EVALUATING ZERO-SHOT LEARNING CAPABILITIES OF VISION-LANGUAGE MODELS

GÖRME-DİL MODELLERİNİN SIFIR-ÖRNEKLE ÖĞRENME YETENEKLERİNİN DEĞERLENDİRİLMESİ

MUSTAFA DOĞAN

PROF. DR. MEHMET ERKUT ERDEM
Supervisor

Submitted to
Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

June 2024

ABSTRACT

EVALUATING ZERO-SHOT LEARNING CAPABILITIES OF VISION-LANGUAGE MODELS

Mustafa Doğan
Master of Science, Computer Engineering
Supervisor: Prof. Dr. MEHMET ERKUT ERDEM
June 2024, 108 pages

Vision-Language Models (VLMs) stand at the forefront of artificial intelligence research, aiming to bridge the gap between visual content and natural language understanding. Their significance lies in their potential to enable machines to comprehend and interact with the world in a more human-like manner. However, the evaluation of VLMs poses twofold challenges that require careful consideration and innovative approaches.

One of the primary challenges in evaluating VLMs revolves around understanding the intricate relationship between visual and linguistic information. While these models are good at processing individual modalities, such as images, videos, or text, effectively integrating these modalities to derive meaningful insights remains a complex task. Particularly in dynamic and context-rich scenarios, VLMs must navigate diverse visual stimuli while interpreting accompanying textual cues, requiring robust mechanisms for cross-modal fusion and comprehension.

Furthermore, the lack of transparency in VLMs adds another layer of complexity to their evaluation. While these models may exhibit high performance on benchmark datasets, understanding the underlying reasoning processes and knowledge representations remains elusive. Deciphering how VLMs leverage their learned knowledge to generate responses and make predictions is essential for gaining insights into their capabilities and limitations.

This thesis addresses these challenges by conducting a comprehensive comparative analysis of Multimodal Large Language Models (MLLMs) and Video-Language Models (VidLMs). It focuses on their ability to bridge the semantic gap between visual inputs and linguistic outputs. Through empirical evaluation, this research examines the strengths and limitations of these models in comprehending and articulating visual content in both static and dynamic contexts.

This thesis makes two main contributions. Firstly, it conducts a comprehensive analysis of few-shot In-Context Learning (ICL) and Chain-of-Thought (CoT) strategies on MLLMs, revealing that these strategies can significantly boost performance compared to zero-shot settings. Secondly, it introduces a novel zero-shot foiling test for VidLMs, designed to assess their proficiency in recognizing actions and actors within dynamic scenes. The findings indicate that current VidLMs face challenges in temporal reasoning and action recognition, performing only marginally better than chance, thereby highlighting the imperative for advancements in VidLMs architectures to effectively handle spatio-temporal tasks.

In conclusion, this thesis sheds light on the performance of MLLMs and VidLMs, offering valuable insights and identifying areas for future improvement. It indicates the importance of ongoing innovation in multimodal architectures to develop more robust and contextually aware language models capable of bridging the gap between visual content and natural language.

Keywords: Chain-of-Thought Reasoning, In-Context Learning, Temporal Reasoning, Multimodal Architectures, Vision Language Models

ÖZET

GÖRME-DİL MODELLERİNİN SIFIR-ÖRNEKLE ÖĞRENME YETENEKLERİNİN DEĞERLENDİRİLMESİ

Mustafa Doğan
Yüksek Lisans, Bilgisayar Mühendisliği
Danışman: Prof. Dr. MEHMET ERKUT ERDEM
Eylül 2021, 108 sayfa

Görme-Dil Modelleri (VLM'ler), görsel içerik ve doğal dil anlayışı arasındaki boşluğu doldurmayı amaçlayan yapay zeka araştırmalarının ön saflarında yer almaktadır. Makinelerin dünyayı daha insani bir şekilde anlamalarını ve etkileşime girmelerini sağlama potansiyelinde VLM'lerin önemi yer almaktadır. Bununla birlikte, VLM'lerin değerlendirilmesi, dikkatli bir değerlendirme ve yenilikçi yaklaşımlar gerektiren iki yönlü zorluklar ortaya çıkarmaktadır.

VLM'lerin değerlendirilmesindeki temel zorluklardan biri, görsel ve dilsel bilgiler arasındaki karmaşık ilişkiyi anlamakla ilgilidir. Bu modeller görüntüler, videolar veya metinler gibi tek tek modaliteleri işlemede iyi olsa da, anlamlı içgörüler elde etmek için bu modaliteleri etkili bir şekilde entegre etmek karmaşık bir görev olmaya devam etmektedir. Özellikle dinamik ve bağlam açısından zengin senaryolarda, VLM'lerin eşlik eden metinsel ipuçlarını yorumlarken çeşitli görsel uyaranlarda gezinmesi gerekir, bu da modlar arası füzyon ve anlama için sağlam mekanizmalar gerektirir.

Ayrıca, VLM'lerdeki şeffaflık eksikliği, değerlendirmelerine başka bir karmaşıklık katmanı ekler. Bu modeller kıyaslama veri kümelerinde yüksek performans gösterebilirken, altta yatan muhakeme süreçlerini ve bilgi temsillerini anlamak zor olmaya devam etmektedir. VLM'lerin yanıt üretmek ve tahminlerde bulunmak için öğrenilmiş bilgilerinden nasıl yararlandıklarını deşifre etmek, yetenekleri ve sınırlamaları hakkında içgörü kazanmak için çok önemlidir.

Bu tez, Çok Modlu Büyük Dil Modelleri (MLLM'ler) ve Video-Dil Modellerinin (VidLM'ler) kapsamlı bir karşılaştırmalı analizini yaparak bu zorlukları ele almaktadır. Tez, görsel girdiler ile dilsel çıktılar arasındaki anlamsal boşluğu doldurma becerilerine odaklanmaktadır. Bu araştırma ampirik değerlendirme yoluyla hem statik hem de dinamik bağlamlarda görsel içeriği anlama ve ifade etmede bu modellerin güçlü yönlerini ve sınırlamalarını incelemektedir.

Bu tezin iki ana katkısı bulunmaktadır. İlk olarak, MLLM'ler üzerinde birkaç atışla bağlam içi öğrenme ve düşünce zinciri stratejilerinin kapsamlı bir analizini yaparak, bu stratejilerin performansı sıfır atış öğrenmeye göre önemli ölçüde artırabileceğini ortaya koyuyor. İkinci olarak, bu tez VidLM'ler için dinamik sahnelerdeki eylemleri ve aktörleri tanıma yeterliliklerini değerlendirmek üzere tasarlanmış yeni bir sıfır atış engelleme testi sunuyor. Bulgular, mevcut VidLM'lerin zamansal muhakeme ve eylem tanıma konusunda zorluklarla karşılaştığını, şanstan yalnızca marjinal olarak daha iyi performans gösterdiğini ve böylece VidLM mimarilerinde uzamsal-zamansal görevleri etkili bir şekilde ele almak için ilerlemelerin zorunluluğunu vurgulamaktadır.

