Evaluating Linguistic Capabilities of Multimodal LLMs in the Lens of Few-Shot Learning

Anonymous ACL submission

Abstract

The linguistic capabilities of Multimodal Large Language Models (MLLMs) are critical for their effective application across diverse tasks. This study aims to evaluate the performance of MLLMs on the VALSE benchmark, focusing on the efficacy of few-shot In-Context Learning (ICL), and Chain-of-Thought (CoT) prompting. We conducted a comprehensive assessment of state-of-the-art MLLMs, varying in model size and pretraining datasets. The experimental results reveal that ICL and CoT prompting significantly boost model performance, particularly in tasks requiring complex reasoning and contextual understanding. Models pretrained on captioning datasets show superior zero-shot performance, while those trained on interleaved image-text data benefit from few-shot learning. Our findings provide valuable insights into optimizing MLLMs for better grounding of language in visual contexts, highlighting the importance of the composition of pretraining data and the potential of few-shot learning strategies to improve the reasoning abilities of MLLMs.

1 Introduction

Multimodal Large Language Models (MLLMs) demonstrate a remarkable ability to interpret both text and other modalities, such as images (Chen et al., 2022b; Alayrac et al., 2022; Tsimpoukelli et al., 2021; Awadalla et al., 2023; Laurençon et al., 2023; Li et al., 2023b). These models integrate visual and textual data, allowing them to perform a wide range of reasoning tasks effectively. Despite their impressive capabilities, optimizing these models through fine-tuning is resource-intensive and costly. To address these challenges, researchers have developed efficient data augmentation techniques and optimization algorithms (Huang et al., 2018; Falcon et al., 2020; Mou et al., 2020). Among these, few-shot learning techniques offer a promising solution by significantly reducing the costs associated with fine-tuning (Chen et al., 2023b; Tsimpoukelli et al., 2021; Wei et al., 2022; Wang et al., 2022).

Few-shot learning is an In-Context-Learning (ICL) strategy that enhances model performance by providing a small number of demonstration examples, introducing a specific context (Brown et al., 2020). This method allows the model to leverage its inherent knowledge, combined with the context provided, to solve complex tasks in various domains without specific prior training. Chain-of-Thought (CoT) (Wei et al., 2022) is, on the other hand, a prompting methodology which involves generating reasoning chains before providing the final answer. This strategy enables models to produce more accurate outputs, especially for tasks that require intermediate steps and reasoning, such as arithmetic and commonsense reasoning. Without these reasoning chains, models often fail when they respond with only the final answer.

Although the individual effects of few-shot ICL and CoT strategies have been studied in multimodal settings, their collective impact on the linguistic capabilities of MLLMs has not been previously explored. Understanding this impact is crucial, as it can significantly influence the development and deployment of more efficient and capable MLLMs. To address this gap, we utilize the VALSE (Vision And Language Structured Evaluation) benchmark (Parcalabescu et al., 2022). VALSE provides a comprehensive framework for evaluating the grounding of linguistic phenomena—from morphosyntax to semantics—in the visual modality. It includes six tasks: Existence, Plurality, Counting, Spatial Relations, Actions, and Coreference. These tasks are designed to test models' abilities to recognize existential quantifiers, semantic number, entity counting, spatial arrangements, actions, and pronominal coreference within images.

The VALSE benchmark is particularly well-suited for this study because it functions as a test-only benchmark without any training data, which

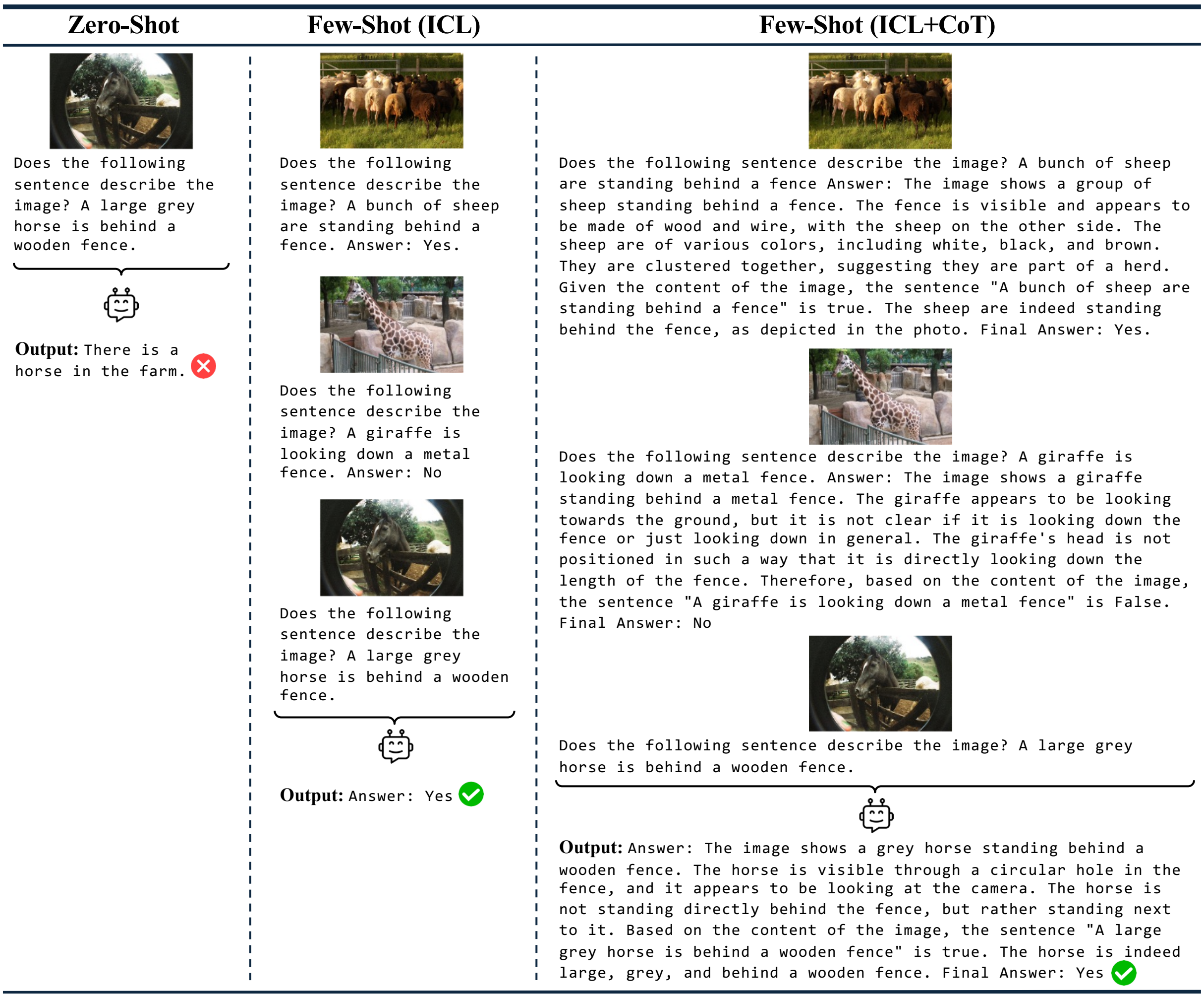


Figure 1: **Zero-Shot, Few-Shot ICL, and Few-Shot ICL+CoT Evaluation Comparison on the *Relations* Task.** In the Zero-Shot approach, the model incorrectly responds to the question. Few-Shot ICL, using prior examples, correctly identifies the horse behind a wooden fence. Few-Shot ICL+CoT, which is beneficial for tasks requiring intermediate reasoning steps, e.g. counting, relational understanding, and coreference resolution, also correctly identifies the horse by employing a detailed step-by-step reasoning process.

aligns perfectly with the ICL setting of our analysis. This allows us to evaluate the models purely based on their pre-existing capabilities and the provided context, without any additional fine-tuning or training. VALSE offers a detailed understanding of how models handle specific linguistic constructs and tasks, highlighting their strengths and deficiencies in visio-linguistic grounding. This makes it an ideal choice for examining the collective impact of ICL and CoT on the linguistic capabilities of MLLMs.

Using VALSE, we aim to investigate the effects of ICL and CoT on the performance of MLLMs. Our study makes the following contributions:

- We conduct a thorough evaluation of 14 different MLLMs on VALSE. This evaluation examines both zero-shot and few-shot settings, providing insights into how demonstration examples and reasoning chains influence model outputs.
- Our results indicate that using demonstration examples in the few-shot ICL setting enhances overall performance. Notably, examples similar to the query image-text pairs significantly boost performance compared to randomly selected examples, as in prior work (Liu et al., 2022; Luo et al., 2023).
- CoT proves highly effective for tasks requiring intermediate reasoning steps, such as counting, relational understanding, and coreference resolution. This highlights the potential of CoT in enhancing the reasoning capabilities of MLLMs.
- We demonstrate that models pretrained on captioning datasets such as MS-COCO (Lin et al., 2014), Conceptual Captions (Sharma et al., 2018), and LAION-5B (Schuhmann et al., 2022)

exhibit superior zero-shot performance compared to those trained on interleaved image-text datasets like Multimodal C4 (Zhu et al., 2023b) and OBELISC (Laurençon et al., 2023). However, with few-shot ICL strategies, lower-capacity models trained on interleaved image-text datasets can achieve similar or even better performance than the larger-capacity models trained on captioning datasets.

The subsequent sections of this paper are organized as follows: In §2, we provide a concise review of relevant literature. §3 outlines our evaluation strategy, offering comprehensive insights into our approach. In §4, we present our results. §5 gives our conclusions, summarizing the key findings and implications derived from this study. Lastly, in §6, we share the limitations of our study.

2 Related Work

In this section, we will explore the specifics of the recent MLLMs (§2.1), current ICL and CoT techniques (§2.2 and §2.3), examining their evolution, applications, and emerging approaches in this rapidly developing area.

2.1 Multimodal Large Language Models

Pretraining Strategies. Multimodal Large Language Models (MLLMs) require different pretraining datasets to support various capabilities. MLLMs often use datasets of image-text pairs due to several advantages: they are easy to use, provide a direct relationship between text and image, and include well-established, widely-used, and standardized datasets (Lin et al., 2014; Plummer et al., 2015; Schuhmann et al., 2022; Changpinyo et al., 2021). Conversely, interleaved image-text datasets (Zhu et al., 2023b; Laurençon et al., 2023; Li et al., 2023a; Zhao et al., 2024) create a context with multiple images and texts, enabling models to leverage this context to solve complex tasks. This approach allows models to tackle new challenges, such as narrating a series of images. Additionally, instruction-tuning datasets (Liu et al., 2024b; Chen et al., 2023a; Li et al., 2023a) are crucial for enhancing the flexibility and responsiveness of these models. By training on a diverse set of instructions paired with corresponding outputs, these datasets enable models to follow specific prompts more accurately and generalize better across different tasks. This improves the models' capabilities in zero-shot and few-shot learning scenarios, making them more versatile and effective for real-world applications where diverse and precise responses are needed.

Models. The development of MLLMs has significantly advanced, leveraging the capabilities of pre-trained autoregressive LLMs and sophisticated visual encoders to handle both text and visual inputs (Chen et al., 2023d; Dong et al., 2024; Zhu et al., 2023a; Bavishi et al., 2023). Notable examples include Flamingo (Alayrac et al., 2022), which has demonstrated remarkable performance across various vision-language tasks. This progress has led to the creation of open-weight models, fostering collaboration and accessibility in the field (Ye et al., 2023; Li et al., 2023b; Sun et al., 2023; Lu et al., 2024; Jiang et al., 2024; Awadalla et al., 2023; Research, 2024; Zhao et al., 2024). IDEFICS models(Laurençon et al., 2024; Laurençon et al., 2023) surpasses inference efficiency and stable training by leveraging pre-trained unimodal backbones. Similarly, Qwen-VL Chat (Bai et al., 2023), based on Qwen-7B, emphasizes fine-grained visual understanding and multilingual support, achieving state-of-the-art performance. In contrast, LLaVA-NeXT (Liu et al., 2024a), an improved version of LLaVA-1.5 (Liu et al., 2023b), employs a surprisingly powerful and data-efficient vision-language integration module, requiring only training a simple fully-connected projection layer on a modest dataset. While Qwen-VL trains specially designed visual resamplers on vast amounts of image-text paired data, LLaVA-NeXT achieves SOTA results with publicly available data, demonstrating efficiency and effectiveness in model design and training. MMICL (Zhao et al., 2024) addresses limitations in current models by efficiently handling multi-modal inputs, including relationships among multiple images and text-to-image references. By introducing a novel context scheme and a comprehensive multi-modal ICL dataset, MMICL significantly improves understanding of intricate text-image relationships and multi-image reasoning.

