.

1 Introduction

Recently, benefiting from notable advancements in large language models (LLMs) [1; 25; 2], the realm of large vision-language models (LVLMs) has undergone a revolutionary transformation. These novel LVLMs [18; 24; 3; 8; 15; 13] try to combine visual signals with textual semantics and spark cognitive brilliance across modalities. Although LVLMs can generate high-quality responses to task prompts, we discover that for correctly answered cases, simply modifying the prompt will result LVLMs in providing contradictory responses. In Figure 1 (a.2), LLaVA-7B [18] properly describes the picture as "It is a man wearing a dinosaur costume.", but when prompted "Is the dinosaur played by humans? Please answer yes or no.", it responds with "No, they are dinosaurs". The above phenomenon of Inconsistency is widely observed across mainstream LVLMs, and a preliminary study was conducted only on LLMs [14]. In practice, in contrast to the fixed patterns of questions, designed in existing multimodal benchmarks, the users tend to pose questions in arbitrary ways. Therefore, it is necessary to ensure the LVLMs in predicting correct and **consistent** answers, even when faced with various formats of queries.

---

*Correspondence to: Shanghang Zhang and Haoyuan Guo.
†State Key Laboratory of Multimedia Information Processing.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

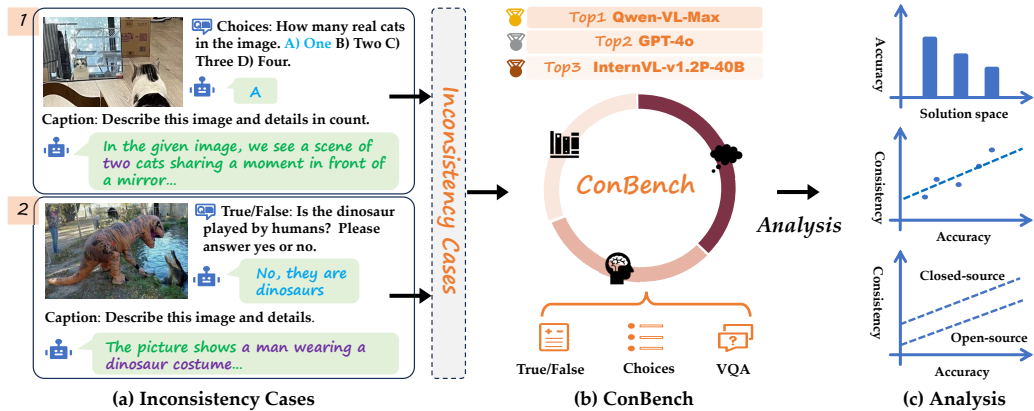



Figure 1: **Here is the overview of our paper.** Part (a) indicates two examples of Inconsistency between discriminative answers and generative captions, where the answers marked in blue contradict the answers marked in purple. Part (b) shows the Consistency evaluation method Conbench and its discriminative top three leaderboard. Part(c) reveals the main three findings built upon ConBench.

However, there are currently no benchmarks or research studies that specifically focus on evaluating the Consistency of LVLMs responses. These single-prompt type evaluation approaches [12; 10; 28; 21; 6] lead to a disconnect between benchmark accuracy and real-world user practical experience.

Based on the above observations, we systematically introduce a **Con**sistency **Bench**mark dubbed ConBench, to estimate the capabilities of LVLMs more thoroughly via diverse question formats. It consists of 1,000 public pictures, and each was manually selected from four multimodal benchmarks [10; 12; 28; 21]. Apart from the original discriminative prompt, we constructed two additional discriminative types of questions[3] by ChatGPT/GPT-4 [1]. Notably, three types of questions of each case are around the **same** knowledge point. Besides, every set is accompanied by a generative question without ground truth. Consequently, ConBench serves as an evaluation tool that observes the Consistency performance of LVLMs and surpasses the limitations of previous assessments.

Furthermore, grounded on the ConBench, we conduct an in-depth analysis and visualization of Consistency on 14 popular LVLMs. In a nutshell, the conclusions of noteworthy insight are threefold:

C1 In the **discriminative** question-answering (QA) domain: (1) A decrease in LVLMs accuracy as the prompt's solution space increases. (2) Instances of erroneous yet consistent answers are scarce.

C2 Extended to the **generative** domain, we establish a connection between discriminative and generative domains by the perspective of Consistency. (1) As the solution space of discriminative questions expands, the Consistency between its answer and caption grows stronger. (2) The accuracy of discriminative answer and its Consistency with the caption exhibit a positive correlation.

C3 Closed-source models exhibit a pronounced bias advantage in terms of Consistency, compared to open-source models. This provides an alternative perspective to demonstrate why closed-source models, despite sometimes having lower accuracy, offer a better user experience in practical applications.

Eventually, leveraging the insights gained from our theoretical discoveries, we enhance the caption performance of LVLMs without any additional costs associated with training. Specifically, we construct discriminative prompts based on the low-confidence words in the answers of LVLMs, forcing the LVLMs to introspect. Then, through iterative refinement in multiple rounds of question-answering, the quality of LVLMs' captions gets an impressive achievement (e.g., our method improves the LLaVA-NeXT-34B [19] by 9.1% and MiniGemini-34B [15] by 9.6% on metric[C] in Sec. 3.4).

In summary, to the best of our knowledge, we are the first to propose a Consistency evaluation method and conduct a comprehensive analysis of Inconsistency in LVLMs. We hope this paper serves as a catalyst for further exploration, and look forward to the community applying the above findings to polish up the usability and practicality of large vision-language models.

---

[3]e.g., Multiple-choice questions and limited VQA questions are generated for MME benchmark.

2

2 Related Work

Large Vison Language Models With the impressive success of large language models (LLMs) [1; 25; 2; 4; 29], recent studies work on generative large vision-language models (LVLMs) [18; 24; 3; 8; 15; 27; 30] to improve multimodal comprehension and generation through utilizing the strong generality of LLMs. Built upon the CLIP [23] image encoder which is somewhat aligned with the language modality, current LVLMs typically utilize vast image-text pairs to connect the vision encoder and LLM, enabling LLM to receive and understand visual content. For instance, LLaVA [20] directly connects the vision encoder and LLM with MLPs, showing proficiency in multi-modal dialogues. Subsequent works have further enhanced LVLMs by improving the multi-modal instruction data [18; 27; 5] and designing novel modules [3; 4; 26] for more sufficient modality alignment.