Sonuç olarak, bu tez MLLM'lerin ve VidLM'lerin performansına ışık tutmakta, değerli içgörüler sunmakta ve gelecekteki iyileştirme alanlarını belirlemektedir. Görsel içerik ve doğal dil arasındaki boşluğu doldurabilecek daha sağlam ve bağlamsal farkındalığa sahip dil modelleri geliştirmek için çok modlu mimarilerde devam eden yeniliklerin önemine işaret etmektedir.

Anahtar Kelimeler: Düşünce Zinciri Akıl Yürütme, Bağlam İçi Öğrenme, Zamansal Akıl Yürütme, Çok Modlu Mimariler, Görme-Dil Modelleri

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Mehmet Erkut Erdem, and my co-advisor, Aykut Erdem. Their exceptional working style and discipline have been a source of inspiration for me throughout this journey. Balancing employment and my Master's studies was a challenging task, but their understanding and support made it manageable. Prior to working with them, I had less enthusiasm for NLP. However, their innovative ideas and unique perspectives ignited my passion for this research area, and for that, I am immensely grateful.

I extend my sincere appreciation to my jury members for taking time out of their busy schedules to evaluate my work. Their invaluable feedback and insights have greatly contributed to the refinement and completion of this thesis.

I am also thankful to my collaborators, İlker Kesen and Emre Can Açıkgöz. Our collaborative efforts were marked by excellent teamwork, and their support and contributions were crucial to the success of our project.

To my team members at the workplace, your support was indispensable. Balancing a full-time job with my Master's studies was a daunting task, and your understanding and assistance made this journey smoother and more manageable.

Special thanks to KUIS AI Center for providing access to their HPC resources. Despite not being obligated to assist, your generosity allowed me to complete the experiments necessary for this thesis using your state-of-the-art hardware.

My heartfelt gratitude goes to my family for their unwavering support throughout this period. Your encouragement and belief in me have been a constant source of strength.

Finally, I would like to thank my fiancée, Elif Kıvrak. Your unwavering support and faith in me, even during the most challenging times when I might have been unbearable, have been invaluable. This thesis would not have been possible without your patience, love, and belief in my abilities.

To my lovely family

CONTENTS

Page

TABLES

FIGURES

ABBREVIATIONS

AI	:	**A**rtifical **I**ntelligence
BERT	:	**B**idirectional **E**ncoder **R**epresen**t**ations
CNN	:	**C**onvolutional **N**eural **N**etworks
CoT	:	**C**hain **o**f **T**hought
DL	:	**D**eep **L**earning
ICL	:	**I**n **C**ontext **L**earning
ILMs	:	**I**mage **L**anguage **M**odel**s**
ITM	:	**I**mage **T**ext **M**atching
LM	:	**L**anguage **M**odel
LSTM	:	**L**ong **S**hort **T**erm **M**emory
MLM	:	**M**asked **L**anguage **M**odeling
MLLM	:	**M**ultimodal **L**arge **L**anguage **M**odel
NLG	:	**N**atural **L**anguage **G**eneration
NLI	:	**N**atural **L**anguage **I**nference
NLP	:	**N**atural **L**anguage **P**rocessing
ResNet	:	**R**esidual Neural **Net**work
VidLMs	:	**Vid**eo **L**anguage **M**odel**s**
VLMs	:	**V**ision **L**anguage **M**odel**s**

1. INTRODUCTION

Vision-Language Models (VLMs) stand at the intersection of computer vision and natural language processing, embodying the pursuit to guide machines with the ability to understand, generate, and interact with visual content through language. This developing field not only encompasses image language models but also extends to video language models, which tackle the temporal dimension of visual data. Over the past few years, VLMs have undergone a transformative evolution, marked by significant advancements in architecture, methodology, and performance. This evolution reflects a paradigm shift from early image captioning systems to sophisticated multimodal models capable of tackling many vision-language tasks with unprecedented accuracy and versatility.

In their developing stages, VLMs primarily relied on conventional machine learning techniques and handcrafted features to bridge the semantic gap between images and natural language. However, the advances in deep learning revolutionized the field, paving the way for data-driven approaches that could learn rich representations directly from raw data [2–5]. This shift catalyzed the development of encoder-decoder architectures [6–11] with attention mechanisms [12–14], enabling models to generate descriptive captions for images with greater fluency and coherence. Through iterative refinements and breakthroughs in neural network architectures, VLMs evolved from simple captioning systems to sophisticated multimodal transformers capable of jointly processing visual and textual inputs. Furthermore, as video data became increasingly common on the internet and the need for video comprehension grew, the scope of VLMs expanded to include temporal dynamics. This evolution paved the way for models capable of not only analyzing individual frames but also understanding the temporal relationships and semantic context across multiple frames over time.

The evolution of VLMs has been accompanied by an augmentation of architectures tailored to address specific vision-language tasks, ranging from image captioning and visual question answering to video summarization and multimodal translation [15–27]. Modern VLMs leverage large-scale pretraining techniques to learn semantically meaningful representations from vast amounts of paired image-text and video-text data. These models exhibit remarkable versatility and generalization across a diverse array of vision-language tasks, indicating their potential to bridge the gap between perception and language in artificial intelligence.

1.1. Scope Of The Thesis

The scope of this thesis encompasses an analysis of the performance of both Multimodal Large Language Models (MLLMs) and Video-Language Models (VidLMs). Through empirical investigation and evaluation, this thesis aims to assess the language understanding capabilities of these models, focusing on their proficiency in bridging the semantic gap between visual content and natural language. This research seeks to provide insights into the strengths and limitations of MLLMs and VidLMs in comprehending and articulating the content of visual inputs. Additionally, this thesis aims to explore how advancements in deep learning and multimodal architectures have influenced the efficacy of these models, contributing to a detailed understanding of their performance in real-world applications.

1.2. Contributions

The primary contributions of this thesis can be outlined as follows:

- Conducted a comprehensive analysis on the impact of few-shot ICL and CoT strategies on the performance of Multimodal Large Language Models, evaluating models pretrained on captioning or interleaved image-text datasets.
- Developed a novel zero-shot foiling test specifically tailored for assessing Video-Language Models, focusing on their ability to recognize actions and actors in a scene. The test includes a proficiency test alongside the main test to provide a range of difficulty levels for evaluation.