2.2 In-Context-Learning (ICL)

ICL was first developed for LLMs, where the goal is to provide a context with examples that the model can use to solve complex tasks (Brown et al., 2020). To transfer ICL for MLLMs, researchers train these models using interleaved image-text datasets. Selecting demonstration examples for ICL is critical, and the multimodal nature of MLLMs makes this selection more challenging, as it requires finding examples that are appropriate both textually and visually. Some studies suggest choosing examples based on their similarity to the query image-text pair (Alayrac et al., 2022; Chen et al., 2023b; Gui et al., 2021; Lin et al., 2022; Liu et al., 2021). However, research (Shukor et al., 2024) indicates that ICL can increase hallucinations and has a limited impact on improving image-text matching and instruction-following abilities. Additionally, Chen et al. 2023b found that while image similarity has a slight effect on model performance in Visual Question Answering (VQA) tasks, it raises questions about the overall effectiveness of ICL in multimodal settings. Several recent studies have begun to explore the In-Context Learning (ICL) capabilities of MLLMs. Shukor et al. (2024) examined the impact of ICL, Chain-of-Hindsight ICL (Liu et al., 2023a), and Self-Correcting ICL (Madaan et al., 2023) on factors such as hallucinations, abstention, compositionality, explainability, and instruction following. Zhao et al. (2024) evaluated the effect of ICL on the performance of a few MLLMs using standard vision-language datasets. In contrast, our study provides a more comprehensive analysis of the grounded linguistic capabilities of fourteen different MLLMs, focusing on ICL and CoT across the tasks available in the VALSE benchmark.

2.3 Chain-of-Thought (ICL) Prompting

Recent research shows that models perform better in reasoning, arithmetic, and commonsense tasks when they develop a reasoning process for their answers (Wei et al., 2022). This method, known as CoT, was initially introduced for LLMs. The core idea behind CoT is that by incorporating intermediate reasoning steps enhances the models' reasoning capabilities, leading to improved results. Models effectively utilize CoT when provided with context, and numerous studies have explored generating context for multimodal tasks to improve both the quality of demonstrations (Rubin et al., 2021; He et al., 2023) and the reasoning chain (Chen et al., 2022a; Wang et al., 2022). However, generating detailed, lengthy, and accurate context can be challenging for humans, which is where MLLMs come into play (Wang et al., 2024; Zhang et al., 2023). Additionally, CoT can be used without context, in a zero-shot manner, where the model is prompted with the phrase, "*Let's think step by step*" (Kojima et al., 2022). In multimodal setting, Mitra et al. (2024) investigated CoT, but their analysis involves generating a scene graph from the query image and use this graph in response generation. On the other hand, in our work, we use detailed CoT descriptions of the images in few-shot setting.

3 Evaluation Strategy

In this study, we investigate the zero-shot and few-shot capabilities of MLLMs through the VALSE benchmark (Parcalabescu et al., 2022). Previous work has separately examined ICL and CoT strategies in multimodal contexts (Mitra et al., 2024; Baldassini et al., 2024; Shukor et al., 2024). This study aims to integrate these approaches and provide a comprehensive analysis regarding how the recent MLLMs tackle with visio-linguistic grounding. Below, we begin by providing a brief review of the VALSE benchmark (§3.1). We then present the ICL methodology (§3.2) employed in our assessment of MLLMs, explaining our demonstration example selection process. Finally, we discuss the application of the CoT approach (§3.3) in our experimental analysis.

3.1 VALSE Benchmark

The VALSE (Parcalabescu et al., 2022) is a zero-shot foiling benchmark designed to assess the capabilities of MLLMs in integrating linguistic constructs with visual contexts. Providing a comprehensive evaluation framework, VALSE encompasses six distinct tasks that thoroughly probe the model's ability to bridge language and vision. These tasks include *Existence, Plurality, Counting, Spatial Relations, Actions*, and *Coreference,* each focusing on a critical linguistic phenomenon necessary for a deep understanding.

- *Existence* task examines the model's ability to identify the presence or absence of entities in an image. Models must differentiate between scenarios where objects exist or not within the visual context, focusing on existential quantifiers.
- *Plurality* task tests the model's understanding of singular and plural forms by requiring it to distinguish between images depicting single and multiple instances of objects. It assesses semantic number comprehension.
- *Counting* task challenges the model to accurately count the number of entities present in an image. The scenarios vary in complexity, demanding precise enumeration capabilities.
- *Spatial Relations* task evaluates the model's ability to recognize and interpret spatial relationships between objects in an image. It focuses on understanding the arrangements and positions of items

relative to each other.

- *Actions* task assesses the model's proficiency in identifying and understanding actions occurring within images. It requires recognizing the activities depicted and understanding the roles and interactions of the participants involved.
- *Coreference* task determines the model's ability to resolve pronoun references within the visual context. It tests whether the MLLM can correctly link pronouns to the corresponding entities in the images, ensuring coherent understanding.

Additionally, VALSE presents foils for *Foil-It!* (Shekhar et al., 2017) dataset which connects objects in the captions to the MS-COCO (Lin et al., 2014) dataset. Refer to Appendix A for further details about VALSE benchmark.

In this work, we aim to investigate the performance of MLLMs on the VALSE benchmark and analyze how few-shot settings can enhance their capabilities in grounding language within visual contexts. Specifically, we focus on models pretrained on interleaved image-text data, which support few-shot learning, to understand the impact of this training strategy. Additionally, we analyze the performance of MLLMs pretrained solely on image captioning data, which do not support few-shot learning, to provide a comprehensive evaluation across different pretraining schemes.

3.2 Few-Shot ICL Strategy

Few-shot ICL aims to increase model performance by providing a few demonstration examples that are contextually related to the query image-text pair. The optimal selection and arrangement of these examples is an active area of research (An et al., 2023; Liu et al., 2022; Lu et al., 2022; Yoo et al., 2022; Min et al., 2022; Chen et al., 2023b). Our investigation examines the impact of in-context demonstrations on model performance by comparing randomly selected examples with those closely matching the visual and textual content of the query pair.

Example Selection. For example selection, we employed the Mixed Modality In-Context Example Selection (MMICES) method (Chen et al., 2023b). This method assesses both textual and visual cosine similarity between the image-text pairs in the demonstration examples and the query pair. Using CLIP as our encoder, we first identified the top K visually similar examples. From these K visually similar examples, we refined the selection to N examples exhibiting textual similarity. The value of N denotes the shot count used in our experiments.

Determining the appropriate value of K proved to be critical and challenging, as it directly influences the model's exposure to textually similar examples. Our analysis revealed that higher K values yielded improved results. Consequently, we set K to a high value of 100 for our experiments, ensuring that the model received suitable contextual information for learning and enhancement.

3.3 CoT Strategy

CoT approach aims to enhance model performance by promoting reasoning during inference, particularly in scenarios with limited data. Initially, we experimented with zero-shot CoT, where the model is asked to generate reasoning without providing additional context. However, we found that without this context, models often generate final answers without engaging in any reasoning process. To address this, we included reasoning information with the demonstration examples.

Given that samples in VALSE lack detailed, fine-grained descriptions for image-text pairs, we employed LLaVA-NeXT (Liu et al., 2024b) to generate CoT descriptions for the context demonstrations. Although this model is capable of generating dense captions, it occasionally fabricates incorrect information and hallucinates details. To mitigate these issues, we adopted a prompt proposed by Nori et al. (2023), instructing the model to generate both reasoning and answers, along with a label-validation step to reduce hallucinations. Despite these measures, some instances still lacked detailed CoT descriptions even when the answers were correct. Hence, we manually discarded instances with incorrect answers or inadequate CoT descriptions. We used only the remaining examples in our few-shot ICL with CoT experiments, as they provide detailed and contextually rich demonstrations. Details of this process are provided in the Appendix.

4 Experiments

4.1 Models

We evaluated fourteen state-of-the-art MLLMs, each varying in model size and trained on distinct pretraining datasets. Five of these models were trained on interleaved image-text data, facilitating to run in few-shot scenarios: OpenFlamingo (Awadalla et al., 2023), Idefics (Laurençon et al., 2023), Idefics2 (Laurençon et al., 2024), xGen-MM (Research, 2024), Qwen-VL-Chat (Bai et al., 2023), and MMICL (Zhao et al., 2024). The remaining four were trained solely on captioning datasets: LLaVA-NeXT (Liu et al., 2024a), PaliGemma (Gemma Team, 2024b), Intern-VL-Chat-V1.5 (Chen et al., 2023d), and InterLM-XComposer2 (Dong et al., 2024). Appendix B describes these models in detail.

4.2 Evaluation Strategy

Shukor et al. (2024) evaluates the effectiveness of the ITM (Image-Text Matching) method, initially examined within CREPE (Ma et al., 2023), which shares several similarities with VALSE. In this method, a sentence is presented to the model, labeled either as a caption or a foil, and the model is asked to determine if the sentence correctly describes the corresponding image. This allows for the measurement of accuracy, providing a quantitative assessment of the model's ability to link visual and linguistic information accurately. In our work, we assess the performance of MLLMs using this strategy and report the average accuracies accross both individual tasks and overall performance.

4.3 Results and Analysis

We show the zero-shot and few-shot capabilities of MLLMs trained on interleaved image-text datasets or captioning datasets in Table 1.

Observation 1. Instruction tuning and ICL help models follow user instructions.

Given our questions, we expect the MLLMs to give a Yes/No response. However, in zero-shot setting, some models struggled in producing outputs containing irrelevant information, leading to notably low scores. Instruction tuning or providing demonstration examples to the models through ICL often help models in following the expected answer templates. For instance, OpenFlamingo-3B and xGen-MM demonstrate this behavior.

Observation 2. Using similar demonstration examples in ICL significantly enhances performance compared to random examples.

Employing demonstration examples in the ICL setting generally improves overall performance. We observe this behavior consistently across the evaluated MLLMs independent from the model size. Notably, examples similar to query image-text pairs significantly enhance performance compared to random examples. For instance, in the 4-shot setting, OpenFlamingo 3B's performance on *Existence* improves from 54.5% (Random) to 67.9% (Similar).

Observation 3. Using more similar demonstration examples generally improves overall performance compared to using random demonstrations.

Shukor et al. (2024) studied atomic foils with the CREPE benchmark (Ma et al., 2023), which is similar to the VALSE benchmark in measuring model performance changes when atomic foils completely alter sentence meanings. They showed that increasing the number of random demonstration examples provides almost no gain in this setup. Our results support this finding and show that increasing the random example count can sometimes even deteriorate performance. However, using a higher number of similar examples helps MLLMs perform better. While more random examples make it difficult to establish a link between the context and query, more similar examples enhance this ability.

Observation 4. The CoT mechanism diminishes the ability to follow instructions acquired through ICL in OpenFlamingo variants, yet enhances the performance of other models in tasks where they struggle under both zero-shot and ICL settings.

CoT descriptions in demonstration examples assist models in reasoning about a given image-text pair, significantly aiding in challenging tasks such as counting, relations, and coreference. For example, in the 4-shot setting for OpenFlamingo 3B, performance on *Relations* improves from 50.1% (S) to 54.6% (S+C). However, CoT sometimes causes OpenFlamingo variants to ignore the expected answer templates. Although they generate reasoning chains as expected, they fail to provide direct answers to the questions, leading to poor performance. However, for the remaining higher capacity models, CoT generally leads to better performances.

Observation 5. With ICL and CoT, lower-capacity models trained on interleaved image-text datasets achieve similar or even better performance than larger-capacity models trained on captioning datasets.

Except for Idefics2, models trained on interleaved image-text datasets exhibit poor zero-shot performance compared to those trained on captioning data. However, with ICL and CoT, these lower-capacity models achieve similar or even better performance than the larger-capacity models trained on captioning datasets. For example, Idefics-9B obtained 77.2% accuracy when 4-shot ICL and

Table 1: Accuracy performance of the evaluated MLLMs, varying in model size and pretraining strategies, evaluated with 0-8 shots across three settings: Random (**R**), Similar (**S**), and Similar with Chain of Thought (**S+C**) settings. In the **R** setting, few-shot demonstrations are randomly selected. In the **S** setting, few-shot examples are selected based on visual and textual similarity. In the **S+C** setting, examples are also selected based on visual and textual similarity but additionally include a CoT description. Models with the suffix 'I' indicate instruction-tuned versions.