Conventional Multimodal Evaluation A multitude of public multimodal benchmarks, such as MME [10], SeedBench [12], and MMBench [21], further advance objective evaluation of LVLMs by only constructing True/False questions or multiple-choice questions, where the absence of diverse question types causes instability. In addition, their objective metrics solely emphasize the LVLM's accuracy, disregarding its robustness and security. The above issues can lead to a situation where some LVLMs have lower accuracy in evaluation results but provide a better user experience. To systematically assess the comprehensive capability of LVLMs, we propose a simple and efficient evaluation approach that relies on checking the Consistency between different kinds of prompts.

Inconsistency in LLMs A amount of prior work has been conducted on investigating Inconsistency in LLMs. [14] is the first to find the Inconsistency phenomenon in question-answering and validator tasks and define GV-consistency. Besides, it leverages consistency pair for training to improve LLMs' performance. While [17] utilizes Consistency to check for hallucination detection in LLMs, a logic consistency-based method that involves logic-related questions and answers. Compared to LLMs, Inconsistency in LVLMs is more likely to occur due to the additional visual modality, which deserves further exploration.

3 ConBench

We propose a novel multimodal evaluation pipeline named ConBench to comprehensively assess LVLMs. The ConBench has a total of 4K questions on 1K images and corresponding 3K discriminative ground truths, guaranteeing evaluation quality in terms of the quantity and diversity of questions. In Sec. 3.1, we present the generation of ConBench and the construction pipeline for prompts. Sec. 3.2 introduces the hierarchical core capabilities and discusses the design philosophy. Sec. 3.3 and 3.4 describe the evaluation strategy for scoring various types of answers.

Figure 2: **Overview of 19 evaluation detailed categories in ConBench.**

3.1 Data Generation Process

Image Filter We manually chose 1K images from four high-quality multimodal benchmarks: MME [10], SeedBench [12], MMBench [21], and MMMU [28]. MME is a true/false question type, while SeedBench and MMBench cover comprehensive multiple-choice questions. Meanwhile, MMMU emphasizes the knowledge level. The criteria for the image filter include: (1) resolution is more than 224×224 (2) the image rarely occurs in the mainstream training dataset (e.g., COCO [16] and Cityscapes [9]) (3) There are more than 3 foreground objects in the image. The above criteria ensure the quality of content in images.

Prompt Construction Each image is accompanied by its original discriminative prompt, and we constructed two extra discriminative questions. Therefore, a case owns three discriminative prompts (true/false, multiple-choice and limited VQA questions) with a generative caption prompt around the same knowledge point. Firstly, we modified the original prompts whose answers can be directly inferred from the text instead of the image, to force LVLMs to utilize information from the visual features. Next, we employed GPT/GPT-4 to generate the extra discriminative types of questions,

3

which were then subjected to the manual review, and the proposed prompt is listed in Figure 3. Finally, to avoid bias in the LVLMs that may affect the evaluation results, the true/false questions have a 50% distribution for both correct and wrong ground truths. For the multiple-choice questions, each option (e.g., A, B, C, D) has an equal probability distribution of 25% for being the correct answer. Notably, to ensure an accurate evaluation parser, limited VQA questions are subject to certain restrictions, like specifying the word count and answer format (e.g., fractions / abbreviations / numbers).

3.2 Hierarchical Core Capabilities

The ConBench comprises three core capabilities, arranged in ascending order of difficulty, namely: Sensation, Cognition, and Knowledge, with nineteen fine-grained dimensions shown in Figure 2.

[Easy Mode] Sensation: What you see is what you get. We assume that sensation is the most fundamental expertise of LVLMs, and it is the "eye" of the LLMs. While perceived questions appear simple and basic, they are nonetheless essential. Therefore, this capability accounts for 50% of the ConBench. Count, color, optical character recognition (OCR) and scene categories focus on subtle details, while poster, attribute recognition and position types emphasize the overall picture.

"Prompt Construction" Prompt :
You are a question expert. Give you a [Discriminative type] question, and you should generate two other kinds questions. The [Discriminative type] question is that [Original Prompt]. Based on the [Discriminative type] question above, a [The other discriminative type] about the [Category] with following answer, and a VQA question about [Category] with following answer are generated for the same knowledge point.

Figure 3: **The prompt for generation of discriminative questions.** Please zoom in to view.

[Medium Mode] Cognition: Go beyond the surface. The cognitive process needs the model to integrate visual and language modalities: observing the content of an image, combining it with the text of question, and retrieving knowledge from within the LLMs. It is more challenging than the single sensation task. This section constitutes 26% of the ConBench, including numerical calculation, code inference, text translation, math, cross-instance reasoning and attribute reasoning categories.

[Hard Mode] Knowledge: Master the art of synthesis and integration. Mastering professional knowledge is an essential pathway for next-generation LVLMs to become Expert AGI, as it requires a higher level of understanding of images and the application of expert knowledge. We carefully selected knowledge from diverse and extensive fields, such as celebrities, chemistry, physics, biology, art and geography. This part takes up 24% of the total, and functions as the upper limit of ConBench.

3.3 Results Parser

For true/false questions, we first extract the "yes" and "no" from the answer. If both of them are absent, the answer would be considered as "none". Then, we strictly compare the extracted answer with the ground truth. If they match exactly, the true/false response is considered correct.

When parsing the outcome of multiple choices, we derive the choice label (e.g., A, B, C, D) from it. If successful, utilize this as the prediction and match the ground truth. If not, we will not proceed with further extracting the answers. Because in each prompt of choices, we specified that only one letter needs to be answered. Doing so would be unfair to LVLMs that excel in following instructions.