1.3. Organization

The organization of the thesis is as follows:

- Chapter 1. introduces the research's motivation, outlines its contributions, and defines the scope of the thesis.
- Chapter 2. provides related work, including an examination of pretrained VLMs as well as few-shot In-Context Learning and Chain-of-Thought strategies.
- Chapter 3. outlines language understanding capabilities of MLLMs using few-shot approaches.
- Chapter 4. explores our zero shot evaluation test for VidLMs.
- Chapter 5. concludes the thesis by summarizing the findings and potential future directions for research in this field.

2. RELATED WORK

2.1. Pretrained Vision-Language Models

This section provides a comprehensive overview of research on pretrained Vision-Language Models, delineated into two distinct domains: 2.1.1. Multimodal Large Language Models and 2.1.2. Video-Language Models.

2.1.1. Multimodal Large Language Models

We classify MLLMs based on three factors: learning approaches, datasets used in pretraining, and model architectures.

Learning Strategies. The evolution of VLMs has transitioned from manual image descriptions to using transformer architectures for integrated image and text processing. Advanced pretraining techniques are crucial to this progress as they give models a rich understanding of both textual and visual content, increasing their versatility across various tasks. The synergy between text and image processing is essential for identifying complex connections and providing accurate representations. We will explore common pretraining methods and techniques that enhance the performance of VLMs, while also discussing key insights and considerations for further improving these advanced systems.

Contrastive Loss. Building on the effectiveness of contrastive learning, recent endeavors have utilized this pretraining objective to bridge the vision and language. Notable examples [28–31] employ contrastive loss to jointly train text and image encoders on extensive datasets comprising image-caption pairs. Contrastive learning minimizes the distance between embeddings of matching image-text pairs while maximizing it for non-matching pairs, thereby aligning the feature spaces of images and texts. While CLIP [28] calculates distance using cosine similarity, ALIGN [30] and DeCLIP [31] designed metrics to accommodate noisy datasets. LiT [32] introduces a method where the text encoder is fine-tuned using

CLIP's pretraining objective while keeping the image encoder fixed. This technique enhances the text encoder's ability to interpret image embeddings, proving to be more sample-efficient than CLIP. Additionally, approaches like FLAVA [33] leverage a blend of contrastive learning and other pretraining strategies to synchronize vision and language embeddings effectively.

PrefixLM. Introducing an alternative method for training VLMs, the PrefixLM objective is employed by models like SimVLM [34] and VirTex [35], featuring a unified multi-modal architecture reminding of autoregressive language models, comprising a transformer encoder and decoder. This approach operates on the premise of predicting subsequent tokens given a preceding text segment. Applied to images, visual transformers (ViT) break down images into patches, sequentially feeding them into the model as inputs. SimVLM, for instance, adopts this concept, where the encoder processes concatenated image patch and text sequence prefixes, with the decoder forecasting the text continuation. While models like SimVLM demonstrates good performance in image-conditioned text generation and VQA tasks, those solely relying on the PrefixLM strategy may find their utility constrained to image captioning and visual question-answering tasks. Conversely, models incorporating multi-modal representations or hybrid approaches exhibit versatility across a spectrum of tasks, encompassing object detection and image segmentation, beyond mere textual inference from visual input.

Multi-modal Fusing with Cross Attention. An emerging approach to utilizing pretrained language models for multi-modal tasks involves directly integrating visual information into the layers of a language model decoder through cross-attention mechanisms, bypassing the need for images as additional prefixes. Pioneered by models like VisualGPT [36], VC-GPT [37], and Flamingo [38], this strategy aims to balance text generation capabilities and visual information fusion, particularly in the absence of extensive multi-modal datasets. VisualGPT employs a visual encoder to embed images, channeling these embeddings into the cross-attention layers of a pretrained language decoder module to produce coherent captions. Recent advancements, such as FIBER [39], push the envelope further by integrating cross-attention layers with gating mechanisms into both vision and language backbones, enhancing the efficiency of multi-modal fusion and enabling diverse downstream tasks such as image-text retrieval and open-vocabulary object detection.

Masked-Language Modeling / Image-Text Matching. Another pretraining approach in VLMs area is Masked-Language Modeling (MLM) and Image-Text Matching (ITM) objectives, aligning specific image segments with corresponding text to facilitate diverse downstream tasks like visual question answering, commonsense reasoning, and text-guided object detection. Models adopting this pretraining approach [33, 40–43] demonstrate the efficacy of combining MLM and ITM objectives. MLM involves predicting masked words in a partially obscured caption based on the accompanying image, typically necessitating richly annotated multi-modal datasets or object detection models for generating region proposals. Conversely, ITM focuses on determining whether a given caption matches the associated image, with negative samples randomly drawn from the dataset. These objectives are often jointly employed during pretraining, as exemplified by VisualBERT [40], which incorporates a BERT-like architecture leveraging pretrained object detection models like Faster-RCNN [44] for implicit alignment of text and image elements through MLM and ITM objectives. Similarly, FLAVA [33] adopts a transformer-based framework comprising image, text, and multi-modal encoders, employing a mixture of pretraining objectives including MLM, ITM, Masked-Image Modeling (MIM), and contrastive learning to enhance multi-modal reasoning and alignment.

No Training. Lastly, various optimization techniques strive to bridge image and text representations by leveraging pretrained models or adapting multi-modal models for new tasks without further training. For instance, MaGiC [45] introduces iterative optimization by employing a pretrained autoregressive language model to generate captions for input images. This process involves computing a “Magic score” based on CLIP embeddings [28] of the generated tokens and the input image. On the other hand, ASIF [46] presents a straightforward approach to transforming pretrained uni-modal image and text models into a multi-modal model for image captioning, leveraging a compact multi-modal dataset without additional training. The underlying principle of ASIF lies in the assumption that captions for similar images exhibit similarity, thus enabling a similarity-based search by constructing a relative representation space using a small dataset of ground-truth multi-modal pairs.

Datasets. VLMs are commonly trained on extensive image and text datasets, each structured according to the pretraining objective. We explore the insights into common pretraining dataset types used for training and evaluating VLMs.

Captioning Datasets. Captioning datasets consist of collections of images paired with descriptive text, enabling models to learn and generate accurate textual descriptions based on visual inputs. Multimodal Large Language Models (MLLMs) frequently utilize image-text pair datasets due to their numerous advantages: they are straightforward to use, establish a direct link between text and images, and include well-established, widely adopted, and standardized datasets such as MS COCO [47], Flickr30k [48], LAION-5B [49], and CC12M [50]. These datasets are pivotal for pretraining MLLMs, enabling them to comprehend and generate accurate text based on visual inputs. By offering a clear correlation between images and their textual descriptions, these datasets facilitate the development of models proficient in precise image captioning and fundamental visual comprehension tasks. Additionally, their standardized nature ensures consistency and comparability across various models and research endeavors, making them indispensable for the progression of multimodal AI research.