Zero-Shot Setting

Model	Existence	Plurality	Counting	Relations	Action	Coreference	Foil-It!	Average
LLaVA-NeXT-34B	**97.0**	71.3	**82.1**	57.4	70.9	**70.4**	**87.6**	76.7
PaliGemma-3B	76.6	63.7	74.1	47.1	64.2	51.2	81.2	65.4
Intern-VL-Chat-V1-5-26B	96.2	**76.5**	76.9	**61.3**	**74.2**	69.5	87.1	**77.4**
InternLM-XComposer2-7B	83.0	66.5	73.7	52.5	68.8	62.2	82.0	69.8
OpenFlamingo-3B	36.4	9.4	14.2	9.0	8.5	32.0	11.0	17.2
OpenFlamingo-3B I	48.3	48.3	45.6	44.1	46.0	25.0	43.3	42.9
OpenFlamingo-4B	46.9	54.6	49.0	47.5	51.6	49.3	49.3	49.7
OpenFlamingo-4B I	48.5	54.8	50.1	47.5	51.9	46.9	49.3	49.9
Idefics-9B	44.2	46.2	47.1	53.8	48.2	26.3	50.4	45.2
Idefics-9B I	58.2	54.6	50.5	49.5	58.1	54.8	56.6	54.6
Idefics2-8B	94.7	70.3	79.1	53.6	59.8	69.1	82.1	72.7
xGen-MM-4.6B	37.2	34.1	37.1	39.6	36.4	37.0	40.9	37.5
Qwen-VL-Chat-9.6B	82.6	46.3	68.3	48.0	41.1	58.7	61.9	58.1
MMICL-12.1B	65.4	57.9	53.1	57.2	59.4	61.9	59.3	59.2

4-Shot Setting

Model	Existence			Plurality			Counting			Relations			Action			Coreference			Foil-It!			Average		
	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C
OpenFlamingo-3B	54.5	67.9	45.7	53.2	52.2	32.7	54.3	59.3	41.5	47.7	52.9	29.9	49.0	51.9	33.0	52.7	57.2	25.4	50.8	52.8	28.4	51.7	56.3	33.8
OpenFlamingo-3B I	52.1	61.6	49.3	53.4	50.5	34.1	53.4	57.4	41.1	51.0	50.1	24.5	54.2	52.7	31.1	51.5	55.0	24.0	50.7	50.2	32.0	52.3	53.9	33.7
OpenFlamingo-4B	53.7	73.1	43.6	50.9	52.3	42.5	54.6	58.4	39.9	50.1	54.6	28.8	57.8	57.5	30.6	50.5	52.9	31.3	48.4	53.8	33.2	52.3	57.5	35.7
OpenFlamingo-4B I	51.9	66.1	44.6	51.9	49.2	37.6	54.1	59.2	41.2	50.5	54.6	27.3	56.2	58.3	33.7	50.8	53.0	33.0	50.0	53.1	30.1	52.2	56.2	35.6
Idefics-9B	59.2	81.0	87.3	49.8	54.8	73.6	54.7	61.2	79.4	50.6	52.1	**72.9**	56.4	60.5	74.5	51.7	53.6	**82.8**	57.0	59.8	69.6	54.2	60.4	**77.2**
Idefics-9B I	74.3	88.3	**87.5**	58.8	58.0	69.0	59.2	65.0	78.3	54.8	57.2	70.5	67.5	72.9	75.7	57.3	59.2	76.5	72.2	77.9	82.7	63.4	68.3	**77.2**
Idefics2-8B	83.2	**94.3**	79.8	**70.3**	**69.7**	**76.6**	**73.4**	**71.4**	**80.1**	**61.7**	**63.2**	70.1	70.3	72.6	**77.0**	63.3	59.8	70.7	**82.6**	**84.9**	**83.1**	**72.1**	**73.7**	76.8
xGen-MM-4.6B-7B	65.2	77.0	73.9	56.8	58.8	71.0	55.6	57.3	72.0	51.6	56.3	69.7	61.2	67.0	67.4	54.6	57.9	67.3	63.3	70.7	78.3	58.3	63.6	71.4
Qwen-VL-Chat-9.6B	**85.2**	92.7	85.7	66.4	64.4	67.5	68.9	69.8	76.7	60.8	60.2	57.0	71.4	72.5	67.0	**64.8**	**62.0**	72.2	79.2	80.1	65.6	71.0	71.7	70.2
MMICL-12.1B	56.6	70.5	37.6	54.4	54.8	16.9	50.1	55.9	32.4	57.2	60.6	25.2	**75.2**	**73.0**	24.9	61.8	60.5	40.2	59.7	56.6	21.7	59.3	61.7	28.4

8-Shot Setting

Model	Existence			Plurality			Counting			Relations			Action			Coreference			Foil-It!			Average		
	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C	R	S	S+C
OpenFlamingo-3B	51.5	72.3	58.4	51.7	51.7	38.4	53.1	58.6	47.9	50.3	49.5	38.5	51.9	56.8	36.3	52.1	56.3	31.6	53.9	50.3	32.2	52.1	56.5	40.5
OpenFlamingo-3B I	51.7	65.3	51.3	50.3	53.1	35.4	53.3	57.4	41.6	53.6	46.9	32.2	49.7	59.7	31.8	52.5	57.2	26.1	52.5	50.8	32.3	51.9	55.8	35.8
OpenFlamingo-4B	52.5	74.1	72.1	52.1	55.6	58.9	56.0	63.6	57.8	52.9	55.9	52.5	59.4	59.4	41.4	49.9	54.2	39.9	52.2	56.5	55.1	53.6	59.9	54.0
OpenFlamingo-4B I	49.9	64.4	56.4	52.1	52.6	47.6	54.4	60.8	53.9	49.7	55.1	41.7	60.1	60.7	47.5	53.4	59.3	44.4	52.4	57.8	39.6	53.1	58.7	47.3
Idefics-9B	57.2	84.4	**92.1**	48.4	55.6	**77.9**	54.8	65.3	**86.9**	53.1	56.1	**83.6**	59.0	66.5	**78.2**	53.2	58.6	70.7	58.1	60.2	75.0	54.8	63.8	**80.6**
Idefics-9B I	76.2	89.9	79.2	57.2	61.0	70.2	58.5	65.2	76.1	56.6	60.8	69.2	68.2	71.4	76.4	55.6	61.5	53.4	74.3	76.3	77.4	63.8	69.4	71.7
Idefics2-8B	**88.5**	94.3	86.7	**70.5**	**71.6**	76.2	**74.5**	**72.1**	83.0	59.6	61.1	71.6	72.0	71.3	75.7	61.0	65.4	68.3	82.6	**83.9**	**81.3**	**72.7**	**74.2**	77.5
xGen-MM-4.6B-7B	65.5	86.1	69.1	56.3	61.5	61.5	55.5	61.6	65.2	54.2	57.6	67.5	65.8	71.0	62.3	56.5	54.1	61.0	64.7	70.4	73.0	59.8	66.0	65.7
Qwen-VL-Chat-9.6B	84.2	**95.3**	72.9	64.2	66.5	65.8	70.0	71.7	76.1	**60.6**	61.5	63.7	72.0	71.5	72.9	62.4	63.9	**76.1**	**84.6**	83.5	66.2	71.1	73.4	70.5
MMICL-12.1B	63.6	78.6	38.6	53.5	56.4	14.3	47.7	52.2	31.9	58.9	**63.4**	21.1	**75.7**	**71.6**	19.6	**63.5**	**65.6**	37.5	61.9	66.3	20.3	60.7	64.9	26.2

CoT are applied while Intern-VL-Chat-V1-5-26B achieved 76.7% overall accuracy.

Observation 6. Models prefer demonstrations that are predominantly textually similar to visual ones, resulting in a slight increase in performance.

Table 2 shows the performance changes of models pretrained on interleaved image-text datasets across different K values within the ICL setting. Increasing the value of K provides a larger pool of visually similar examples. Subsequently, when N examples are selected from this pool based on textual similarity, the final demonstration examples tend to exhibit higher textual similarity to the query image-text pair, albeit potentially lower visual similarity. The results indicate a marginal performance improvement with higher K, suggesting that models prefer more textually similar examples.

For additional analyses and qualitative examples of few-shot learning settings, see the Appendix.

Table 2: Accuracy performance of the MLLMs pretrained on interleaved image and text data, varying in model size, in the few-shot ICL setting. Demonstrations are selected based on their similarity to the query. For each setting, (N) textual similar examples are chosen from (K) visual similar examples. The table shows performance across different (K) values, specifically 20, 50, and 100. Models with the suffix 'I' indicate instruction-tuned versions.

Zero-Shot Setting

Model	Existence	Plurality	Counting	Relations	Action	Coreference	Foil-It!	Average
OpenFlamingo-3B	36.4	9.4	14.2	9.0	8.5	32.0	11.0	17.2
OpenFlamingo-3B I	48.3	48.3	45.6	44.1	46.0	25.0	43.3	42.9
OpenFlamingo-4B	46.9	54.6	49.0	47.5	51.6	49.3	49.3	49.7
OpenFlamingo-4B I	48.5	54.8	50.1	47.5	51.9	46.9	49.3	49.9
Idefics-9B	44.2	46.2	47.1	53.8	48.2	26.3	50.4	45.2
Idefics-9B I	58.2	54.6	50.5	49.5	58.1	54.8	56.6	54.6
Idefics2-8B-8B	**94.7**	**70.3**	**79.1**	53.6	**59.8**	**69.1**	**82.1**	**72.7**
xGen-MM-4.6B	37.2	34.1	37.1	39.6	36.4	37.0	40.9	37.5
Qwen-VL-Chat-9.6B	82.6	46.3	68.3	48.0	41.1	58.7	61.9	58.1
MMICL-12.1B	65.4	57.9	53.1	**57.2**	59.4	61.9	59.3	59.2

4-Shot Setting

Model	Existence			Plurality			Counting			Relations			Action			Coreference			Foil-It!			Average		
	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100
OpenFlamingo-3B	65.0	67.7	67.9	55.5	52.4	52.2	57.5	59.3	59.3	52.5	49.4	52.9	53.9	50.9	51.9	56.0	52.3	57.2	54.2	57.0	52.8	56.4	55.6	56.3
OpenFlamingo-3B I	53.1	58.8	61.6	53.1	49.2	50.5	60.0	58.2	57.4	53.3	50.3	50.1	53.1	54.1	52.7	55.3	53.7	55.0	50.0	52.5	50.2	54.0	53.8	53.9
OpenFlamingo-4B	63.8	69.3	73.1	53.1	49.2	52.3	57.6	58.8	58.4	52.3	53.8	54.6	54.9	54.1	57.5	51.1	51.8	52.9	52.8	55.6	53.8	55.1	56.1	57.5
OpenFlamingo-4B I	62.4	63.8	66.1	50.3	45.6	49.2	57.8	59.6	59.2	51.0	53.3	54.6	55.3	57.2	58.3	51.4	52.2	53.0	52.9	53.7	53.1	54.4	55.1	56.2
Idefics-9B	76.0	79.6	81.0	57.6	57.0	54.8	58.3	59.9	61.2	57.6	52.1	52.1	61.6	62.1	60.5	53.6	53.7	53.6	58.2	60.1	59.8	60.4	60.6	60.4
Idefics-9B I	86.3	86.7	88.3	58.0	56.0	58.0	61.4	63.3	65.0	59.1	57.9	57.2	71.5	71.9	72.9	58.5	55.0	59.2	76.7	79.1	77.9	67.4	67.1	68.3
Idefics2-8B	**92.7**	**94.3**	**94.3**	**71.2**	**68.2**	**69.7**	**71.7**	**71.9**	71.4	**63.4**	**63.0**	**63.2**	**72.4**	73.8	72.6	62.1	58.5	59.8	**84.7**	**84.2**	**84.9**	**74.0**	**73.4**	**73.7**
xGen-MM-4.6B	74.7	78.8	77.0	61.3	61.0	58.8	55.5	56.1	57.3	59.8	60.6	56.3	68.3	66.9	67.0	56.6	54.2	57.9	69.0	71.6	70.7	63.6	64.2	63.6
Qwen-VL-Chat-9.6B	85.2	92.7	85.7	66.4	64.4	67.5	68.9	69.8	**76.7**	60.8	60.2	57.0	71.4	72.5	67.0	**64.8**	**62.0**	**72.2**	79.2	80.1	65.6	71.0	71.7	70.2
MMICL-12.1B	65.5	70.9	70.5	52.2	50.1	54.8	52.6	53.0	55.9	59.8	60.8	60.6	72.1	**74.8**	**73.0**	61.0	60.4	60.5	59.9	61.2	56.6	60.4	61.6	61.7

8-Shot Setting

Model	Existence			Plurality			Counting			Relations			Action			Coreference			Foil-It!			Average		
	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100	20	50	100
OpenFlamingo-3B	65.5	66.9	72.3	51.7	52.5	51.7	56.0	60.0	58.6	47.1	52.9	49.5	56.9	56.8	56.8	53.9	58.4	56.3	52.0	51.5	50.3	54.7	57.0	56.5
OpenFlamingo-3B I	56.4	62.2	65.3	49.0	53.4	53.1	56.6	58.3	57.4	48.8	52.1	46.9	57.7	56.8	59.7	53.9	58.6	57.2	51.5	54.5	50.8	53.4	56.6	55.8
OpenFlamingo-4B	59.8	69.5	74.1	52.5	51.7	55.6	60.7	61.5	63.6	52.3	53.1	55.9	63.0	60.8	59.4	52.8	55.6	54.2	55.6	57.4	56.5	56.7	58.5	59.9
OpenFlamingo-4B I	54.6	59.8	64.4	50.9	50.2	52.6	57.5	57.8	60.8	51.8	50.3	55.1	62.5	60.5	60.7	54.4	57.0	59.3	52.7	53.0	57.8	54.9	55.5	58.7
Idefics-9B	73.1	79.6	84.4	53.4	57.0	55.7	60.7	66.6	65.3	54.0	56.3	56.1	65.9	64.7	66.5	54.2	57.2	58.6	58.9	61.8	60.2	60.0	63.3	63.8
Idefics-9B I	81.6	84.8	89.9	61.1	61.2	61.0	62.2	65.9	65.2	59.4	57.4	60.8	72.2	72.0	71.4	56.4	60.5	61.5	76.7	76.0	76.3	67.1	68.3	69.4
Idefics2-8B	**92.5**	**93.7**	94.3	**70.9**	**68.7**	**71.6**	**72.2**	**72.5**	**72.1**	63.0	**62.1**	61.1	72.7	71.6	71.3	63.0	62.7	65.4	**82.9**	**84.2**	**83.9**	**73.9**	**73.6**	**74.2**
xGen-MM-4.6B	79.6	85.0	86.1	57.9	60.3	61.5	59.6	62.8	61.6	59.4	57.9	57.6	72.8	70.9	71.0	54.4	56.5	54.1	69.9	70.0	70.4	64.8	66.2	66.0
Qwen-VL-Chat-9.6B	90.7	92.3	**95.3**	63.9	63.6	66.5	71.8	72.3	71.7	**63.4**	59.8	61.5	72.2	73.1	71.5	**66.4**	**67.2**	63.9	80.8	83.1	83.5	72.7	73.1	73.4
MMICL-12.1B	74.3	77.8	78.6	55.9	55.1	56.4	49.8	51.8	52.2	63.0	61.5	**63.4**	**74.0**	**73.2**	**71.6**	62.4	64.6	**65.6**	61.3	61.6	66.3	63.0	63.7	64.9

5 Conclusion

This work evaluates MLLMs using the VALSE benchmark to assess the impact of ICL and CoT. Our findings show that these strategies significantly enhance model performance, especially in tasks requiring complex reasoning and context understanding. We identified specific areas where MLLMs excel and where they struggle, emphasizing the importance of training data composition, pretraining strategies, and effective prompting techniques.