We still utilize character matching for the answer of limited VQA instead of GPTs. On one hand, we have taken strict formatting constraints on the prompts. For instance, in physics and math, there are restrictions on answering with fractions (e.g., 1/2), while in geography at the city level. On the other hand, the cost of the GPT's judgment is high and the waiting time is delayed. Specifically, the parser is based on the Average Normalized Levenshtein Similarity (ANLS) [22], where the threshold τ is set to 0.95 and $M = N = 1$. When parsed result $s > 0.4$, we consider the answer to be exactly right.

3.4 Multidimensional Evaluation Metric

Here we provide two evaluation metrics, each from the perspective of discriminative and generative domains, aiming to provide a more comprehensive understanding of LVLMs consistency. The former does not rely on AI tools and quickly produces Consistency results among discriminative responses via Sec. 3.3, primarily evaluating the knowledge. The latter employs GPT to indirectly assess the quality of captions, by judging the consistency between discriminative responses and captions.

4

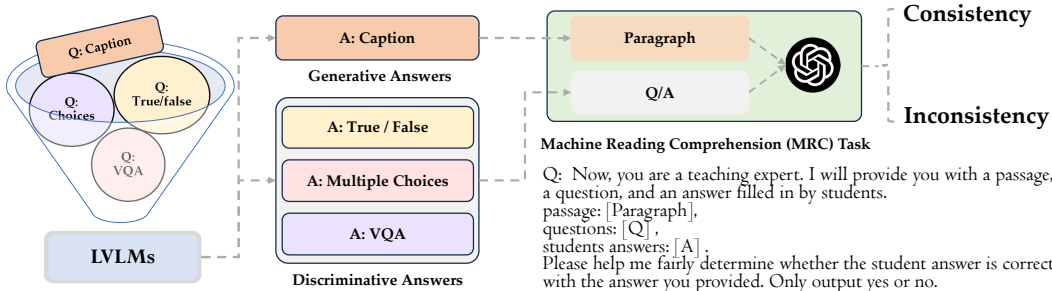



Figure 4: **The pipeline of judging Consistency between caption and discriminative answers via GPT/GPT4.** Please zoom in to view the prompt.

Table 1: **Evaluation[D] of mainstreams series of LVLMs on ConBench.** The detailed results of the Sensation, Cognition, and Knowledge core capabilities are listed below. T, C, and V represent true-false, multiple-choice, and limited VQA questions, respectively. The ranking can be found below the respective numbers. †: Due to safety considerations, GPT-4V declined to answer the celebrity category.

Method	ConScore[D]	Sensation				Cognition				Knowledge			
		T	C	V	Con	T	C	V	Con	T	C	V	Con
Closed-source Vision Language Models													
GPT-4V† [1]	29.20_6	80.4	79.0	61.7	48.3	68.8	53.2	39.9	20.4	63.1	57.2	30.0	14.2
GPT-4-Omni [1]	$\underline{35.70}_2$	89.2	79.4	64.4	$\underline{55.0}$	71.8	62.8	44.9	27.8	64.7	61.7	39.7	$\underline{23.3}$
Gemini-Pro-Vision [24]	25.00_{10}	85.2	60.7	63.4	39.3	71.8	45.0	44.2	15.1	65.0	51.4	39.7	15.8
Gemini-Ultra-Vision [24]	33.10_4	78.9	78.6	66.3	50.3	68.1	58.5	47.9	$\underline{28.5}$	62.9	62.2	44.7	19.7
Qwen-VL-Plus [3]	28.10_7	82.7	74.9	60.4	45.0	64.2	41.7	30.8	16.3	63.6	54.2	33.3	15.8
Qwen-VL-Max [3]	$\mathbf{37.00}_1$	86.4	80.7	65.4	$\mathbf{56.3}$	72.9	51.4	51.3	28.1	68.3	58.6	38.9	$\mathbf{24.2}$
7B Vision Language Models													
LLaVA-v1.5-7B [18]	16.60_{14}	79.3	56.8	44.3	28.3	51.4	33.5	15.8	4.7	61.7	44.4	16.9	7.8
Qwen-VL-Chat [3]	26.40_9	81.0	79.6	54.2	39.0	55.0	46.3	33.2	13.5	60.3	54.2	28.9	14.7
∼ 13B Vision Language Models													
LLaVA-v1.5-13B [18]	24.00_{11}	82.9	77.1	49.6	39.5	53.6	37.8	20.1	10.4	65.6	50.3	17.2	9.7
MiniGemini-13B [15]	21.80_{13}	81.9	69.7	52.8	39.3	51.9	38.2	21.1	6.9	52.8	36.7	17.5	9.2
InternVL-v1.5-26B [7]	31.40_5	85.6	84.8	65.0	54.3	59.7	58.6	44.4	19.4	58.1	55.8	25.3	12.2
∼ 34B Vision Language Models													
LLaVA-NeXT-34B [19]	27.70_8	82.4	81.7	55.6	43.6	50.7	47.5	25.6	9.9	60.4	56.1	27.8	12.8
MiniGemini-34B [15]	23.00_{12}	80.8	76.8	48.2	39.7	36.9	30.7	18.9	6.0	58.1	42.3	20.8	8.2
InternVL-v1.2P-40B [8]	34.70_3	83.7	83.2	66.6	53.4	74.2	67.6	57.1	$\mathbf{34.9}$	72.2	58.3	28.6	13.6

Discriminative Domain Evaluation Metric We define the **ConScore[D]** as that: when **all three discriminative** types of questions within the same case are answered correctly, the model gets one point. The maximum score is 1000 points. The final format is presented as a percentage (%).

Generative Domain Evaluation Metric Due to the high variability in captions, it is not possible to calculate Consistency based on character matching alone. Therefore, we rely on GPT/GPT4 for judgment. The judging process and the constructed prompts are shown in Figure 4. We formulate it as a machine reading comprehension task. We manually sample the judgment results, and GPT4 achieved an accuracy rate of 95%, which is reliable and trustworthy. Next, we define the **ConScore[C]** as the average score of Consistency between the caption and the other three discriminative responses.