Interleaved Image-Text Datasets. In contrast, interleaved image-text datasets [51–54] provide a context involving multiple images and texts, enabling models to utilize this context to solve more complex tasks. This approach empowers models to address new challenges, such as narrating a sequence of images. Training with interleaved image-text datasets allows MLLMs to develop a deeper understanding of the interactions between multiple visual and textual elements, resulting in more sophisticated and contextually relevant outputs. These datasets are essential for tasks that require a comprehensive understanding of the sequence and interplay of images and texts, enhancing the models' performance in scenarios that mirror real-world complexities. Furthermore, interleaved datasets support the advancement of narrative generation and complex scene understanding, extending the capabilities of MLLMs in contextual comprehension and generation.

Pretraining Datasets. VLMs typically are pretrained on vast multi-modal datasets sourced from the internet, comprising matched image or video content alongside textual annotations. Textual data within these datasets may include human-generated captions, automatically generated descriptions, image metadata, or object labels. Prominent examples of such expansive datasets include PMD PMD [33] and LAION-5B [49]. PMD combines smaller datasets like Flickr30K [48], COCO [47], and Conceptual Captions [55]. COCO encompasses over 330,000 images annotated with object labels and natural sentence descriptions, while Conceptual Captions and Flickr30K feature millions of images scraped from the web, each paired with descriptive captions. Even datasets solely comprising human-generated captions, like Flickr30K, exhibit inherent noise due to varied user practices in captioning. To mitigate this, datasets such as LAION-5B employ CLIP [28] or other pretrained multi-modal models for noise filtration, ensuring the creation of high-quality multi-modal datasets. Additionally, some models like ALIGN [30] propose preprocessing steps and curate their datasets to enhance quality. Other datasets, such as LSVTD [56] and WebVid [57], integrate video and text modalities on a smaller scale.

Downstream Datasets. Pretrained VLMs are commonly adapted to various downstream tasks spanning visual question-answering [58, 59], text-guided object detection, image inpainting, multi-modal classification, as well as standalone NLP and computer vision tasks. Models tailored for question-answering tasks predominantly rely on datasets that feature images paired with open-ended questions and answers [60–65]. Notably, datasets like Vizwiz [65] and TextCaps [64] extend their utility to image segmentation and object localization tasks. Other noteworthy multi-modal datasets include Hateful Memes for classification, SNLI-VE [66] for visual entailment prediction, and Winoground [67] for visio-linguistic compositional reasoning. Furthermore, VLMs find applications in classical NLP and computer vision tasks like text or image classification, often leveraging uni-modal datasets [68, 69]. Moreover, datasets like COCO and Conceptual Captions serve dual purposes, being utilized both in model pretraining and downstream caption generation tasks.

Evaluation Benchmarks. Several novel benchmarks have been introduced to evaluate the capabilities of pretrained VLMs. VALSE (Vision And Language Structured Evaluation) [1] offers a suite of tests covering various linguistic constructs, providing a finer-grained evaluation than previously possible. CREPE [70] focuses on evaluating compositionality, measuring systematicity and productivity across different datasets. The Visual-Spatial Reasoning (VSR) [71] dataset provides a comprehensive evaluation of models' abilities to understand spatial relations expressed in natural language. Lastly, the Basic Language Abilities (BLA) benchmark [72] evaluates models on basic linguistic constructions such as active-passive voice, coordination, and relative clauses, shedding light on their understanding of image-text interaction. These diverse benchmarks contribute to a deeper understanding of the strengths and limitations of pretrained VLMs across different linguistic and visual domains.

Models. The development of Multimodal Large Language Models (MLLMs) has seen significant progress, with advancements leveraging pretrained autoregressive LLMs and sophisticated visual encoders to process both text and visual inputs. Prominent examples such as Flamingo [38] have shown exceptional performance across various vision-language tasks. This progress has facilitated the creation of open-weight models, promoting collaboration and accessibility within the field [54, 73–79]. Models like IDEFICS [52, 80] surpass inference efficiency and stable training by utilizing pretrained unimodal backbones. Similarly, Qwen-VL Chat [81], built on Qwen-7B, focuses on fine-grained visual understanding and multilingual support, achieving state-of-the-art results. In contrast, LLaVA-NeXT [82], an enhanced version of LLaVA-1.5 [83], uses a highly effective and data-efficient vision-language integration module that requires only a simple fully-connected projection layer trained on a modest dataset. While Qwen-VL employs specially designed visual resamplers trained on large amounts of image-text paired data, LLaVA-NeXT achieves state-of-the-art performance using publicly available data, demonstrating both efficiency and effectiveness in model design and training. MMICL [54] addresses current model limitations by efficiently managing multi-modal inputs, including the relationships among multiple images and text-to-image references. By introducing a novel context scheme and a comprehensive multi-modal ICL dataset, MMICL enhances the understanding of complex text-image relationships and multi-image reasoning. Models like MANTIS [77] address multi-image visual language tasks through instruction tuning with academic-level resources, leading to notable performance improvements. InternVL [84] bridges the gap between LLMs and vision-language foundation models, scaling up vision foundation models and achieving state-of-the-art results across various visual-linguistic benchmarks. InternLM-XComposer2 [85] demonstrates advancements in free-form text-image composition and comprehension, using a Partial LoRA approach to balance vision understanding and text composition, thus surpassing multimodal content creation and understanding.

Despite the progress, challenges remain, particularly in optimizing techniques like In-Context Learning for MLLMs, as highlighted by the limited research in this area [86, 87]. Nonetheless, the rapid evolution of MLLMs holds promise for advancing multimodal AI systems and their applications in various domains. These diverse models signify the ongoing progress and exploration within the field of multimodal AI, offering promising avenues for future research and development.

2.1.2. Video-Language Models

Research on pretrained Video-Language Models (VidLMs) spans various dimensions, including modalities, datasets, learning paradigms, modeling architectures, and performance assessments. By synthesizing insights from these diverse perspectives, researchers can develop a detailed understanding of VidLMs and their potential impact.

Modalities and Datasets. Recent developments in VidLMs have shifted towards leveraging a diverse array of modalities and datasets for more robust pretraining. Unlike earlier models that predominantly relied on images, videos, and textual data, contemporary VidLMs are incorporating a wider spectrum of inputs, including speech and audio, and even exploring emerging modalities like haptic or sensor data [88–100]. This expansion broadens the scope of applications and presents new challenges in understanding the interaction between different modalities and datasets.