One key insight is that MLLMs trained on captioning datasets perform better in zero-shot settings, while those trained on interleaved image-text data benefit more from few-shot learning. This suggests that targeted pretraining and few-shot strategies are crucial for improving model performance in complex tasks. ICL and CoT prompting enable MLLMs to leverage contextual information and reason through intermediate steps. Future research should optimize these strategies and explore additional methods to enhance model robustness and reasoning capabilities. By refining sophisticated reasoning mechanisms, we can develop MLLMs that are more flexible and effective across a wider range of tasks and settings.

6 Limitations

While the VALSE benchmark provides a comprehensive framework, it may not cover all possible linguistic phenomena or real-world scenarios, potentially limiting the generalizability of the findings to other datasets or applications. Moreover, our study evaluates only fourteen state-of-the-art MLLMs, which, although representative, may not encompass the full spectrum of available models and their respective training datasets. For instance, closed-source proprietary models such as GPT-4o (OpenAI, 2024), Gemini 1.5 Pro (Gemini Team, 2024), and Claude 3 Opus (Anthropic, 2024) are intentionally left out due to their restricted access, which limits the ability to conduct comprehensive and reproducible evaluations.

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*.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. 2022. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.

AI@Meta. 2024. Llama 3 model card.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052, Toronto, Canada. Association for Computational Linguistics.

Anthropic. 2024. Introducing the next generation of claude. Available at: `https://www.anthropic.com/news/claude-3-family`.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? *arXiv preprint arXiv:2404.15736*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for VQA are image captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1947–1963, Seattle, United States. Association for Computational Linguistics.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. 2023b. Understanding and improving in-context learning on vision-language models. *Preprint*, arXiv:2311.18021.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023c. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. 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, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. 2024. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Alex Falcon, Oswald Lanz, and Giuseppe Serra. 2020. Data augmentation techniques for the video question answering task. In *European Conference on Computer Vision*, pages 511–525. Springer.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Google Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arxiv preprint arXiv:2403.05530*.

Google Gemma Team. 2024a. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Google Gemma Team. 2024b. Paligemma model. Available at: https://github.com/google-research/big_vision/tree/main/big_vision/configs/proj/paligemma.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Jiabang He, Lei Wang, Yingpeng Hu, Ning Liu, Hui juan Liu, Xingdong Xu, and Hengtao Shen. 2023. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. arxiv abs/2303.05063 (2023).

Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. 2018. Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pages 57–64.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2023. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

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, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023a. Languages are rewards: Hindsight finetuning using human feedback. *arXiv preprint arXiv:2302.02676*.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. Dr.ICL: Demonstration-retrieved in-context learning. In *NeurIPS 2023 Workshop on the Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (R0-FoMo)*.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Xiangyang Mou, Brandyn Sigouin, Ian Steenstra, and Hui Su. 2020. Multimodal dialogue state tracking by qa approach with data augmentation. *arXiv preprint arXiv:2007.09903*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2024. Hello gpt-4o. Available at: `https://openai.com/index/hello-gpt-4o/`.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. 2022. Pre-training image-language transformers for open-vocabulary tasks. *arXiv preprint arXiv:2209.04372*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Salesforce AI Research. 2024. xgen-mm-phi3-mini-instruct model card.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2024. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

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

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

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

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

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

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. MMICL: Empowering vision-language model with multi-modal in-context learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023b. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

Appendix

In the following sections, we provide a comprehensive set of supplementary notes detailing various aspects of our work:

- **Detailed Review of VALSE Benchmark (§A):** This section elaborates on the VALSE benchmark, outlining the specific tasks it encompasses.
- **Detailed Review of Evaluated Multimodal LLMs (§B):** We offer an in-depth review of all evaluated MLLMs, emphasizing their unique characteristics and capabilities.
- **Demonstrations (§C):** This section describes our methodology for selecting demonstrations and constructing Chain-of-Thought (CoT) descriptions.
- **Further Analysis (§D):** We expand on our key findings, providing additional analyses and insights into individual tasks within the VALSE benchmark.
- **Qualitative Examples (§E):** We present qualitative examples that illustrate the few-shot learning settings considered in our study.

A VALSE Benchmark

The VALSE benchmark (Parcalabescu et al., 2022) is a pioneering effort to evaluate the abilities of general-purpose pretrained vision and language models in grounding linguistic constructs within a visual context. It consists of six tasks—*Existence*, *Plurality*, *Counting*, *Spatial Relations*, *Actions*, and *Coreference*—each targeting a key linguistic phenomena (see Figure 2). These tasks assess models' capabilities in recognizing existential quantifiers, semantic number, entity counting, spatial arrangements, actions, and pronominal coreference within images, providing a thorough evaluation framework for exploring the complexities of language grounding in visual contexts. The benchmark contains 6795 examples in total.

To develop VALSE, rigorous methodologies were applied to ensure the benchmark's validity and effectiveness (Lan et al., 2019). This included establishing robust criteria for generating valid foils (Xie et al., 2019), which are crucial for accurately assessing model performance. Through detailed experimentation and evaluation of five widely-used MLLMs, the original VALSE paper provided insights into the current challenges faced by pretrained models in understanding and interpreting linguistic phenomena in visual contexts.

B Evaluated MLLMs

Here, we describe the models used in our experiments. We tested models trained on datasets containing image-text pairs (§B.1) as well as models trained on interleaved image-text datasets (§B.2). Figure 3 demonstrates sample data that are utilized in each training strategy.

B.1 MLLMs pretrained on Captioning Datasets

Recently, there has been considerable interest in NLP regarding models capable of handling single image-text pairs (Li et al., 2023c; Dai et al., 2024; Liu et al., 2024b; Zhu et al., 2023a; Bavishi et al., 2023; Ge et al., 2023). These models demonstrate a remarkable ability to understand and generate textual descriptions for given images, which greatly aids tasks such as image captioning, visual question answering, and image retrieval. By employing sophisticated architectures and multimodal learning techniques, these models effectively integrate visual and textual data to deduce semantic meaning and context. Consequently, they hold significant potential for diverse applications in image comprehension, multimedia analysis, and human-computer interaction.

LLaVA (Liu et al., 2024b), also known as Large Language and Vision Assistant, model family, including LLaVA 1.5 (Liu et al., 2023b) and LLaVA-NeXT (Liu et al., 2024a), represents a significant leap forward in large multimodal models research. These models surpass natural instruction-following and visual reasoning tasks, with LLaVA 1.5 setting new standards across 12 datasets. The latest iteration, LLaVA-NeXT, enhances reasoning, OCR, and world knowledge capabilities, even outperforming Gemini Pro 1.0 (Gemini Team, 2023) on certain benchmarks. LLaVA-NeXT achieves these improvements while maintaining a minimalist design and high data efficiency, requiring fewer than 1M visual instruction tuning samples for training. Notably, it demonstrates leading performance among open-source large multimodal models, with significantly lower training costs. During our evaluation, we decided to use the LLaVA-NeXT 34B variant.

PaliGemma, created by Google, is another powerful MLLM featuring a Transformer decoder and a Vision Transformer image encoder, having 3 billion parameters. Built from Gemma-2B (Gemma Team, 2024a) and SigLIP-So400m/14 (Zhai et al., 2023),

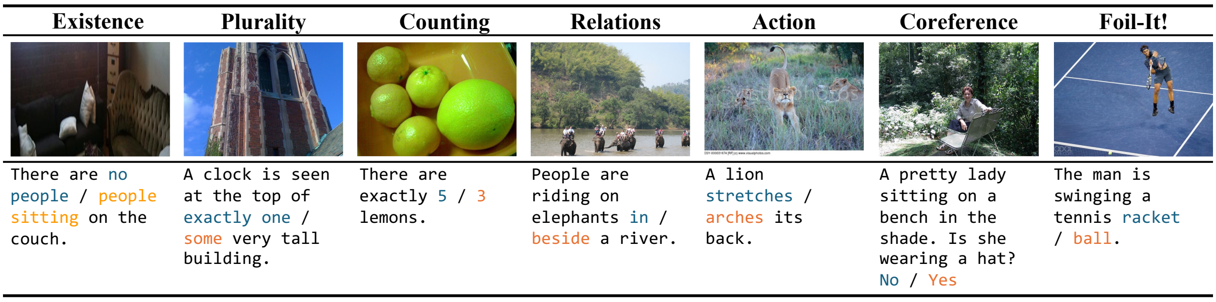


Figure 2: Sample instances from the VALSE benchmark (Parcalabescu et al., 2022).

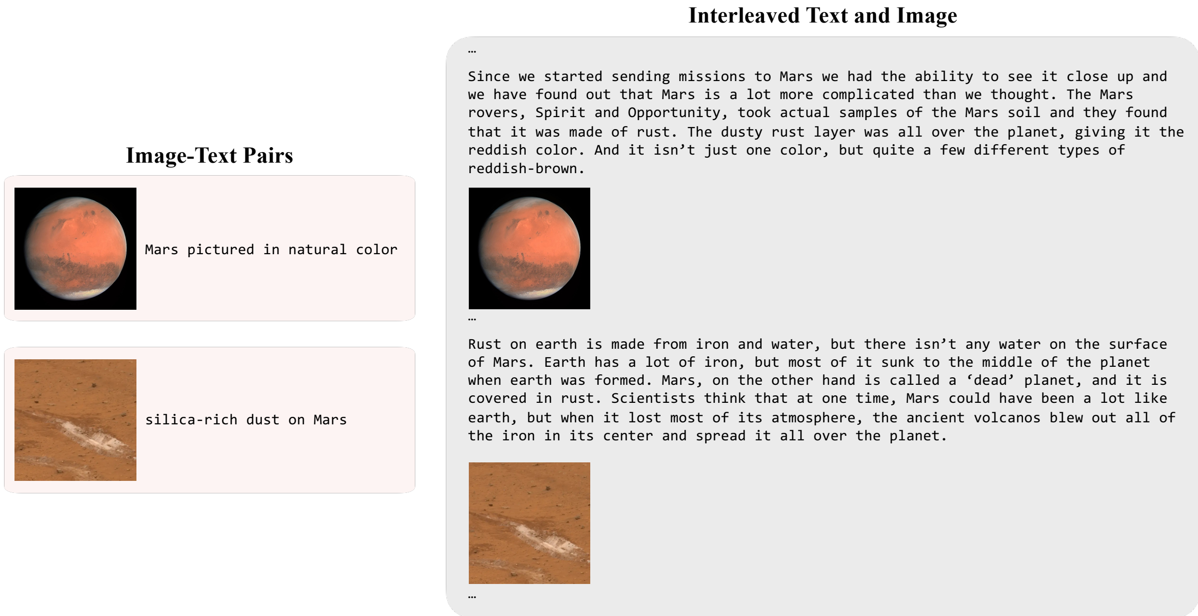


Figure 3: Sample data demonstrating the differences between image-text pairs, and interleaved text and image data used in training MLLMs.

it follows the PaLI-3 training protocol (Chen et al., 2023c). This model accepts images and text strings as inputs, generating outputs like image captions, answers to questions, object bounding box coordinates, or segmentation codewords. Pre-trained on a variety of datasets including WebLI (Chen et al., 2023c), CC3M-35L (Chen et al., 2022b), VQ²A-CC3M-35L/VQG-CC3M-35L (a subset of VQ2A-CC3M (Changpinyo et al., 2022)), OpenImages (Piergiovanni et al., 2022), and WIT (Srinivasan et al., 2021), PaliGemma surpasses in visual semantic understanding and multilingual tasks. Rigorous data responsibility filters are applied to ensure the training data is safe, clean, and respects privacy by removing inappropriate or sensitive content using advanced filtering techniques.