4 Analysis

4.1 Evaluation Results

In this section, 6 closed-source and 8 open-source representative LVLMs with varying sizes and architectures are evaluated on our Consistency benchmark, including GPT-4V [1], GPT4-Omni [1], Gemini-Vision [24], Qwen-VL series [3], LLaVA series [18; 19], MiniGemini series [15] and InternVL series [8]. The evaluation results on ConBench are listed in Table 1 and 2. In the metric[D]

5

Table 2: **Evaluation of Consistency between caption and three discriminative types of answer on ConBench.** The "rank diff" means the difference between ConScore[D] and Score[C]. The Con[X] is the Consistency ratio between discriminative answer type X and caption. The "ordered" represents whether Con[T] < Con[C] < Con[V] is in its line.

Method	Rank Diff	ConScore[C]	Con[T]	Con[C]	Con[V]	Ordered
Closed-source Vision Language Models						
GPT-4V [1]	↑ **3**	55.6_3	51.20	56.50	59.10	Y
GPT-4-Omni [1]	↑ **1**	$\mathbf{62.2}_1$	58.00	62.50	66.10	Y
Gemini-Pro-Vision [24]	↑ **1**	48.4_9	43.30	45.20	56.80	Y
Gemini-Ultra-Vision [24]	–	54.6_4	47.80	55.20	60.70	Y
Qwen-VL-Plus [3]	–	50.2_7	47.10	49.10	54.30	Y
Qwen-VL-Max [3]	↓ **1**	$\underline{58.4}_2$	54.30	58.00	62.90	Y
7B Vision Language Models						
LLaVA-v1.5-7B [18]	–	38.4_{14}	39.20	36.60	39.50	N
Qwen-VL-Chat [3]	↓ **2**	48.0_{11}	42.00	50.80	51.30	Y
~ 13B Vision Language Models						
LLaVA-v1.5-13B [18]	↓ **1**	44.4_{12}	41.50	45.80	46.00	Y
MiniGemini-13B [15]	–	41.7_{13}	38.80	42.90	43.30	Y
InternVL-v1.5-26B [7]	↓ **1**	50.9_6	44.50	53.90	54.20	Y
~ 34B Vision Language Models						
LLaVA-NeXT-34B [19]	↓ **2**	48.3_{10}	46.00	52.20	46.80	N
MiniGemini-34B [15]	↑ **4**	49.6_8	56.80	48.00	44.10	N
InternVL-v1.2P-40B [8]	↓ **2**	53.7_5	49.80	55.50	55.80	Y

1, Qwen-VL-Max [3] secures the top position, leading the second-place GPT4-Omni [1] by a margin of 1.3%. The InternVL-v1.2P-40B [8] performs best in the open-sourced community, especially in cognition capability. The LLaVA series did not make it to the top ten. In the metric[C], the newest GPT4-Omni [1] leads the leaderboard, which is the only model that surpasses 60. It has a significant advantage over the second-place model Qwen-VL-Max [3], with a gap of 3.8. We observed that although the GPT series slightly underperforms Qwen-Max in metric[D], it significantly outperforms the Qwen series in metric[C], which aligns with our actual user experience. Actually, ConScore[C] provides an alternative quality description of captions, because higher recall and precision rates usually match better Consistency. Besides, rankings of LVLMs show a slight variation between metric[C] and metric[D]. The GPT series models claim better performance of caption generation.

4.2 Discriminative Domain

To investigate what causes the Inconsistency between different types of prompts, we first conduct analyses on the discriminative domain to compare the performance differences. We summarize our findings into the following facts:

Fact 4.2.1 (Inconsistency in Accuracy). *The accuracy of the answer decreases as the solution space of the discriminative prompt increases.*

As shown in the columns of "T", "C", and "V" in Table 1, the accuracy decreases as the solution space expands in all core capabilities. For instance (e.g., the Sensation of GPT-4-Omni), the double-choice true-false questions achieve an accuracy of 89.2, whereas the accuracy for multiple-choice and VQA questions on the same case declines to 79.4 and 61.7, respectively. This is understandable, as the number of potential choices increases, the difficulty in identifying the correct answer also rises.

Figure 5: The confidence and logits of answers of LLaVA-13B and MGM-13B.

Fact 4.2.2 (Inconsistency in Wrong Answers). *Cases of erroneous yet consistent answers are scarce.*

6

Table 3: **The Consistency between multiple choices and VQAs, including both correct and wrong.** Each case is picked up in order from top to bottom from the 14 LVLMs in Table 1.

Con[Correct] (%)	35.00	39.90	31.20	34.10	41.60	37.60	39.40	29.30	39.70	25.90	28.70	22.20
Con[Wrong] (%)	0.30	0.40	0.20	0.50	0.40	0.50	0.30	0.40	0.20	0.50	0.10	0.20

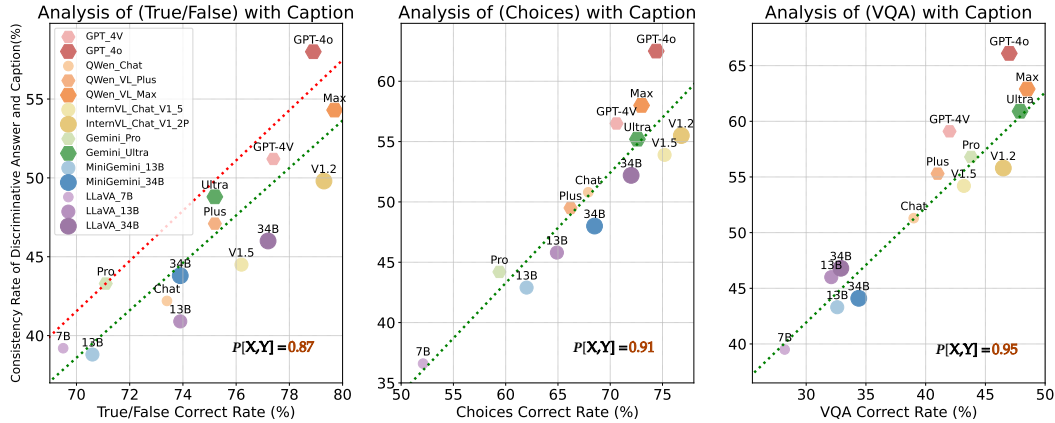



Figure 6: Visualization of the relationship between the correct rate of discriminative answer and its Consistency with the caption on different **answer types**.