The choice of training data for VidLMs is crucial and often depends on the type of pretraining utilized for the visual modality. Earlier models leaned towards datasets like HowTo100M [101], which provided linguistic input via automatic speech recognition (ASR) or manually written subtitles. Recent advancements have shifted towards more extensive datasets such as WebVid-2M [102], Conceptual Captions [55], and large-scale image-text corpora like SBU captions [103]. HowTo100M offers a large-scale dataset comprising 136 million video clips sourced from instructional web videos, while WebVid-2M proposes an end-to-end trainable model designed to leverage large-scale image and video captioning datasets. Conceptual Captions provides a dataset of 3.3 million images annotated with captions harvested from the web's alt-text HTML attribute, offering a variety of styles for image captioning tasks. VidSitu [15] introduces a framework for understanding salient events in videos through visual semantic role labeling, presenting a benchmark dataset of movie clips richly annotated with verbs and semantic roles to advance research in video understanding.

Recent advancements in VidLMs have expanded beyond conventional visual and linguistic modalities to encompass a broader spectrum of sensory inputs. For instance, emerging models integrate haptic data [99, 100], such as tactile and kinesthetic feedback, to enrich multi-modal understanding and enable applications in virtual reality, robotics, and human-computer interaction.

Learning Paradigms and Architectures. The training and adaptation of VidLMs involve a combination of pretraining objectives and fine-tuning strategies, guiding the initial learning process and adapting the pretrained models to downstream tasks. These objectives include video-text contrastive loss, video-text matching, masked language modeling, masked frame modeling, natural language generation, masked visual-token modeling, and temporal reordering [92, 93, 104, 105]. Architectures for VidLMs encompass temporal modeling techniques, multimodal fusion mechanisms, and computational efficiency considerations, reflecting advancements in recurrent neural networks, attention mechanisms, and graph-based models. Some methods employ joint space-time attention to process video [102, 104, 106], while others rely on a multi-modal attention mechanism between patches and word embeddings [92, 96, 107]. Models may incorporate additional multi-modal transformers or fuse a visual prefix into text-only language models [105, 108].

Combining pretraining objectives and fine-tuning strategies into a single category focusing on learning paradigms can elucidate the overarching training and adaptation methodologies of VidLMs [91, 96, 98, 102]. On the other hand, combining temporal modeling and multimodal fusion into a single category centered on modeling architectures can provide a unified perspective on how VidLMs process temporal dynamics and integrate information across modalities [93, 109, 110].

Performance Assessment and Robustness. Evaluation metrics are pivotal in assessing the proficiency and effectiveness of VidLMs in understanding and generating textual descriptions for video content. Metrics commonly employed include BLEU (Bilingual Evaluation Understudy) [111], which measures the similarity between generated descriptions and human-written references. METEOR (Metric for Evaluation of Translation with Explicit Ordering) [112] gauges the quality of generated text by considering exact word matches, synonyms, and paraphrases. Additionally, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [113] evaluates the overlap between generated and reference texts based on n-gram overlap and recall. Furthermore, CIDEr (Consensus-based Image Description Evaluation) [114] considers consensus among human annotators and the diversity of generated descriptions. These metrics collectively provide a comprehensive framework for assessing VidLM performance, guiding advancements in their development and refinement [115–118].

Beyond numerical assessments, evaluating VidLMs also involves analyzing their robustness and generalization capacities, critical for their practical utility across real-world scenarios. This includes examining their performance under diverse conditions, such as domain shifts, adversarial perturbations, and encounters with noisy or incomplete data. To enhance VidLMs' resilience, researchers employ techniques like data augmentation and adversarial training, aimed at fortifying the models against unforeseen challenges [119, 120]. By embracing this comprehensive evaluation framework, the development of VidLM advances with a focus on ensuring their reliability, adaptability, and effectiveness in diverse applications and contexts, driving progress toward more sophisticated and versatile models for video understanding and generation.

2.2. In-Context Learning

In-Context Learning (ICL) involves crafting prompts for models that include task examples expressed in natural language. With ICL, it is possible to leverage pretrained models to address novel tasks without the need for fine-tuning. This examination of ICL will be divided into two main sections.

ICL in Multimodal Setting. The first attempt at ICL in multimodal setting is Frozen [86], which uses a fixed GPT-like language model. Flamingo [38] demonstrated improved ICL performance, handling a variety of interleaved text and image sequences. By using a masked cross-attention mechanism, Flamingo can manage an unlimited number of visual inputs and seamlessly integrate visual data into already trained language models. Due to certain limitations, several VLMs like BLIP [121] and MiniGPT-4 [122] are considered unsuitable for ICL. However, open-source Flamingo adaptations, such as OpenFlamingo [78] and IDEFICS [52], compete effectively in ICL. While utilizing Flamingo's model architecture and masked cross-attention for visual integration, Otter [74] employs instruction tuning to enhance task capability. Furthermore, [123] focused solely on image captioning trials, exploring improved in-context configurations without evaluating the importance of visual and textual data. Other efforts, like SINC [124] and MetaVL [125], attempt to reduce the need for extensive pretraining but still fall short of Flamingo's performance with pretrained VLMs.

Understanding ICL. LLMs have made significant progress in ICL [126–129], where models learn new tasks from a few contextual examples without requiring gradient updates. Studies aim to identify the critical elements of ICL examples for LLMs [130–134]. While the precision of the input-label mapping is not critical, exposure to the label space and the distribution of demonstrations substantially impact ICL performance. [131] emphasized how proper demonstration labels affect ICL performance, particularly under specific conditions. [131] also discovered that semantically similar instances to a test query can improve ICL performance, whereas [132] demonstrated the importance of order sensitivity in ICL performance. [130] examined the effects of demonstration complexity, variety, and similarity on ICL proficiency. Additionally, a few studies looked at ICL in LLMs from the perspective of model architecture [135, 136], clarifying the tight connection between model elements and ICL performance. However, ICL in VLMs differs due to unique model elements and the extra visual data in demonstrations. This thesis focuses on ICL with respect to VLMs to determine which information in multimodal demonstrations is more crucial.

2.3. Chain-of-Thought

Chain-of-Thought (CoT) prompting facilitates complex reasoning abilities by incorporating intermediate steps of reasoning. Employing CoT allows the utilization of pretrained models to tackle new tasks without requiring fine-tuning. When combined with ICL, it enhances performance on complex tasks requiring reasoning prior to generating responses. Our analysis of CoT will be segmented into two primary sections.