Intern-VL-Chat-V1-5 (Chen et al., 2024) is an advanced vision-language model with 26B parameters aimed at closing the performance gap between open-source and commercial models. It utilizes the InternViT-6B (Chen et al., 2023d) vision foundation model and InternLM2-20B (Cai et al., 2024) language model, enhanced by three key features: continuous learning with high-quality image-text data, a dynamic high-resolution strategy for detailed image analysis, and a diverse multilingual dataset pipeline. In tests across 18 multimodal benchmarks, InternVL 1.5 achieved top results in 8 benchmarks, surpassing leading models like GPT-4V (Achiam et al., 2023) in OCR-related tasks, showcasing its ability to narrow the gap between open-source and commercial multimodal models.

InternLM-XComposer2 (Dong et al., 2024), with 7B parameters, surpasses in generating and comprehending free-form text-image content. By combining text and graphics from diverse inputs such as outlines and reference images, it allows for highly flexible content production beyond traditional comprehension. Utilizing a Partial LoRA (PLoRA) approach to strategically apply additional parameters to image tokens, InternLM-XComposer2 preserves language understanding while enhancing vision comprehension, leading to superior performance in various evaluations compared to existing multimodal models like GPT-4V (Achiam et al., 2023) and Gemini Pro (Gemini Team, 2023).

B.2 MLLMs pretrained on Interleaved Image-Text Data

The development of models capable of handling multiple image-text pairs has become a critical focus in research (Awadalla et al., 2023; Laurençon et al., 2023; Laurençon et al., 2024; Jiang et al., 2024; Ye et al., 2023; Li et al., 2023b; Bai et al., 2023; Alayrac et al., 2022). These frameworks demonstrate the ability to analyze and comprehend several instances of image-text pairs simultaneously, enabling a more thorough understanding and interpretation of multimodal data. Through the utilization of advanced multimodal fusion techniques and attention mechanisms, these models seamlessly integrate information from various sources to extract nuanced semantics and context across multiple modalities. This expanded capability broadens the range of applications to tasks such as image album summarization, cross-modal retrieval, and interactive storytelling, where the analysis of multiple image-text pairs enriches the depth and complexity of information processing and comprehension.

OpenFlamingo (Awadalla et al., 2023) introduces a fresh approach to vision and language modeling, enabling autoregressive models to process sequences of mixed images and text for enhanced flexibility, including few-shot learning and multi-round chatbot interactions. Unlike proprietary models such as Flamingo (Alayrac et al., 2022), CM3 (Aghajanyan et al., 2022), Kosmos-1 (Huang et al., 2024), PALME (Driess et al., 2023), and multimodal GPT-4 (Achiam et al., 2023), OpenFlamingo provides an open-source alternative, promoting research accessibility. By leveraging pretrained language models with cross-modal attention to vision encoders, OpenFlamingo achieves competitive performance, with models ranging from 3B to 9B parameters. Evaluation across seven datasets indicates that OpenFlamingo models reach 85% to 89% of the performance of their corresponding Flamingo models, underscoring their effectiveness and adaptability.

Idefics (Laurençon et al., 2023; Laurençon et al., 2024) , includes two versions: Idefics1 and Idefics2. Idefics1, an open-access multimodal model inspired by DeepMind's Flamingo, processes sequences of images and text to generate textual outputs. Utilizing publicly available data and models like CLIP-ViT-H-14 (Schuhmann et al., 2022) and LLaMA-65B (Touvron et al., 2023), it comes in two sizes (80B and 9B parameters) and surpasses image captioning and visual question-answering benchmarks. Idefics2, with 8B parameters, offers improved OCR capabilities, document understanding, and visual reasoning. It handles images in their native resolutions with the NaViT strategy (Dehghani et al., 2024) and incorporates new training data for enhanced OCR and document comprehension.

xGen-MM (Research, 2024) series, developed by Salesforce AI Research, builds on the successful BLIP series, aligned with Salesforce's XGen initiative for large foundational models. These models, trained on diverse datasets including high-quality image captions, demonstrate state-of-the-art performance in contextual learning. Notably, the xGen-MM mini base model achieves superior performance with under 5 billion parameters, while the fine-tuned xGen-MM mini instruction-tuned model surpasses high-resolution image encoding. Training data sources range from CC12M (Changpinyo et al., 2021) to academic VQA tasks, ensuring versatility and robustness. We used the xGen-MM mini base with a model size of 4.6B variant during our experiments.

Qwen-VL (Bai et al., 2023) series expands on the Qwen language model, overcoming the limitations of traditional LLMs by integrating visual understanding capabilities. These models, including Qwen-VL-Chat, 9.6B parameters, enable interaction with users through both text and images. They surpass tasks like image captioning and question answering, boasting superior performance and supporting multiple languages. Additionally, Qwen-VL models handle multiple images and demonstrate strong performance across various benchmarks, particularly in fine-grained visual understanding.

MMICL (Zhao et al., 2024), Multi-Modal In-Context Learning, is designed to address the shortcomings of existing MLLMs in processing complex prompts that involve multiple images and text. MMICL, with a model size of 12.1B, introduces a new method for handling multi-modal inputs, proposes a unique context scheme to improve in-context learning, and utilizes the Multi-modal In-Context-Learning (MIC) dataset to enhance the model's ability to understand complex multi-modal prompts. This model effectively tackles challenges such as understanding text-to-image references and the relationships between multiple images. Additionally, MMICL reduces language bias, which often causes MLLMs to produce hallucinations when dealing with extensive textual contexts.

For our experiments, we follow the model implementations in the HuggingFace repository. We used half-precision to run Idefics1, MMICL, and full precision to run OpenFlamingo variants and xGen-MM. For InterVL-Chat, we applied 8-bit quantization, while the rest of the models were tested with 4-bit quantization. We conducted our experiments on a single Tesla T4, Quadro P4000, V100 or A40 GPU.

C Demonstration Examples

Similar Example Selection. Given the relatively modest size of the VALSE dataset, we opted against partitioning it for creating a demonstration example set. Instead, we leveraged the remaining dataset, excluding the query image-text pair under examination.

Chain-of-Thought Generation. CoT approach aims to enhance model performance by promoting reasoning during inference, especially in scenarios with limited data. Initially, we experimented with zero-shot CoT, where the model generates reasoning without additional context. However, in this setup, models often produced final answers without engaging in reasoning. To address this, we incorporated reasoning information into the demonstration examples. In particular, we employed MLLMs to generate these CoT descriptions. The prompt that is used to generate CoT descriptions is given below:

```
"Given an image and a corresponding sentence, analyze the image to determine if the sentence is true or false. Provide the answer in the format: Final Answer: Yes (if the sentence is true for the image) / No (if the sentence is false for the image). Sentence: ..."
```

During this process, we encountered challenges such as fabricated information and hallucinated details. To mitigate these issues, we filtered out descriptions yielding incorrect answers. Despite these measures, some instances still lacked CoT descriptions even when the answers were correct, eventually leading us to discard those with inaccurate or inadequate descriptions and the corresponding samples while selecting the demonstrations for few-shot (ICL + CoT) experiments.

To generate CoT reasonings and avoid hallucinations, we applied an automatic filtering approach to eliminate some responses. We tested three MLLMs: LLaVA-NeXT 34B (Liu et al., 2024a), InternLM-XComposer2, and LLaVA-LLaMA3 (Contributors, 2023), a LLaVA-1.5-7B (Liu et al., 2024b) model finetuned from LLaMA-8B Instruct (AI@Meta, 2024). Table 3 shows the rate of successful description generation for each model. The results indicate that LLaVA-NeXT clearly surpasses the other models, and larger models generate better reasoning chain descriptions.

D Further Analysis

In this section, we provide a detailed analysis of the results for each task in VALSE.

D.1 Existence

The Existence task is the most basic yet fundamental task in VALSE, assessing a model's ability to determine the presence or absence of an object in an image. All models demonstrated higher accuracy on this task compared to others, indicating that MLLMs effectively represent objects and determine their existence in a scene. However, when CoT descriptions were introduced, the performance of all models, except for Idefics-9B, deteriorated. This decline is attributed to the models hallucinating and generating irrelevant reasoning chains in response to the actual question, ultimately leading to incorrect answers. Additionally, as shown in Table 2, an increase in textually similar examples significantly boosts model performance more than in other tasks.

D.2 Plurality

The Plurality task is challenging because the models must not only recognize the given object but also determine its plural form. Results reveal that demonstration examples do not improve the models' understanding of pluralism, although the models correctly recognize the objects. For this task, CoT reasoning is useful as it directly provides reasoning chains that describe what a plural form is. With this context, models are able to develop an understanding of the task.

Table 3: Rate of valid Chain-of-Thought (CoT) descriptions generated by the corresponding models.

Model	Existence	Plurals	Counting	Relations	Action	Coreference	Foil-It!
LLaVA-NeXT-34B	88.3	55.2	62.4	42.2	45.8	70.9	69.8
LLaVA-LLAMA3-8B	5.9	20.6	6.0	17.2	15.6	16.5	7.6
InternLM-XComposer2-7B	1.8	10.3	10.8	9.7	8.3	13.8	2.3

D.3 Counting

The Counting task, similar to Plurality, evaluates a model's understanding of the exact count of an item in a scene. The model must identify both the object and the number of its appearances. Models trained on captioning datasets outperform those trained on interleaved image-text data. However, the combination of few-shot ICL and CoT reasoning enhances the performance of these models, bringing them closer to those trained on captioning data. As seen in qualitative examples, models are guided to count each occurrence, allowing for a direct comparison between the actual and stated occurrences.

D.4 Spatial Relations

The Spatial Relations task evaluates models' abilities to recognize interactions between objects. Zero-shot performance shows that all models struggle with this task, as it requires a deep understanding of the interactions and relationships between objects. Results indicate that providing demonstration examples through ICL helps models achieve a certain performance level, but increasing the number of demonstrations does not lead to further improvement. Performance gains saturate with a higher example count. However, using few-shot ICL combined with CoT reasoning, it is possible to achieve up to a 30% performance increase (Idefics-9B).

D.5 Action

The Action task aims to assess how successfully models detect actions and actors in a scene. This task is relatively hard as it requires models to accurately identify dynamic interactions and context-specific activities within an image, which demands a deeper understanding beyond static object recognition. In this task, models trained on captioning data performed better compared to those trained on interleaved image-text datasets. Few-shot ICL successfully elevated these models' performance to up to 73%. However, except for the Idefics model family, none of the models benefited from CoT descriptions. Additionally, increasing the number of demonstration examples did not always positively impact performance.

D.6 Coreference

The Coreference task evaluates a model's capability to resolve pronoun references within a visual context, examining whether MLLMs can accurately associate pronouns with their corresponding entities in images to maintain coherent understanding. This task is challenging as the models need to accurately interpret and maintain contextual relationships between pronouns and their antecedents within a visual scene. Results show that models trained on captioning datasets outperformed those trained on interleaved image-text datasets. The Idefics model family, in particular, substantially benefited from CoT descriptions, which are crucial for solving the coreference task as they provide explicit reasoning pathways to link pronouns correctly. However, the general model performance with CoT descriptions does not improve much as the number of demonstration examples increases.

D.7 Foil-It!

The Foil-It! task is designed to evaluate a model's understanding of objects by replacing the target object with an irrelevant one to create a foil. This task demands models to not only recognize objects accurately but also to detect subtle inconsistencies in the context. Similar to the Counting task, the zero-shot performance of models trained on captioning data surpassed that of models trained on interleaved image-text datasets. Additionally, these models could not be outperformed even with the application of few-shot ICL and CoT techniques.

E Qualitative Examples

In this section, we show example model responses from our evaluation. Figure 4-10 illustrate the process of evaluating In-Context Learning (ICL). In this setup, demonstration examples are selected based on their similarity to the query and are provided with their ground truth answers before presenting the actual query image-text pair to the model. Similarly, Figure 11-17 demonstrate the evaluation of Chain-of-Thought (CoT) reasoning combined with ICL. In this setting, CoT descriptions are included with the demonstration examples. These detailed reasoning chains guide the model in making inferences for the query image-text pair.

`Does the following sentence describe the image? There are people surfing.`
`Answer: Yes`

`Does the following sentence describe the image? There is no woman.`
`Answer: No`

`Does the following sentence describe the image? There is a surfer.`

OpenFlamingo-3B: `Answer: Yes Does the following sentence`
OpenFlamingo-3B I: `Answer: Yes`
OpenFlamingo-4B: `Answer: Yes`
OpenFlamingo-4B I: `Answer: Yes`
Idefics-9B: `Answer: Yes`
Idefics-9B I: `Answer: Yes`
Idefics2-8B: `Answer: Yes`
xGen-MM-4.6B-7B: `Answer: Yes`
Qwen-VL-Chat-9.6B: Yes
MMICL-12.1B: Yes

`Does the following sentence describe the image? There are no propellers on the plane.`
`Answer: No`

`Does the following sentence describe the image? There are no tires on the car.`
`Answer: No`

`Does the following sentence describe the image? There are no wheels visible on the plane.`

OpenFlamingo-3B: `Answer: No`
OpenFlamingo-3B I: `Answer: No`
OpenFlamingo-4B: `Answer: No`
OpenFlamingo-4B I: `Answer: No`
Idefics-9B: `Answer: No`
Idefics-9B I: `Answer: No`
Idefics2-8B: No
xGen-MM-4.6B-7B: `Answer: No`
Qwen-VL-Chat-9.6B: No
MMICL-12.1B: No

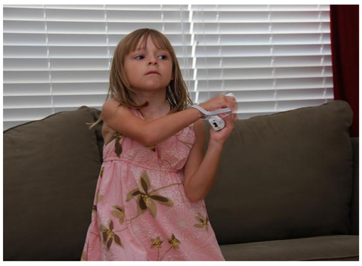
Figure 4: Example model predictions on instances from the *Existence* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

Does the following sentence describe the image? Some people race their horses down a beach.
Answer: Yes

Does the following sentence describe the image? The child sits on a number of horses in the pasture.
Answer: No

Does the following sentence describe the image? Exactly one horse stands on rocks near a river.