We analyze the answers that fail in all three question types and find that, despite all resulting in incorrect predictions, they do not demonstrate a consistent understanding of the same images, leading to distinct answers. For example, we calculated the proportion of consistent incorrect responses in VQA and multiple-choice questions. We found a very small consistency, and it did not exceed 0.50% across the entire benchmark. This indicates that the models struggle to interpret the visual content uniformly, revealing significant variability in their failure modes.

Fact 4.2.3 (Inconsistency in Confidence). *The confidence of models in their answers reveals signs of inconsistent and incorrect predictions.*

Taking Fact 4.2.1 and Fact 4.2.2 into account, we perform a deeper analysis of the model's predictions by measuring their confidence in the answers. We use the predicted probabilities and logits of the answer tokens to represent confidence (see Appendix B for details). As summarized in Figure 5, we measure the average probabilities and logits of the correct and incorrect answers[4], respectively. The three types of questions share similar confidence levels for the correct answers. However, for the incorrect answers, their confidence levels vary significantly with a clear trend: the larger the solution space, the smaller the confidence. This analysis provides crucial insights for our method in enhancing the consistency and accuracy of LVLMs, which we will further discuss in Sec. 5.

4.3 Generative Domain

Next, we extend our attention to the generative domain. Based on Consistency, we first build a bridge between the discriminative and generative domains. We consolidate our findings as the below facts:

Fact 4.3.1 (Inconsistency to Generative Answers). *As the solution space of discriminative questions increases, the Consistency between their answers and generative answers increases.*

As indicated in the last column of Table 2, "Ordered" means Con[T] < Con[C] < Con[V]. The answers of all closed-source models and most open-source models adhere to this pattern. Here is the theoretical explanation. Assume the distribution for the generative domain (Caption) is $\mathcal{S}$, and the sample space of $\mathcal{S}$ is W. For the discriminative domain, the sample space is limited to W', which only contains some candidates from W. Assume the model handles the discriminative domain by creating another distribution $\mathcal{S}'$ according to $\mathcal{S}$ and W'. Then the total variation distance (TVD) [11]

---

[4]MGM_Y and LLaVA_Y mean the correct, while MGM_N and LLaVA_N represent the wrong.

7

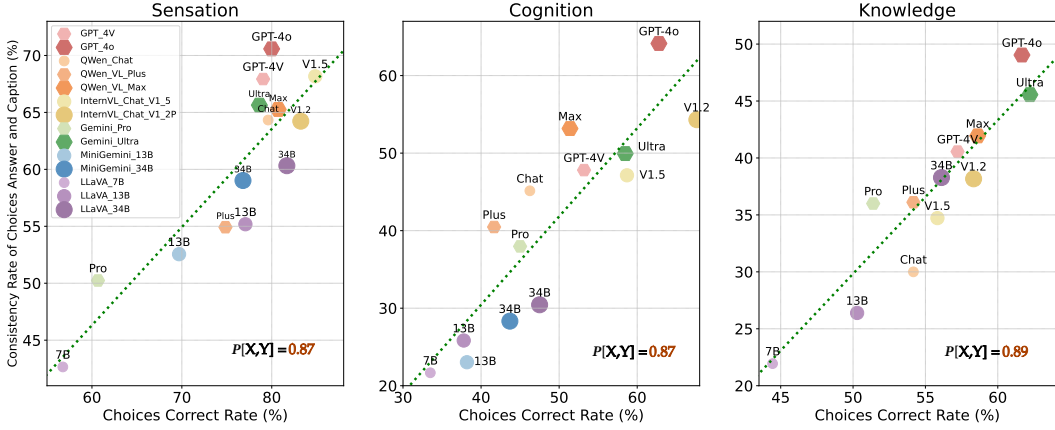



Figure 7: Visualization of the relationship between the correct rate of discriminative answer and its Consistency with the caption on different **capability types**.

between $\mathcal{S}$ and $\mathcal{S}'$ is

$$\frac{1}{2}\left\|\mathcal{S}-\mathcal{S}'\right\|_1.\tag{1}$$

This becomes larger when $|W \setminus W'|$ becomes larger. For instance, if the model creates $\mathcal{S}'$ by simply doing reject sampling[5], then

$$\frac{1}{2}\left\|\mathcal{S}-\mathcal{S}'\right\|_1 = Pr[\mathcal{S} \in W \setminus W'].\tag{2}$$

It is obvious that when W' is more "different" from W, the distance will be larger.

Fact 4.3.2 (Connection between Discriminative and Generative Domain). *The accuracy of the discriminative answer exhibits a strong positive correlation with its Consistency with the generative.*

As shown in Figure 6 and 7, we conduct visualizations for all tested LVLMs: The vertical axis represents the accuracy of their discriminative answers, while the horizontal axis represents the consistency of the answers with caption. Figure 6 displays the distribution across different question types, while Figure 7 illustrates the distribution across different core capabilities. The green lines represent a fitted linear equation. Additionally, we utilize the Pearson coefficient $P[X,Y]$ to quantitatively analyze the degree of linear correlation, and the 6 coefficients in the above figures are all more than 0.85.

4.4 Consistency Bias

Fact 4.4.1 (Consistency Bias). *Closed-source models exhibit a pronounced bias advantage on Consistency, compared to open-source models.*

When we fit a linear regression to all evaluated models and get the green line in Figure 6 (a):

$$\mathcal{L}_1 : y = kx + b,\tag{3}$$

where x is the accuracy, and y means Consistency between its answer and caption. We found that the majority of open-source models lie below this line, while closed-source models lie above it. In other words, at the same level of accuracy, the responses from closed-source models tend to exhibit better consistency with their captions. So we fit a linear regression to closed-source models and get the red line. The line they reside on has a higher bias b_c (e.g., $b_c - b = 3.24$ in Figure 6 (a)), which aligns with our experience where closed-source models provide more comprehensive and reliable answers.