Multimodal CoT. The Multimodal Chain of Thought (M-CoT) framework extends the CoT idea, consisting of a sequence of intermediate thinking phases useful in challenging reasoning problems [137–139]. M-CoT extends this concept to multimodal contexts, attempting to mimic human cognitive processes by asking LLMs to explain not only their final replies but also the rationale behind them. This modification is crucial as it improves the interpretability and transparency of LLMs' outputs. Researchers investigate M-CoT through various learning paradigms, such as few-shot learning, zero-shot learning, and fine-tuning [140–143]. Each paradigm has unique benefits and drawbacks, considering sample size needs and computing performance.

The arrangement of reasoning chains in M-CoT is a key research area, with methods ranging from single-chain to tree-shaped structures, as well as decision points for either adaptive or pre-established chain lengths [141, 144–148]. Additionally, research is being conducted on reasoning chain generation patterns, exploring infilling or prediction-based techniques to ensure the accuracy and coherence of produced stages [140–142, 144–148]. Researchers aim to enhance multimodal reasoning skills and improve the interpretability and performance of AI systems in various fields by clarifying these aspects of M-CoT.

Understanding CoT. [149] investigates the effects of CoT prompting on LLMs' multi-step reasoning skills using isolated factor trials. Although CoT has successfully improved reasoning abilities by providing a series of stages in demonstrations, its overall efficacy and the specific contributions of these steps are still not well understood. The study shows that CoT reasoning is still possible with flawed examples, achieving significant performance gains comparable to correctCoT prompting. Critical components affecting CoT prompting efficacy include appropriate reasoning step sequencing and relevance to the input inquiry. Additionally, these results highlight the potential of contextual reasoning in LLMs, as further research has proven the usefulness and efficacy of contextual learning [150, 151].

3. PROBING LANGUAGE UNDERSTANDING CAPABILITIES OF MULTIMODAL LARGE LANGUAGE MODELS

3.1. Introduction

In this study, we explore the zero-shot and few-shot capabilities of pretrained MLLMs utilizing the VALSE benchmark [1]. VALSE is a pioneering benchmark designed to analyze the visio-linguistic grounding abilities of general-purpose pretrained VLMs across a spectrum of linguistic phenomena. Comprising six distinct tasks, VALSE offers a multifaceted suite for detailed evaluations, probing the models' capacity to bridge the gap between visual and linguistic modalities.

We conducted a comprehensive assessment of 14 state-of-the-art MLLMs, varying in model size and pretraining datasets. Four of these models were trained solely on captioning datasets, which consist of image descriptions, enabling models to generate captions for new images without additional examples. The remaining ten models were trained on both captioning and interleaved image-text datasets. Interleaved image-text datasets contain pairs of images and their associated text within a larger context. While models trained solely on captioning datasets only support zero-shot settings, those trained on interleaved image-text datasets can perform few-shot learning by leveraging the contextual relationships between images and text.

We investigate the linguistic capabilities of MLLMs on the VALSE benchmark, emphasizing the effectiveness of few-shot In-Context Learning (ICL) and Chain-of-Thought (CoT). ICL, a methodological cornerstone, involves acquainting the models with demonstration examples prior to presenting query examples. This capability first emerges in large language models [126], enabling them to adapt and respond effectively to diverse prompts. Notably, ICL is also utilized in multimodal settings [38, 102, 152, 153], where models integrate both text and image data to perform complex tasks such as image captioning and visual question answering. Our experiments under two distinct settings, random demonstration examples and similar demonstration examples to the query example, provide valuable insights into the efficacy of this approach in augmenting model comprehension and performance.

Chain-of-Thought [137] facilitates multi-step problem-solving by guiding models through intermediate steps of reasoning. This technique is particularly beneficial for tasks requiring logical thinking and multiple steps, such as arithmetic or commonsense reasoning questions. The integration of CoT into our methodology marks a significant advancement, facilitating complex reasoning abilities by incorporating intermediate steps of reasoning. By leveraging CoT, pretrained models can tackle new tasks with a heightened level of sophistication, all without the need for fine-tuning. We conduct experiments in few-shot scenarios, employing CoT in conjunction with ICL to amplify performance on tasks necessitating advanced reasoning prior to response generation.

Figure 3.1 illustrates our comprehensive evaluation of MLLMs using both In-Context Learning and Chain-of-Thought techniques.

3.2. Evaluation Strategy

In this section, we elaborate on our evaluation strategy on MLLMs. Section 3.2.1. offers a brief exploration of the VALSE benchmark, highlighting its significance as an evaluation framework for pretrained VLMs. Section 3.2.2. introduces the evaluation metric, outlining the criteria guiding our assessment of model performance. Sections 3.2.3. and 3.2.4. delve into the details of the ICL and CoT methodologies, elucidating their implementation and implications for model comprehension and performance enhancement. Finally, in Section 3.2.5., we provide an overview of the pretrained models under evaluation, highlighting their architectures and relevance to our experimental framework.