OpenFlamingo-3B: Answer: No Does the following sentence ✓
OpenFlamingo-3B I: Answer: No ✓
OpenFlamingo-4B: Answer: No ✓
OpenFlamingo-4B I: Answer: Yes ✗
Idefics-9B: Answer: Yes ✗
Idefics-9B I: Answer: Yes ✗
Idefics2-8B: Answer: No ✓
xGen-MM-4.6B-7B: Answer: No ✓
Qwen-VL-Chat-9.6B: No ✓
MMICL-12.1B: Yes ✗



Does the following sentence describe the image? A number of little girls are intently playing the video game. Answer: No

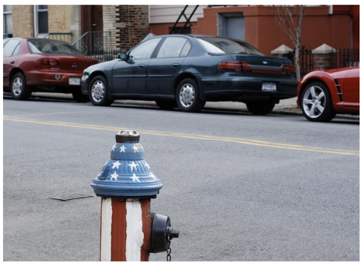
Does the following sentence describe the image? The woman is handing a single package to another person. Answer: Yes

Does the following sentence describe the image? Exactly one woman in uniform is talking on a cell phone.

OpenFlamingo-3B: Answer: Yes ✓
OpenFlamingo-3B I: Answer: No ✗
OpenFlamingo-4B: Answer: Yes ✓
OpenFlamingo-4B I: Answer: Yes ✓
Idefics-9B: Answer: Yes ✓
Idefics-9B I: Answer: Yes ✓
Idefics2-8B: Yes ✓
xGen-MM-4.6B-7B: Answer: Yes ✓
Qwen-VL-Chat-9.6B: Yes ✓
MMICL-12.1B: Yes ✓

Figure 5: Example model predictions on instances from the *Plurality* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

`Does the following sentence describe the image? There are exactly 8 vehicles on the street.`
`Answer: Yes`



`Does the following sentence describe the image? There are exactly 11 cars parked.`
`Answer: No`

`Does the following sentence describe the image? There are exactly 4 cars parked.`

OpenFlamingo-3B: `Answer: No Does the following sentence` ✗
OpenFlamingo-3B I: `Answer: No` ✗
OpenFlamingo-4B: `Answer: Yes` ✓
OpenFlamingo-4B I: `Answer: Yes` ✓
Idefics-9B: `Answer: No` ✗
Idefics-9B I: `Answer: No` ✗
Idefics2-8B: `Answer: No` ✗
xGen-MM-4.6B-7B: `Answer: No` ✗
Qwen-VL-Chat-9.6B: Yes ✓
MMICL-12.1B: No ✗

`Does the following sentence describe the image? There are exactly 3 lights above the mirror.`
`Answer: No`

`Does the following sentence describe the image? There are exactly 6 chairs.`
`Answer: Yes`

`Does the following sentence describe the image? There are exactly 6 lamps.`

OpenFlamingo-3B: `Answer: Yes Does the following sentence` ✗
OpenFlamingo-3B I: `Answer: No` ✓
OpenFlamingo-4B: `Answer: Yes` ✗
OpenFlamingo-4B I: `Answer: Yes` ✗
Idefics-9B: `Answer: Yes` ✗
Idefics-9B I: `Answer: No` ✓
Idefics2-8B: Answer: No ✓
xGen-MM-4.6B-7B: `Answer: No` ✓
Qwen-VL-Chat-9.6B: No ✓
MMICL-12.1B: No ✓

Figure 6: Example model predictions on instances from the *Counting* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

```
Does the following sentence describe the image?
Two small clocks sit behind a glass window.
Answer: Yes
```

```
Does the following sentence describe the image?
A woman walking into a shop filled with
merchandise.
Answer: No
```

```
Does the following sentence describe the image?
There are many vases on display outside the
building.
```

OpenFlamingo-3B: `Answer: Yes Does the following sentence` ✓

OpenFlamingo-3B I: `Answer: No` ✗

OpenFlamingo-4B: `Answer: Yes` ✓

OpenFlamingo-4B I: `Answer: Yes` ✓

Idefics-9B: `Answer: Yes` ✓

Idefics-9B I: `Answer: No` ✗

Idefics2-8B: `Answer: No` ✗

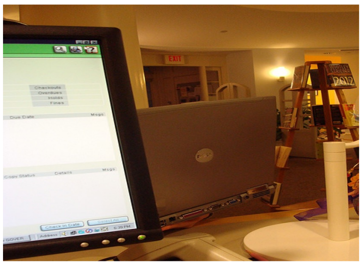
xGen-MM-4.6B-7B: `Answer: No` ✗

Qwen-VL-Chat-9.6B: `No` ✗

MMICL-12.1B: `No` ✗

```
Does the following sentence describe the image?
A computer mouse is beside a notebook computer.
Answer: Yes
```

```
Does the following sentence describe the image?
The laptop was left open under the desk.
Answer: No
```



```
Does the following sentence describe the
image? A computer is lit down on the checkout
counter.
```

OpenFlamingo-3B: `Answer: No Does the following sentence` ✓

OpenFlamingo-3B I: `Answer: No` ✓

OpenFlamingo-4B: `Answer: No` ✓

OpenFlamingo-4B I: `Answer: Yes` ✗

Idefics-9B: `Answer: No` ✓

Idefics-9B I: `Answer: No` ✓

Idefics2-8B: `No` ✓

xGen-MM-4.6B-7B: `Answer: No` ✓

Qwen-VL-Chat-9.6B: `Yes` ✗

MMICL-12.1B: `No` ✓

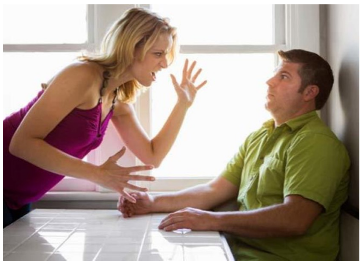
Figure 7: Example model predictions on instances from the *Spatial Relations* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

Does the following sentence describe the image?
A man burns his rubbish.
Answer: Yes

Does the following sentence describe the image?
A man leaps to the ground.
Answer: No

Does the following sentence describe the image?
A man ducks the cruiser.

OpenFlamingo-3B: Answer: No ✅
OpenFlamingo-3B I: Answer: Yes ❌
OpenFlamingo-4B: Answer: No ✅
OpenFlamingo-4B I: Answer: No ✅
Idefics-9B: Answer: No ✅
Idefics-9B I: Answer: No ✅
Idefics2-8B: No ✅
xGen-MM-4.6B-7B: Answer: No ✅
Qwen-VL-Chat-9.6B: No the man is walking ✅
MMICL-12.1B: No ✅



Does the following sentence describe the image?
A woman gardens a man.
Answer: No

Does the following sentence describe the image?
A man interrogates a man.
Answer: Yes

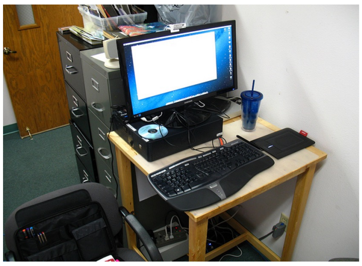
Does the following sentence describe the image?
A man confronts a man.

OpenFlamingo-3B: Answer: Yes ❌
OpenFlamingo-3B I: Answer: No ✅
OpenFlamingo-4B: Answer: No ✅
OpenFlamingo-4B I: Answer: Yes ❌
Idefics-9B: Answer: No ✅
Idefics-9B I: Answer: No ✅
Idefics2-8B: Answer: No ✅
xGen-MM-4.6B-7B: Answer: Yes ❌
Qwen-VL-Chat-9.6B: No ✅
MMICL-12.1B: Yes ❌

Figure 8: Example model predictions on instances from the *Actions* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

Does the following sentence describe the image? A computer monitor expressing disappointment on the screen with mouse. Is it laptop? Yes.
Answer: No

Does the following sentence describe the image? The computer is on a wooden computer desk. Is it a laptop? Yes.
Answer: No



Does the following sentence describe the image? A computer on a very small table in an office. This computer is a laptop? Yes.

OpenFlamingo-3B: Answer: Yes Does the following ✗
OpenFlamingo-3B I: Answer: Yes ✗
OpenFlamingo-4B: Answer: Yes ✗
OpenFlamingo-4B I: Answer: Yes ✗
Idefics-9B: Answer: No ✓
Idefics-9B I: Answer: Yes ✗
Idefics2-8B: Answer: No ✓
xGen-MM-4.6B-7B: Answer: No ✓
Qwen-VL-Chat-9.6B: No ✓
MMICL-12.1B: Yes ✗

Does the following sentence describe the image? An image of a person slicing pizza with a knife. Is it pepperoni pizza? No.
Answer: Yes

Does the following sentence describe the image? A open box of pizza placed on a kitchen counter. Is this a whole pizza? Yes.
Answer: Yes

Does the following sentence describe the image? A couple sitting at a table having pizza and beverages. Are they outside? No.

OpenFlamingo-3B: Answer: Yes ✓
OpenFlamingo-3B I: Answer: Yes ✓
OpenFlamingo-4B: Answer: Yes ✓
OpenFlamingo-4B I: Answer: Yes ✓
Idefics-9B: Answer: Yes ✓
Idefics-9B I: Answer: Yes ✓
Idefics2-8B: Answer: Yes ✓
xGen-MM-4.6B-7B: Answer: Yes ✓
Qwen-VL-Chat-9.6B: Yes ✓
MMICL-12.1B: Yes ✓

Figure 9: Example model predictions on instances from the *Coreference* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

```
Does the following sentence describe the image?
A large commercial airplane parked on the
runway.
Answer: Yes
```

```
Does the following sentence describe the image?
An airplane that is sitting in the water.
Answer: No
```

```
Does the following sentence describe the image?
A single prop red and white car with four
wings.
```

OpenFlamingo-3B: `Answer: Yes` ✗

OpenFlamingo-3B I: `Answer: Yes` ✗

OpenFlamingo-4B: `Answer: Yes` ✗

OpenFlamingo-4B I: `Answer: Yes` ✗

Idefics-9B: `Answer: No` ✓

Idefics-9B I: `Answer: Yes` ✗

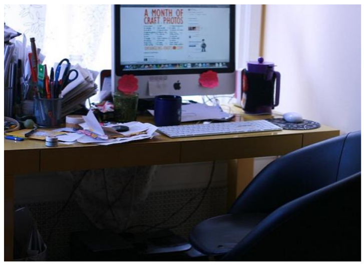
Idefics2-8B: `Yes` ✗

xGen-MM-4.6B-7B: `Answer: Yes` ✗

Qwen-VL-Chat-9.6B: `No` ✗

MMICL-12.1B: `Yes` ✗

```
Does the following sentence describe the image?
There is a mp3 player that is next to a cell
keyboard.
Answer: No
```



```
Does the following sentence describe the image?
A cluttered desk with a black chair next to it.
Answer: Yes
```

```
Does the following sentence describe the image?
A cell phone in a holder attached to a desk.
```

OpenFlamingo-3B: `Answer: Yes Does the following sentence` ✓

OpenFlamingo-3B I: `Answer: Yes` ✓

OpenFlamingo-4B: `Answer: Yes` ✓

OpenFlamingo-4B I: `Answer: Yes` ✓

Idefics-9B: `Answer: Yes` ✓

Idefics-9B I: `Answer: Yes` ✓

Idefics2-8B: `Yes` ✓

xGen-MM-4.6B-7B: `Answer: Yes` ✓

Qwen-VL-Chat-9.6B: `Yes` ✓

MMICL-12.1B: `Yes` ✓

Figure 10: Example model predictions on instances from the *Foil-It!* task, with demonstrations selected based on both visual and textual similarity (setting **S**).

Does the following sentence describe the image? There are servings on the plate.
Answer: The image shows a plate with what appears to be flatbread or pizza topped with various ingredients, including what looks like greens, possibly arugula. The sentence "There are servings on the plate" is a bit ambiguous, but it seems to suggest that there are portions of food on the plate. Based on the image, it is clear that there are portions of food, which could be considered servings. Therefore, the sentence is true for the image. Final Answer: Yes

Does the following sentence describe the image? There are pieces of food on the plate.
Answer: The image shows a young girl sitting at a table with a plate in front of her. On the plate, there are several pieces of what appear to be rolled-up food items, possibly burritos or similar dishes. The girl is looking at the plate with a thoughtful or contemplative expression. Given the content of the image, the sentence "There are pieces of food on the plate" is true. The plate contains multiple pieces of food, which are clearly visible and identifiable as food items. Final Answer: Yes

Does the following sentence describe the image? There are plates shown.