5 Trigger-based Diagnostic Refinement

In light of the previous findings, we summarize two key insights: (1) LVLMs exhibit higher accuracy when operating within a narrower discriminative solution space; (2) Incorrect answers are usually associated with significantly lower confidence and logits. Consequently, we propose a simple but efficient method dubbed Trigger-based Diagnostic Refinement (TDR) to ameliorate the generation skill of LVLMs without any additional training. The proposed pipeline is presented in Figure 8.

---

[5]e.g, projecting or clustering $\mathcal{S}$ to W'

8

Table 4: **Results on LLaVA-34B and MiniGemini-34B via Trigger-based Diagnostic Refinement.**

Method	ConScore[C]	Con[T]	Con[C]	Con[V]
LLaVA-NeXT-34B [19]	48.3	46.00	52.20	**46.80**
+ TDR	**57.4** (9.1 ↑)	**69.10**	**57.40**	45.70
MiniGemini-34B [15]	49.6	56.80	48.00	44.10
+ TDR	**60.2** (9.6 ↑)	**76.10**	**53.80**	**50.80**

Method We start by making the LVLM generate a caption, with each word accompanied by its corresponding probability. Next, uninformative words are dropped based on their parts of speech, and we only keep nouns, adjectives and quantifiers. When the remaining words with probabilities below a threshold τ (we set $\tau = 0.85$ here), trigger subsequent diagnostic processes. Since low probabilities of words indicate a lack of confidence, we formulate True/False discriminative questions to force the LVLM to self-verify (e.g., Is there {cat} in the picture?). The self-diagnostic prompt and its response will be drafted into a new prompt, which is fed back into the LVLM to generate a higher-quality caption.

Results We carried out experiments on the LLaVA-NeXT-34B and MiniGemini-34B and evaluated them on the metric[C] of ConBench. The experimental results are detailed in Table 4. Notably, the LLaVA-NeXT-34B sees an improvement of 9.1 points, while

Figure 8: **The Trigger-based Diagnostic Refinement pipeline.**

the MiniGemini experiences an overall enhancement of 9.6 points. Although our approach primarily employs True/False questions for self-verify, there is still a noticeable improvement in ConScore[C]. Hence, our method effectively boosts the quality of captions by triggering the model to self-check.

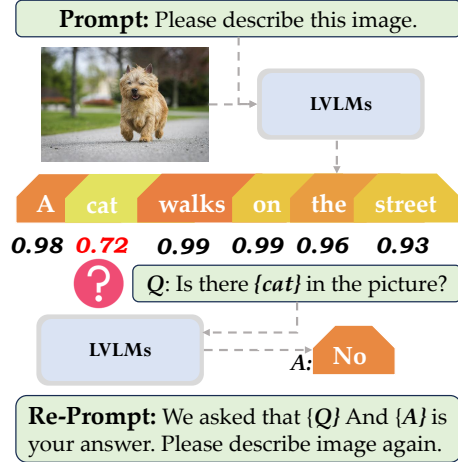
In theory, we can further construct multiple discriminative questions for the caption, enabling the model to verify multiple elements within the caption. Additionally, the process can be iterated multiple rounds, leading to ongoing enhancements in the quality of the generated output. Our method is a simplified implementation of the above approaches.

6 Conclusion

In this study, we investigate the Consistency issues in large vision-language models (LVLMs). Consistency reflects the overall ability of LVLMs, as it not only requires LVLMs to provide correct answers but also demands sufficient confidence in their knowledge point, regardless of the type of question encountered. We first introduce the ConBench, a benchmark that fills the gap in assessing Consistency. It includes 1K images with 4K prompts and two evaluation metrics: ConScore[D] and ConScore[C]. Then, our findings shed light on the nature of Consistency in LVLMs according to the ConBench. We observe that as the solution space of a prompt increases, the accuracy of the answers tends to decrease. Besides, we establish a relationship between the discriminative and generative realms, highlighting the importance of Consistency between the discriminative answer and caption. Furthermore, we discover that closed-source models exhibit a bias advantage over open-source models in terms of consistency. Finally, we propose a solution by forcing LVLMs to self-think, where a discriminative prompt is constructed via uncertain words in the caption. Our method makes the quality of LVLMs' captions an impressive achievement. We believe that our research contributes to the evaluation of LVLMs and encourages future advancements for achieving Consistency in LVLMs.

7 Acknowledgment

Shanghang Zhang was supported by the National Science and Technology Major Project (No. 2022ZD0117800).

9

References

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

[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] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[5] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[6] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[7] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[8] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[10] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[11] J. Kennedy and M. Quine. The total variation distance between the binomial and poisson distributions. *The Annals of Probability*, pages 396–400, 1989.

[12] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[13] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[14] X. L. Li, V. Shrivastava, S. Li, T. Hashimoto, and P. Liang. Benchmarking and improving generator-validator consistency of language models. *arXiv preprint arXiv:2310.01846*, 2023.

[15] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] Z. Lin, S. Trivedi, and J. Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

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

[19] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.

[21] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

10

[22] M. Mathew, D. Karatzas, and C. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[23] 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*, pages 8748–8763. PMLR, 2021.

[24] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[26] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[27] 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.

[28] 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. *arXiv preprint arXiv:2311.16502*, 2023.

[29] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[30] Y. Zhang, C.-K. Fan, J. Ma, W. Zheng, T. Huang, K. Cheng, D. Gudovskiy, T. Okuno, Y. Nakata, K. Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

11

Appendix

A Showcases of Trigger-based Diagnostic Refinement

MGM-34B

Image. Count-000000236721.jpg

Original Caption. *The image contains a variety of objects, which can be counted as follows: one tray, one cake, one bowl of sauce, one carton of milk, **one banana**, one glass of orange juice, **one cup of coffee**, and one sugar packet. The objects are arranged neatly on the tray, which is placed on a table.*

Refined Caption *The image contains a total of **two bananas**, which are positioned on the right side of the tray. The bananas are yellow, indicating that they are ripe, and they are the only fruits visible in the image. They are placed next to a carton of milk, ice cream, a cake, **two cups of coffee**, a glass of orange juice, which are a common accompaniment in a breakfast setting.*