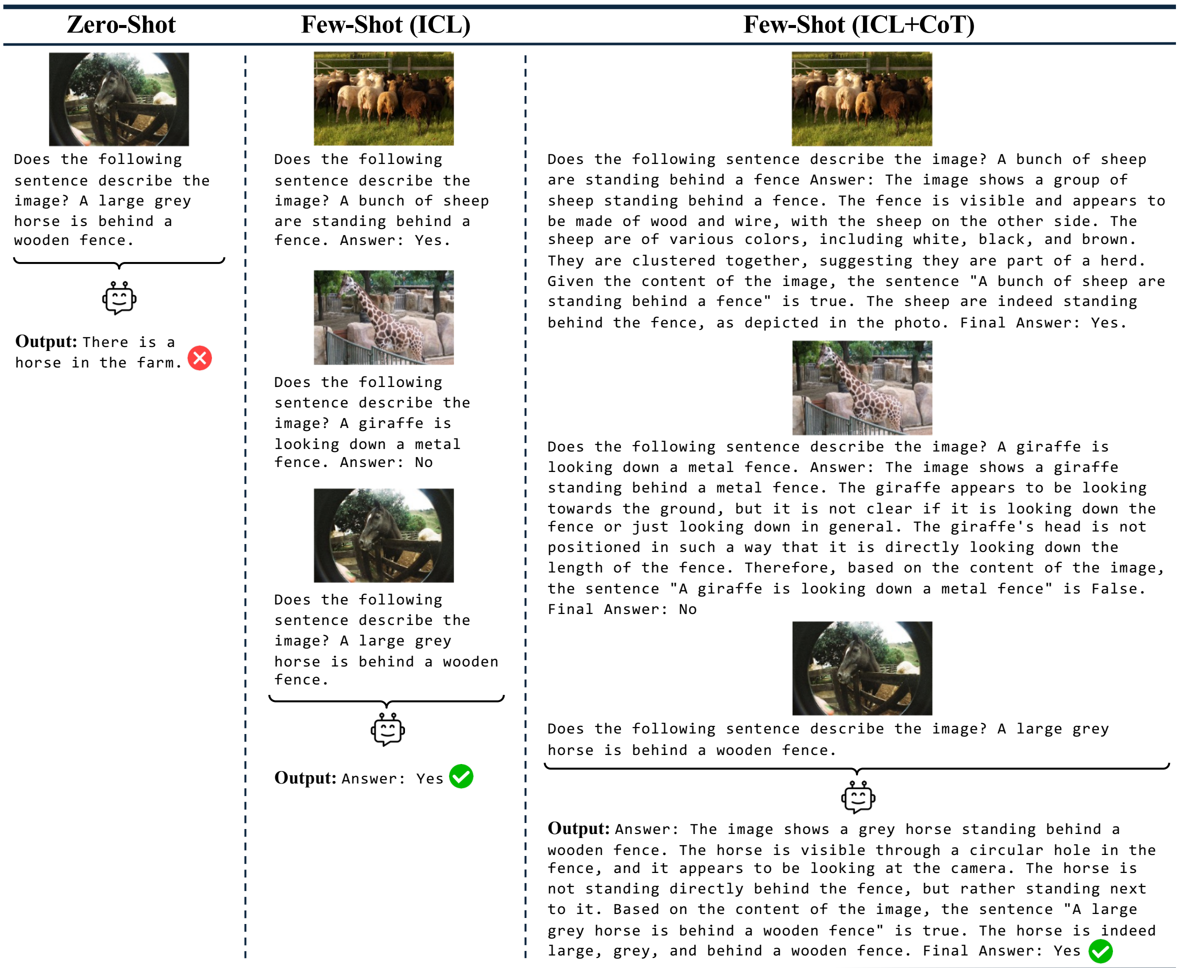


Figure 3.1 **Zero-Shot, Few-Shot ICL, and Few-Shot ICL+CoT Evaluation Comparison on the *Relations* Task.** In the Zero-Shot approach, the model incorrectly responds to the question. In the Few-Shot ICL approach, using prior examples, the model correctly identifies the horse behind a wooden fence. In the Few-Shot ICL+CoT approach, which is beneficial for tasks requiring intermediate reasoning steps such as counting, relational understanding, and coreference resolution, the model also correctly identifies the horse by employing a detailed step-by-step reasoning process.

3.2.1. VALSE Dataset

The VALSE benchmark [1] stands as a groundbreaking initiative aimed at assessing the capabilities of general-purpose pretrained VLMs in grounding linguistic constructs within the visual context. Developed to offer a comprehensive evaluation framework, VALSE encompasses six distinct tasks, each designed to probe the model's ability to bridge the gap between language and vision. These tasks include Existence, Plurality, Counting, Spatial

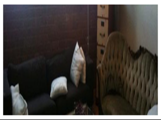
Existence	Plurality	Counting	Relations	Action	Coreference	Foil-It!
						
There are no people / people sitting on the couch.	A clock is seen at the top of exactly one / some very tall building.	There are exactly 5 / 3 lemons.	People are riding on elephants in / beside a river.	A lion stretches / arches its back.	A pretty lady sitting on a bench in the shade. Is she wearing a hat? No / Yes	The man is swinging a tennis racket / ball.

Figure 3.2 Sample instances from the VALSE benchmark [1].

Relations, Actions, and Coreference, each targeting specific linguistic phenomena essential for comprehensive understanding (see Figure 3.2).

- *Existence* task evaluates the model's capability to detect whether entities are present or absent in an image. It requires the model to distinguish between situations where objects are present or missing, focusing on existential quantifiers.

- *Plurality* task assesses the model's ability to recognize singular and plural forms by identifying whether images show single or multiple instances of objects. It evaluates the model's comprehension of semantic number distinctions.

- *Counting* task challenges the model to accurately enumerate the entities within an image. The complexity of the scenarios varies, testing the model's precise counting skills.

- *Spatial Relations* task examines the model's ability to identify and interpret spatial relationships between objects in an image, focusing on the arrangement and positions of items relative to each other.

- *Actions* task measures the model's proficiency in recognizing and understanding actions depicted in images. It involves identifying the activities and understanding the roles and interactions of the participants.

- *Coreference* task tests the model's ability to resolve pronoun references within the visual context. It evaluates whether the model can correctly link pronouns to the appropriate entities in the images, ensuring coherent interpretation.

Additionally, VALSE is based on the Foil-It dataset, which provides image-caption pairs from the COCO dataset. The Foil-It [154] dataset connects objects in the captions to the COCO [47] dataset. Leveraging these connections, VALSE generated foils and validated 943 out of 1000 captions.

To construct VALSE, detailed methodologies were employed to ensure the validity and effectiveness of the benchmark [7]. This involved developing robust criteria for generating valid foils [66], which are crucial for accurately assessing the model's performance. Through experimentation and evaluation of five widely-used V&L models, VALSE provides insights into the current challenges faced by pretrained models in understanding and interpreting linguistic phenomena in visual contexts. The benchmark not only highlights the existing limitations but also serves as a catalyst for driving future advancements in pretrained V&L models.

3.2.2. Evaluation Metric

[155] examines the effectiveness of the Image-Text Matching (ITM) prompting method within the CREPE benchmark [70], which shares similarities with the VALSE benchmark. This method presents a sentence to a model, labeling it as either a caption or a foil, and then inquires whether it accurately describes the corresponding image. Through this process, we measured the accuracy of models to gauge their performance.

3.2.3. In-Context Learning Approach

Few-shot ICL strives to enhance model performance by providing in-context demonstration examples related to the query image-text pair. The process of selecting these examples and determining their sequence remains a subject of ongoing research [130–132, 134, 156, 157]. In our investigation into the impact of in-context demonstrations on model performance, we experimented with models using both randomly chosen examples and those similar in visual and textual content [157]. This comparative analysis sheds light on the effectiveness of different approaches in leveraging contextual information to improve model understanding and performance.

Example Selection. We used the Mixed Modality In-Context Example Selection (MMICES) method outlined by [157] and detailed in Algorithm 1. Following this method, we assessed the textual and visual cosine similarity between each pair of image and text in the demonstration examples and the query pair. Utilizing Clip[1] as our encoder, we initially selected the top K visually similar examples. Given the relatively small size of the VALSE dataset, we opted not to partition it for creating a demonstration example set. Instead, we utilized the remaining dataset, excluding the query image-text pair under examination. From the selected K visually similar examples, we further filtered down to N examples exhibiting textual similarity. This N value ultimately represents the shot-count utilized in our experiments. However, determining the appropriate K value is a crucial and challenging task. As K increases, the model receives more examples with textual similarity. To investigate the impact of K on model performance, we conducted an ablation study. As higher K values yielded improved results, we set K to a high value of 100 for our experiments.