OpenFlamingo-3B: The image shows a woman sitting at a table with a plate in front of her. On the plate, there are several pieces of what appears to be flatbread or pizza topped with various ingredients, including what looks like greens, possibly arugula or spinach, and possibly some vegetables like broccoli and red peppers.

OpenFlamingo-3B I: The image shows a woman sitting at a table with a plate in front of her. On the plate, there are several pieces of what appear to be rolled-up food items, possibly burritos or similar dishes. The woman is looking at the plate with a thoughtful or contemplative expression.

OpenFlamingo-4B: The image shows a woman sitting at a table with a plate in front of her. On the plate, there are several pieces of what appear to be rolled-up food items, possibly burritos or similar dishes.

OpenFlamingo-4B I: The image shows a woman sitting at a table with a plate in front of her. On the plate, there are several pieces of what appear to be rolled-up food items, possibly burritos or similar dishes. The woman is looking at the plate with a thoughtful or contemplative expression.

Idefics-9B: The image shows a woman sitting at a table with a plate in front of her. On the plate, there is a cake or other dessert item, as well as a cup of tea. The sentence "There are plates shown" is a bit ambiguous, but it seems to suggest that there are plates visible in the image. Based on the image, it is clear that there are plates, which could be considered shown. Therefore, the sentence is true for the image. Final Answer: Yes

Idefics-9B I: Yes, there are plates in the image. The woman in the picture is holding a plate with a cake on it, and there are other plates on the table, including a white plate with a cake on it, a green plate with a cake on it, and a red plate with a cake on it. The sentence "There are plates shown" is true, as there are multiple plates visible in the image. Final Answer: Yes

Idefics2-8B: The image shows a woman sitting at a table with a plate in front of her. On the plate, there are two slices of cake, a cup of tea, and a small bowl. The cake slices are clearly visible and identifiable as food items. Given the content of the image, the sentence "There are plates shown" is true. The image contains a plate with food items on it, including cake, tea, and a small bowl. Final Answer: Yes

xGen-MM-4.6B-7B: The image shows a round loaf of bread with raisins and other ingredients mixed in. There are no people visible, but there is an empty plate next to the cake that suggests it might have been eaten or served recently by someone who isn't pictured here (the person taking this photo). The sentence "There arent any pictures" could be interpreted as false for some contexts because we can see what appears like food on top/underneath something else; however since nothing specific about those items being consumed has yet appeared within our viewfinder then I would say they don

Qwen-VL-Chat-9.6B: Yes, there are two plates shown in the image. One is placed on the left side of the table and the other one is placed on the right side of the table. Both the plates contain food items. Final Answer: Yes

MMICL-12.1B: The image shows a man standing in front of a table with several plates of food. Therefore, the sentence "There are plates shown" is true. Final Answer: Yes

Figure 11: Example model predictions on instances from the *Existence* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).

Does the following sentence describe the image? There is a single skateboarder doing tricks on the street.
Answer: The image shows two skateboarders. One skateboarder is in the air, performing a trick, while the other skateboarder is on the ground, holding a skateboard. The skateboarder in the air is the one doing the trick, and the other skateboarder is not performing any tricks at the moment. Therefore, the sentence "There is a single skateboarder doing tricks on the street" is false. Final Answer: No

Does the following sentence describe the image? A young man riding a single skateboard with red wheels.
Answer: The image shows a young man standing on a skateboard with red wheels. He is not actively riding the skateboard, but rather standing still. The skateboard is positioned on the ground, and the young man appears to be looking at something or someone to his left. The wheels are indeed red, and there is no indication of multiple skateboards in the image. Final Answer: Yes

Does the following sentence describe the image? A skateboarder is turning his board at the top of exactly one ramp.

OpenFlamingo-3B: The image shows a skateboarder performing a trick at the top of a ramp. The skateboard is positioned on a surface, and the person's other foot is not visible in the frame. The skateboard has red wheels and appears to be in a state of use, as indicated by the wear and tear on the deck.

OpenFlamingo-3B I: The image shows a skateboarder turning his board at the top of a ramp. The ramp appears to be a concrete ramp, and there are no other skateboards visible in the image. The sentence "A skateboarder is turning his board at the top of exactly one ramp" is false. Final Answer: No

OpenFlamingo-4B: The image shows a skateboarder performing a trick on a ramp. The skateboard is positioned on the ground, and the skateboarder appears to be turning his board at the top of the ramp. There is no indication of multiple skateboards in the image, so the statement "A skateboarder is turning his board at the top of exactly one ramp" is false. Final Answer: No

OpenFlamingo-4B I: The image shows a skateboarder performing a trick on a ramp. The skateboarder is in mid-air, and he is turning his skateboard at the top of the ramp. There are no other skateboarders visible in the image, so the statement "A skateboarder is turning his board at the top of exactly one ramp" is not accurate for this particular image

Idefics-9B: The image shows a skateboarder performing a trick at the top of a ramp. The skateboarder is in mid-air, and the skateboard appears to be in a state of use, as indicated by the wear and tear on the deck. The sentence "A skateboarder is turning his board at the top of exactly one ramp" is true for the image. The skateboarder is in mid-air, and the skateboard appears to be in a state of use, as indicated by the wear and tear on the deck. The skateboarder is turning the board at the top of a ramp, and there is only one ramp visible in the image. Final Answer: Yes

Idefics-9B I: The image shows a skateboarder performing a trick at the top of a ramp. The skateboarder is in mid-air, and the skateboard is positioned at an angle, indicating that the skateboarder is in the middle of a turn. The ramp is clearly visible in the image, and there is no indication of multiple ramps or skateboarders. Final Answer: Yes

Idefics2-8B: The image shows a skateboarder performing a trick on a ramp. The skateboarder is turning the board at the top of the ramp, and there is no indication of multiple ramps or skateboards in the image. Therefore, the sentence is true for this image. Final Answer: Yes

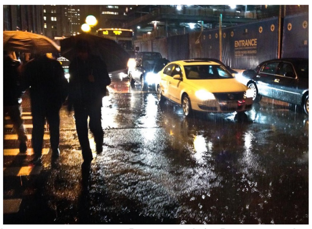
xGen-MM-4.6B-7B: The image shows a skateboarder in the middle of performing an action on his board. He is not turning it at all, but rather appears to be balancing or preparing for another trick while standing upright with one foot forward and holding onto both sides simultaneously (as indicated by "at least two hands"). The ramp he stands upon has graffiti written across its surface; however this detail does nothing towards determining whether there are multiple ramps present as suggested within Sentence 2's statement that only 'exactly ONE RAMP'. Final answer = No

Qwen-VL-Chat-9.6B: The image shows a skateboarder performing a trick on a ramp. The skateboarder is at the top of the ramp, and the ramp is visible in the image. The skateboarder is not turning the board, but rather performing a trick. Therefore, the sentence "A skateboarder is turning his board at the top of exactly one ramp" is false. Final Answer: No. There is no turning of the board in the image. The skateboarder is performing a trick on the ramp.

MMICL-12.1B: The image shows a skateboarder turning his board at the top of a ramp. The skateboarder is positioned in the middle of the

Figure 12: Example model predictions on instances from the *Plurality* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).

Does the following sentence describe the image? There are exactly 4 buildings.
Answer: The image shows a scene with several buildings. There are at least four buildings visible in the image: 1. The tall building with a clock tower on the left. 2. The red brick building to the right of the clock tower. 3. The building with the "BAHNHOF POTSDAM" sign in the center. 4. The building with the red brick facade to the right of the sign. There may be more buildings out of the frame, but based on what is visible in the image, there are at least four buildings. Final Answer: Yes



Does the following sentence describe the image? There are exactly 5 vehicles on the road.
Answer: To analyze the image and the sentence, let's count the vehicles on the road: 1. There is a yellow car in the foreground. 2. There is a black car behind the yellow car. 3. There is a white car behind the black car. 4. There is a bus behind the white car. 5. There is a truck behind the bus. So, there are indeed five vehicles on the road. Final Answer: Yes

Does the following sentence describe the image? There are exactly 0 cars parked.

OpenFlamingo-3B: To analyze the image and the sentence, let's count the number of cars in the image:\n\n1. There is a white car parked on the left side of the image. 2. There is a red car parked on the right side of the image. 3. There is a black car parked on the left side of the image. 4. There is a white car parked on the right side of the image. 5. There is a red car parked on the left side of the image…

OpenFlamingo-3B I: To analyze the image and the sentence, let's count the number of parked cars:

OpenFlamingo-4B: To determine if the sentence is true or false, let's count the number of cars in the image: 1. There is a white car on the left side of the image. 2. There is a white car in the center of the image. 3. There is a white car on the right side of the image…

OpenFlamingo-4B I: To determine if the sentence is true or false, let's count the number of cars in the image: 1. There is a white car on the left side of the image. 2. A red car is visible in the middle of the image. 3. A green car is on the right side of the image. 4. There is a white car in the center of the image. 5. A red car is on the left side of the image. 6. A white car is on the right side of the image. 7. A red car is on the right side of the image…

Idefics-9B: To analyze the image and the sentence, let's count the cars parked: 1. There is a car parked on the left side of the image. 2. There is a car parked on the right side of the image. 3. There is a car parked in the distance on the right side of the image. 4. There is a car parked in the distance on the left side of the image. 5. There is a car parked in the distance on the right side of the image. 6. There is a car parked in the distance on the left side of the image…

Idefics-9B I: To analyze, let's count the number of cars parked on the street: 1. There are two cars parked on the street. 2. The cars are parked in a line, and I can clearly see the number plates on some of them. 3. I will count the visible cars from the left side of the image to the right. Based on the image, there are indeed two cars parked on the street. Therefore, the sentence "There are exactly 0 cars parked" is false. Final Answer: No

Idefics2-8B: To analyze the image and the sentence, let's count the visible cars: 1. There is a white car on the left side of the image. 2. There is a white car on the right side of the image. 3. There is a red car on the left side of the image. 4. There is a red car on the right side of the image. 5. There is a white car on the left side of the image. 6. There is a white car on the right side of the image. Based on the image, there are indeed five cars visible. Final Answer: No

xGen-MM-4.6B-7B: To determine if the sentence is true or false, let's count how many cars are parked in this image. 1) There appears to be a car on each side of where these two people stand (one red and one white). 2.) The street has no visible traffic lights so we can assume that there arent any other vehicles driving down it at present time; however they may have been stopped before entering our viewpoint area which would make them invisible from here but still technically considered cparking because their wheels were touchin

Qwen-VL-Chat-9.6B: To analyze the image and the sentence, let's count the number of cars parked: 1. There is a white car parked on the left side of the image. 2. There is a white car parked on the right side of the image. 3. There is a white car parked behind the white car on the right. 4. There is a white car parked behind the white car on the right. 5. There is a white car parked behind the white car on the right. 6. There is a white car parked behind the white car on the right. 7. There is a white car parked behind the white car on the right. Adding up the cars, there are more than 7 cars parked. Final Answer: No. There are exactly 0 cars parked.

MMICL-12.1B: To determine if the sentence is true or false, let's count the number of cars in the image: 1. There are no cars visible in the

Figure 13: Example model predictions on instances from the *Counting* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).

Does the following sentence describe the image? A cow stands on a sidewalk in a building.

Answer: The image shows a cow standing on a sidewalk next to a building with a blue door. The cow is not inside the building; it is outside, on the sidewalk. Therefore, the sentence "A cow stands on a sidewalk in a building" is false. Final Answer: No

Does the following sentence describe the image? A man dressed up as a clown is holding a cellphone to his ear.

Answer: The image shows a person dressed in a clown costume, which includes a colorful jacket, a hat, and face paint. The clown is holding a cellphone to their ear. The person is standing next to a vehicle with a sign that reads "CALIFORNIA" on it. Based on the content of the image, the sentence "A man dressed up as a clown is holding a cellphone to his ear" is true. The clown is indeed holding a cellphone to their ear, as depicted in the image. Final Answer: Yes

Does the following sentence describe the image? A lady walks down the street with bananas on her head.