LLava-Next-34B

Image. Attribute Reason-1724416_2902943546.jpg

Original Caption. *The image captures a vibrant live music scene. At the center of the frame, a musician is seated on a stool, engrossed in playing a red electric guitar. He is dressed casually in a white t-shirt and blue jeans, adding to the relaxed atmosphere of the performance. The stage, adorned with a colorful banner, is set against a backdrop of a brick wall, lending an urban feel to the setting. The banner prominently features the text "DOWNTOWN 5", possibly indicating the name of the band or the event. The musician is **surrounded by other band members, each engrossed in their own instruments**, contributing to the overall harmony of the performance. The stage is well-equipped with a microphone and a speaker, essential for amplifying the sound and reaching the audience. In the background, a crowd of people can be seen, their attention riveted on the performance.*

Refined Caption *The image captures a vibrant scene of a live music performance. At the center of the stage, a musician is seated on a stool, engrossed in playing an electric guitar. He is dressed in a casual white t-shirt, and his focus is evident as he strums the strings. The stage is set against a backdrop of a colorful banner that proudly proclaims "DOWNTOWN 5". The banner's bold colors contrast with the musician's white t-shirt, drawing attention to both the performer and the event. The stage **is equipped with various musical instruments**, including a drum set and a keyboard, suggesting a full band performance. A microphone stands ready, capturing the musician's performance for the audience. The stage is situated in a space that appears to be partially open to the elements, as suggested by the presence of a roof structure in the background. This setup indicates that the concert is being held in a venue that can accommodate outdoor or semi-outdoor performances, adding a unique ambiance to the event. Overall, the image portrays a dynamic and engaging live music event, with the musician at the heart of the performance, surrounded by his instruments.*

B Computation for Probabilities and Logits

For each model (MGM-13B and LLaVA-13B), we randomly selected 50 consistent and 50 inconsistent samples, respectively. The JSON samples are in support materials. To measure the LVLMs' response confidence accurately, we manually simplified the answers and maintained the words that are the most critical to the judgment. The specific principles are as listed:

For true/false questions, we only keep the word 'yes' or 'no' and their probabilities.

```
e.g., [Yes], there is a cat.
```

For multiple-choice questions, we only keep the choice labels (e.g., A, B, C, D) and their probabilities.

12

```
e.g., ([A]) Cats.
```

For limited VQA questions, we manually picked out keywords that matched ground truth from the answers, and computed the average probabilities of them as the final probability.

```
e.g., A [Cat] walks on the street.
```

C   Limitations

The introduced ConBench offers a new perspective on evaluating model performance through the consistency between multiple types of questions, providing a more comprehensive measurement and understanding of existing LVLMs. However, due to the distinct response forms of captions, assessing the consistency between captions and discriminative answers is judged by GPT, posing a risk of inaccurate evaluations. Besides, by delving deeper into our benchmark analysis, we propose trigger-based diagnostic refinement to improve the consistency and accuracy of LVLMs. This, however, introduces additional computational costs and is limited by the inherent capabilities of the LVLMs. Further improvements can be achieved by designing and training LVLMs with a focus on consistency.

D   Broader Impacts

Overall, this research has broader impacts on the evaluation, performance, fairness, and future development of LVLMs, fostering progress and advancements in the field of vision-language models.

Advancing Evaluation: The introduction of ConBench, a benchmark for assessing Consistency in LVLMs, fills a crucial gap in the evaluation of these models. This benchmark provides a standardized framework for measuring the performance and reliability of LVLMs across different prompts.

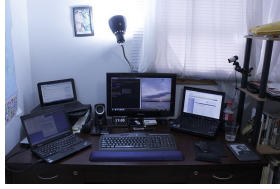
Novel Insights: we are the first to reveal the tapestry and get the following findings: (1) In the discriminate realm, the larger the solution space of the prompt, the lower the accuracy of the answers. (2) Establish the relationship between the discriminative and generative realms. (3) Compared to open-source models, closed-source models exhibit a bias advantage in terms of Consistency.

Inspiring Future Research: By contributing to the evaluation and understanding of Consistency in LVLMs, this research paves the way for future advancements in the field. It encourages researchers to explore new techniques, methodologies, and approaches to achieve higher levels of Consistency in LVLMs, ultimately pushing the boundaries of language and vision understanding.

E   Detailed Cases in ConBench

We have uploaded the ConBench dataset, including images and their prompts, to the Hugging Face platform. The dataset can be accessed at the following URL: https://huggingface.co/datasets/ConBench/ConBench. Here, we enumerate several representative cases from ConBench. Arrange in order from easy to difficult, respectively, based on sensation, cognition, and knowledge.

13

Count

T: *Are there three laptops in the picture? Please only answer yes or no.*
A: *Yes*

V: *How many laptops are depicted? Please answer with a number.*
A: *3*

C: *How many laptops are in the picture?*
A) *One*
B) *Two*
C) *Three*
D) *Four.*
Please choose an answer from [A, B, C, D].
A: *C*
Caption: *You are an expert in image description. You need to describe this picture with accurate object count information.*

Scene

T: *Is this photo taken in a place of corridor? Please answer yes or no.*
A: *Yes*

V: *Where was the photo taken? Answer within a word.*
A: *Corridor*

C: *Where was this photo taken?*
A) *Corridor*
B) *Park*
C) *Office*
D) *Street.*
Please choose an answer from [A, B, C, D].
A: *A*

Caption: *You are an expert in scene. You need to describe the scene in the picture.*

OCR

T: *Is the word in the logo \"angle's\"? Please answer yes or no.*
A: *No*

V: *What is the word in the referenced logo? Answer within a word.*
A: *Angie's*

C: *What is the word in the referenced logo?*
A) *Angie's*
B) *Angie*
C) *Agnes's*
D) *Anjie's.*
Please choose an answer from [A, B, C, D].
A: *A*
Caption: *You are an expert in image OCR. What is the word in the referenced logo?*

Position

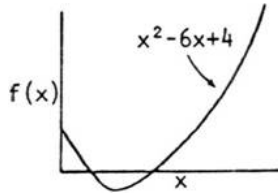
T: *Is the white couch positioned behind the glass coffee table? Please answer yes or no.*
A: *Yes*

V: *Where is the white couch located in relation to the glass coffee table? Please answer within 4 words.*
A: *behind the coffee table*

C: *What is the position of the white couch relative to the glass coffee table?*
A) *The couch is in front of the coffee table*
B) *The couch is to the right of the coffee table*
C) *The couch is to the left of the coffee table*
D) *The couch is behind the coffee table.*
Please choose an answer from [A, B, C, D].
A: *D*

Caption: *You are an expert in position. You need to describe this picture with accurate information about position about objects in the image.*

14

Math

T: *Is the function f(x) = x^2 - 6x + 4 convex? Please answer yes or no.*
A: *Yes*

V: *Is the function f(x) = x^2 - 6x + 4 convex or concave? Answer within a word.*
A: *Convex*

C: *Is the function f(x) = x^2 - 6x + 4 convex or concave?*
A) Convex
B) Concave
C) Neither
D) Both.
Please choose an answer from [A, B, C, D].
A: *A*
Caption: *You are an expert in math. Describe the concave and convex properties of the function.*

Code

```python
x = 10
if x < 20:
    print("Hello")
else:
    print("World")
```

T: *The image shows a python code. Is the output of the code 'World'? Please only tell me 'yes' or 'no' without any other words.*
A: *No*

V: *What is the output of the Python code? Please answer with one word without any other words.*
A: *Hello*

C: *What is the output of the Python code?*
A) Goodbye
B) Hello
C) Error
D) Nothing.
Choose one from the four letters [A, B, C, D] without any other words.
A: *B*

Caption: *You are an expert in code reasoning. What is the programming language of the code and tell me the output without any other words.*

Translation

美味的晚餐

T: *Is it appropriate to translate the Chinese in the image into English 'a delicious dinner' in the picture? Please answer yes or no.*
A: *Yes*

V: *Translate the Chinese in the picture to English. Answer within 3 words.*
A: *a delicious dinner*

C: *How to translate the Chinese in the image into English?*
A) a delicious dinner
B) traditional flavor
C) hamburger and chips
D) vintage taste.
Please choose an answer from [A, B, C, D].
A: *A*

Caption: *You are an expert in translation. You need to translate the Chinese in this picture into English.*

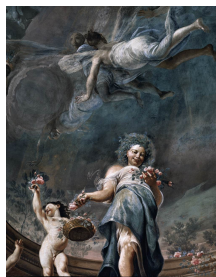
Attribute Reasoning

T: *Is the position and activity of the horses indicative of them engaging in a competitive or playful interaction rather than stationary activities like grazing or standing still? Please answer yes or no.*
A: *Yes*

V: *Based on the attributes and positions of the horses, which conclusion could be drawn? Answer within 5 words.*
A: *The horses are grazing*

C: *Based on the attributes and positions of the horses, which conclusion could be drawn?*
A) The horses are grazing
B) The horses are all standing still
C) The horses are playing together
D) The horses are racing.
Please choose an answer from [A, B, C, D].
A: *A*

Caption: *You are an expert in attribute reason, please describe this image in detail*

15

Artwork

T: *Does this artwork exist in the form of painting? Please answer yes or no.*
A: *Yes*

V: *What is the form of this artwork? Please answer within 3 words.*
A: *Painting*

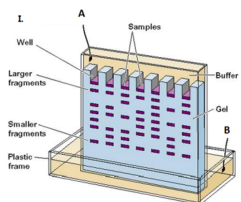
C: *In which form does this artwork exist?*
A) Sculpture
B) Painting
C) Digital Art
D) Performance Art.
Please choose an answer from [A, B, C, D].
A: *B*
Caption: *You are an expert in artwork. You need to describe this picture with accurate information about its title, actor, form and the location of display.*

Celebrity

T: *Is the actor inside the red bounding box called Hugh Jackman? Please answer yes or no.*
A: *Yes*

V: *Who is the actor identified inside the red bounding box? Please answer with a name within 3 words.*
A: *Hugh Jackman*

C: *Which actor is identified within the red bounding box?*
A) Tom Cruise
B) Robert Downey Jr
C) Hugh Jackman
D) Chris Hemsworth.
Please choose an answer from [A, B, C, D].
A: *C*
Caption: *You are an expert in image description. You need to describe this picture with accurate information about the actor.*

Biology

T: *Does letter A indicate the anode? Please answer yes or no.*
A: *No*

V: *The apparatus in figure is used for SDSPAGE (polyacrylamide gel electrophoresis). Which letter indicates the anode?*
A: *B*

C: *The apparatus in figure is used for SDSPAGE (polyacrylamide gel electrophoresis). Which letter indicates the anode?*
A) A
B) B
C) Neither
D) Unknown.
Please choose an answer from [A, B, C, D].
A: *B*
Caption: *You are an expert in biology. The apparatus in figure is used for SDSPAGE (polyacrylamide gel electrophoresis). Which letter indicates the anode?*

Landmark

T: *Is this a photo of Great Palace Mosaic Museum? Please answer yes or no.*
A: *No*

V: *In which city can you find the landmark shown in the picture? Answer within 2 words.*
A: *Utrecht*

C: *Which of the following landmarks is pictured in this photo?*
A) The Eiffel Tower in Paris, France
B) The Leaning Tower of Pisa in Pisa, Italy
C) The Dom Tower in Utrecht, Netherlands
D) Big Ben in London, England.
Choose one from the four letters [A, B, C, D].
A: *C*

Caption: *You are an expert in landmark. You need to describe this picture with accurate location information.*

16

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope, including a benchmark, three main analyses and a method to improve the caption of LVLMs.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We point out our limitations of the work in Appendix C.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

17

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our paper releases a dataset and evaluation method, and the experimental results are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

18

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the dataset and evaluation code after the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our paper only includes LVLMs evaluation and specifies all the test details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We investigate the robustness of our results in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

19

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The evaluation in our paper only needs an A100-80GB GPU.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss both potential positive societal impacts and negative societal impacts of the work in Appendix D.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

20

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper re-packages the public datasets, and provides both the original license and the license of the derived asset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

21

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new benchmark, and we created an anonymized URL in Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

22