OpenFlamingo-3B: The image shows a woman walking down the street with a large bunch of bananas on her head. The woman is wearing a straw hat, and the bananas are on her head. The woman is walking down the street with a large bunch of bananas on her head…

OpenFlamingo-3B I: The image shows a woman walking down the street with bananas on her head. The woman is wearing a hat, and the bananas are on her head. The woman is on the left side of the image, and she is indeed wearing a hat. The woman is on the right side of the image. Based on the content of the image, the sentence "A lady walks down the street with bananas on her head" is true. Final Answer: Yes

OpenFlamingo-4B: The image shows a woman walking down the street with a large bunch of bananas on her head. The woman is walking in the same direction as the street, but she is walking on the sidewalk, which is separate from the road where the street is. The woman is not carrying the bananas on her head in the sense that she is not overtaking them; they are simply moving in parallel

OpenFlamingo-4B I: The image shows a lady walking down the street with bananas on her head. The bananas are not attached to the lady's head, and they are not attached to each other either. The bananas are not attached to the lady's head, and they are not attached to each other either

Idefics-9B: The image shows a woman walking down the street with bananas on her head. She is wearing a hat, and the bananas appear to be wrapped in a cloth. She is also carrying a bag, which suggests she might be out for errands or returning from work. Based on the content of the image, the sentence "A lady walks down the street with bananas on her head" is true. The woman is indeed walking down the street with bananas on her head, as depicted in the image. Final Answer: Yes

Idefics-9B I: The image shows a woman walking down the street with bananas on her head. She is wearing a hat made out of banana leaves, and the bananas are resting on top of the hat. The woman is also wearing a pink shirt, and there are other people visible in the background. Final Answer: Yes

Idefics2-8B: The image shows a woman carrying bananas on her head. She is walking down a street, and the bananas are visible on her head. The woman is wearing a hat, and the bananas are piled on top of her head. The sentence provided states, "A lady walks down the street with bananas on her head. " This sentence is accurate for the image. Final Answer: Yes.

xGen-MM-4.6B-7B: The image shows a woman walking down the street with bananas on her head. The lady is wearing glasses and has long black hair, which she ties up in pigtails to keep it out of their way while carrying them…

Qwen-VL-Chat-9.6B: The image shows a woman walking down the street with a basket of bananas on her head. The woman is indeed wearing a hat, and the bananas are not in a bag. The woman is walking down the street, and the bananas are on her head, not in a bag. Final Answer: Yes

MMICL-12.1B: The lady is walking down the street with bananas on her head. This is an unusual and creative way to transport bananas. Final Answer: Yes

Figure 14: Example model predictions on instances from the *Spatial Relations* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).

Does the following sentence describe the image? A man stumbles for a cellular telephone.

Answer: The image shows a man holding a cellular phone in his hand. He appears to be looking at the phone, possibly checking it or using it. There is no indication in the image that the man is stumbling or struggling to find the phone. The phone is clearly visible in his hand, and he seems to be in control of it. Final Answer: No

Does the following sentence describe the image? A man instructs a colleague.

Answer: The image shows two individuals in a setting that appears to be a workshop or factory, with one person pointing at a machine or control panel while the other looks on. The person pointing seems to be explaining or showing something to the other person. Based on this visual context, the sentence "A man instructs a colleague" is consistent with the scene depicted in the image. Final Answer: Yes

Does the following sentence describe the image? A man lifts his torso.

OpenFlamingo-3B: The image shows a person lifting a barbell. The person's torso is visible, and the barbell appears to be resting on the person's chest, which is consistent with the action described in the sentence.

OpenFlamingo-3B I: The image shows a man lifting his torso. The man's torso is bent at the waist, which is consistent with the action described in the sentence

OpenFlamingo-4B: The image shows a person lifting a barbell with their torso. The barbell appears to be resting on the ground, and the person appears to be lifting the barbell with their torso. Based on the content of the image, it is plausible that the person is lifting the barbell with their torso. However, without additional context or information, it is not possible to definitively confirm that the person is lifting the barbell with their torso.

OpenFlamingo-4B I: The image shows a person lifting a barbell with their torso. The barbell appears to be resting on the person's chest, and the person's head is turned slightly to the side. The person appears to be engaged in lifting the barbell, which is consistent with the sentence.

Idefics-9B: Answer:The image shows a man lifting a weighted barbell over his head. The barbell is resting on the man's shoulders, and the man's torso is bent at the waist. Now, let's analyze the sentence: "A man lifts his torso". The image depicts a person lifting a weighted barbell over their head, which is consistent with the action described in the sentence. Final Answer: Yes

Idefics-9B I: The image shows a man lifting a weighted barbell off the ground while standing on a field. The man's torso is lifted off the ground as he performs the lift. Now, let's analyze the sentence: "A man lifts his torso". The image depicts a man lifting his torso, which is consistent with the action described in the sentence. Final Answer: Yes

Idefics2-8B: The image shows a man lifting a barbell with two large tires attached to it. The man is wearing a shirt that says "Science" on it, and he is lifting his torso to lift the weight. Based on the content of the image, it is plausible that the man is lifting his torso to lift the weight. However, without additional context or information, it is not possible to definitively confirm that the man is lifting his torso. Final Answer: Yes

xGen-MM-4.6B-7B: The image shows a person lifting weights, with their torso and arms prominently displayed. The individual is wearing athletic clothing suitable for weightlifting or…

Qwen-VL-Chat-9.6B: The image shows a man lifting a barbell off the ground. The man is standing and holding the bar with both hands, while his torso is upright. The focus of the image is on the man and the barbell, so it is not accurate to say that the man is "lifting his torso." Instead, he is engaged in the activity of lifting the barbell. Final Answer: No A man lifts his torso. Answer: The image shows a man lifting a barbell off the ground. The man is standing and holding the bar with both hands, while his torso is upright. The focus of the image is on the man and the barbell, so it is not accurate to say that the man is "lifting his torso." Instead, he is engaged in the activity of lifting the barbell.

MMICL-12.1B: The image shows a man lifting his torso while standing on a bench. The man is bending his knees and extending his

Figure 15: Example model predictions on instances from the *Actions* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).

Does the following sentence describe the image? There is a woman taking a bite out of some bread. Is she wearing glasses? Yes
Answer: The image shows a woman taking a bite out of what appears to be a piece of bread or pastry. However, the sentence provided states that she is wearing glasses, which is not true based on the image. The woman in the image is not wearing glasses. Final Answer: No

Does the following sentence describe the image? A woman sits on the grass with dogs all around. Is the woman in her backyard? Yes
Answer: The image shows a woman sitting on the grass with several dogs around her. She is not in her backyard, as there are no indications of a residential setting or personal property in the background. The woman is in a park or a similar public space, as evidenced by the open grassy area and the presence of other people and dogs in the background. Final Answer: No

Does the following sentence describe the image? Woman in purple shirt examines her cellphone in the open field. Is it big field? Yes

OpenFlamingo-3B: The image shows a woman in a purple shirt examining her cellphone in an open field. The woman is standing on a hill with a cityscape in the background. The sky is blue, and the clouds are visible in the sky. The woman is holding a cellphone in her hand, which is visible in the image. The sky is blue, and the clouds are visible in the sky. Therefore, the sentence "is it big field? yes" is true. Final Answer: Yes.

OpenFlamingo-3B I: The image shows a woman in a purple shirt examining her cell phone in the open field. The woman is standing on a grassy hill, and there is a blue sky in the background. The field is large, as indicated by the open space and the presence of trees in the background. The woman is holding a cell phone in her right hand, and she appears to be looking at the screen of the phone.

OpenFlamingo-4B: The sentence provided is: "A woman in a purple shirt examines her cell phone in the open field. is it sunny? yes. Answer:The image shows a woman in a purple shirt examining her cell phone in the open field. The woman is standing on a grassy hill, and there is a blue sky in the background. The field is large, as indicated by the open space and the presence of trees in the background. The woman is holding a cell phone in her right hand, and she appears to be looking at the screen of the phone. The sentence provided is: "A woman in a purple shirt examines her cell phone in the open field. is it sunny? yes." Based on the image, the first part of the sentence is true, as there is a blue sky in the background. The second part of the sentence is also true, as the woman is holding a cell phone in her right hand, and she appears to be looking at the screen of the phone. Final Answer: Yes

OpenFlamingo-4B I: The image shows a woman in a purple shirt standing in the open field, looking at her cellphone. The field appears to be large, as indicated by the presence of trees and mountains in the background. The woman appears to be standing in the middle of the field, and there is a clear sky in the background. The image is clear and well-lit, indicating that it was taken during the day

Idefics-9B: The image shows a woman in a purple shirt examining her cellphone in an open field. The woman appears to be standing on a grassy hill, and there are other people and objects in the background. The woman is wearing sunglasses, which suggests it might be sunny. The presence of a car and a clear sky further supports the daytime setting. The sentence provided is: "Woman in purple shirt examines her cellphone in the open field. is it big field? yes". Based on the image, the first part of the sentence is true, as the woman is examining her cellphone in an open field. The second part of the sentence is also true, as the field appears to be large. Final Answer: Yes

Idefics-9B I: The image shows a woman in a purple shirt examining her cellphone in an open field. The field appears to be quite large and spacious, with mountains visible in the background. The woman seems to be enjoying her time outdoors while using her phone. The sentence provided is: "Woman in purple shirt examines her cellphone in the open field. is it big field? yes". Based on the image, the first part of the sentence is true, as the woman is examining her cellphone in the open field. The second part of the sentence is also true, as the field is indeed big. Final Answer: Yes

Idefics2-8B: The image shows a woman in a purple shirt examining her cell phone in an open field. The field is indeed big, as it stretches out in front of her and includes various trees and plants. The sky is visible in the background, and the woman is standing on a grassy area. Final Answer: Yes

xGen-MM-4.6B-7B: The image shows a woman standing in an open field with green grass and trees. She is wearing glasses, earphones around her neck or headset on the phone's handsfree feature while she examines it closely using both hands to hold up one side of its screen for better viewing angle as well look at other features like buttons etc.. … Final answer yes

Qwen-VL-Chat-9.6B: Yes, the image shows a woman standing in a large open field. The field is surrounded by trees and hills, and the woman is holding a cell phone in her hand. The woman is wearing a purple shirt and appears to be taking a picture of the landscape with her phone. The image is clear and well-lit, indicating that it was taken during the day. Final Answer: Yes…

MMICL-12.1B: The woman is standing in a field with a cellphone. The field is open and large enough for her to see the phone clearly. Final Answer:

Figure 16: Example model predictions on instances from the *Coreference* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).

Does the following sentence describe the image? An horse walking through underbrush next to a fence.

Answer: The image shows an elephant, not a horse, walking through underbrush next to a fence. Therefore, the sentence is false. Final Answer: No

Does the following sentence describe the image? Two girls bathe an elephant lying on its side.

Answer: The image shows two individuals, who appear to be young girls, interacting with an elephant that is lying on its side in water. The girls are standing in the water and seem to be washing or bathing the elephant. The elephant is partially submerged, and the girls are reaching out to it. The girls are indeed bathing the elephant, which is lying on its side in the water. Final Answer: Yes

Does the following sentence describe the image? Man riding an elephant into water surrounded by forest.

OpenFlamingo-3B: The image shows a man riding an elephant into water surrounded by forest. The man is standing on the elephant's back, and the elephant appears to be moving through the water. The elephant is surrounded by greenery, and there are trees in the background. The elephant appears to be moving through the water, which could be described as a forest or a jungle-like environment

OpenFlamingo-3B I: The image shows a man riding an elephant into a body of water, which is surrounded by trees. The man is sitting on the elephant's back, and he appears to be holding on to the elephant's mane. The elephant is partially submerged, and the man is reaching out to it. Based on the content of the image, the sentence "Man riding an elephant into water surrounded by forest" is true. The man is indeed riding the elephant into the water, which is surrounded by trees. Final Answer: Yes

OpenFlamingo-4B: The image shows a woman sitting at a table with a plate in front of her. On the plate, there are several pieces of what appear to be rolled-up food items, possibly burritos or similar dishes.

OpenFlamingo-4B I: The image shows a man riding an elephant into a body of water surrounded by trees. The man is sitting on the elephant's back, and the elephant appears to be swimming in the water. The man is wearing a blue shirt and orange pants, and he appears to be smiling.

Idefics-9B: The image shows a man riding an elephant into water surrounded by forest. The man is sitting on the elephant's back, and the elephant is walking into the water. There are trees and other vegetation in the background. The sentence provided is false because it describes a man riding an elephant, which is not present in the image. The correct description would be "A man riding an elephant into water surrounded by forest". Final Answer: No

Idefics-9B I: The image shows a man riding on the back of an elephant that is walking into a body of water. The elephant and the man are surrounded by trees and vegetation, indicating that they are in a forested area. The man appears to be guiding the elephant into the water. Based on the content of the image, the sentence "Man riding an elephant into water surrounded by forest" is true. The man is indeed riding on the back of an elephant, and they are walking into a body of water surrounded by trees and vegetation Final Answer: Yes

Idefics2-8B: The image shows a man riding an elephant into water surrounded by a forest. The man is sitting on the elephant's back, and the elephant is walking into the water. The forest is visible in the background, and the elephant is partially submerged in the water. The sentence provided is true for the image. Final Answer: Yes

xGen-MM-4.6B-7B: The image shows a man riding an elephant into water surrounded by trees. The scene appears to be in the jungle or forest, and there are other animals visible as well such\nas monkeys on branches above them . There is also another person sitting next him who seems like he might have some sort of connection with this activity (perhaps they work together).

Qwen-VL-Chat-9.6B: The image shows a man riding an elephant into a body of water, with trees surrounding the area. The elephant is walking through the water, and the man appears to be enjoying the ride. The scene is quite picturesque, with the elephant and rider creating a unique and interesting image. The presence of trees and the water body also adds to the beauty of the scene.

MMICL-12.1B: The image shows a man riding an elephant into a body of water surrounded by forest. The man is positioned on the elephant's back

Figure 17: Example model predictions on instances from the *Foil-It!* task, where demonstrations are selected based on visual and textual similarity, and Chain-of-Thought (CoT) reasoning is employed (setting **S+C**).