Algorithm 1 Mixed Modality In-Context Example Selection (MMICES) [157]

Require: QITP (Query Image-Text Pair), DES (Demonstration Example Set), K, N, V (Visual Encoder), T (Textual Encoder), cos_sim (Cosine Similarity)
Ensure: SES (Similar example set)

procedure MMICES
  SES ← []
  Similarities ← []
  for each example in DES **do**
    Encode the example image: $v_{example} \leftarrow V(I_{example})$.
    Encode example text: $t_{example} \leftarrow T(T_{example})$.
    Calculate cosine similarity: $sim_{image} \leftarrow cos\_sim(v_{example}, V(I_{QITP}))$.
    Calculate cosine similarity: $sim_{text} \leftarrow cos\_sim(t_{example}, T(T_{QITP}))$.
    Similarities $+=$ image and text similarity scores.
  end for
  Visual similar examples ← select top K visual similar example among similarities.
  SES ← select top N textual similar example among visual similar examples.
  return SES
end procedure

[1] https://huggingface.co/openai/clip-vit-base-patch32

Table 3.1 Rate of valid Chain-of-Thought (CoT) descriptions generated by the respective 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

3.2.4. Chain-of-Thought Approach

The CoT methodology is designed to boost model performance by encouraging reasoning during inference, particularly in scenarios with limited data. Initially, our exploration focused on zero-shot CoT, where models were tasked with generating reasoning independently, devoid of additional context. However, we observed that in such settings, models often arrived at final answers without engaging in meaningful reasoning. To remedy this, we integrated reasoning cues into our demonstration examples.

In addressing the need for detailed, nuanced descriptions in VALSE samples, we leveraged LLaVA-NeXT [82] for generating CoT descriptions within the context demonstrations. While this model surpasses producing rich captions, it occasionally introduces inaccuracies and fabricated details. To mitigate these challenges, we adopted a strategy proposed by [158], which directed models to generate both reasoning chains and answers, coupled with a validation step to curb hallucinations. Despite these precautions, some instances still lacked sufficiently detailed CoT descriptions even when the answers were correct. Consequently, we automatically removed examples with inaccurate answers or inadequate CoT descriptions, retaining only those that provided comprehensive and contextually rich demonstrations.

To refine the generation of CoT reasoning and minimize hallucinations, we implemented an automated filtering mechanism to discard unreliable responses. Our evaluation encompassed three MLLMs: LLaVA-NeXT 34B [82], InternLM-XComposer2 [159], and LLaVA-LLaMA3 [160], a LLaVA-1.5-7B [161] model derived from fine-tuning LLaMA-8B Instruct [162]. Table 3.1 illustrates the success rates of these models in generating descriptive reasoning chains. Notably, LLaVA-NeXT outperformed the others, demonstrating superior performance in producing coherent reasoning chains. This highlights the advantage of

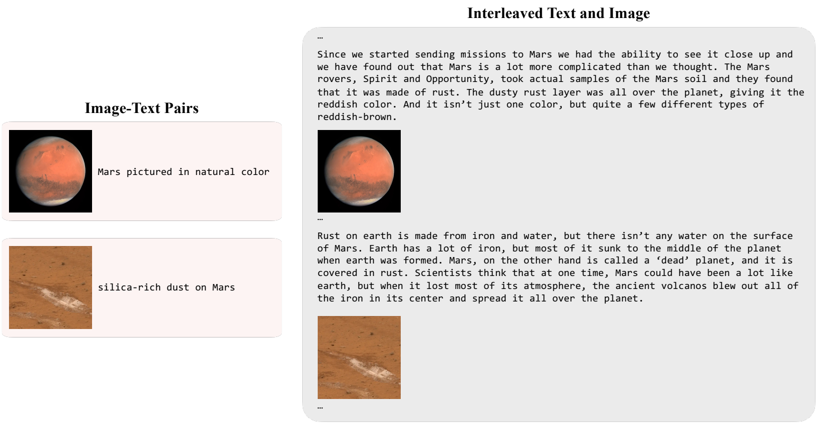


Figure 3.3 Sample data illustrating the differences between image-text pairs and interleaved text-image data used in training MLLMs.

employing larger models for enhancing the quality of generated reasoning. The prompt fed to MLLMs 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: ...*"

3.2.5. Pretrained Models

In this study, we evaluate MLLMs trained on both captioning datasets and interleaved image-text datasets. Captioning datasets consist of individual images paired with descriptive text, which allows models to develop strong zero-shot capabilities, meaning they can generate responses based on a single image-text pair without prior examples. This approach is advantageous for applications requiring immediate and contextually relevant descriptions, as demonstrated in Figure 3.3. In contrast, interleaved image-text datasets contain sequences of images and corresponding text, enabling models to understand and generate coherent responses across multiple image-text pairs. This few-shot ability is beneficial for tasks demanding deeper contextual understanding and continuity. Figure 3.4 illustrates the training process for models using these dataset types, highlighting the difference between processing

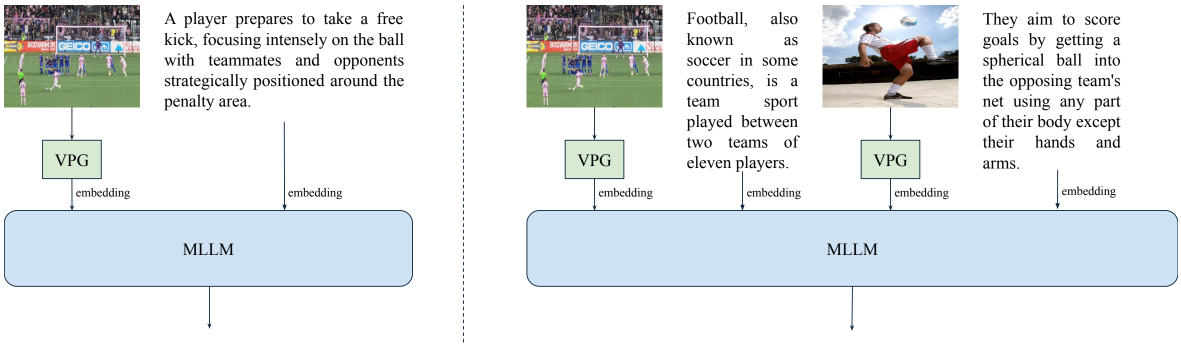


Figure 3.4 An overview illustrating the training procedure of MLLMs on captioning datasets (left) and interleaved text-image datasets (right). On the left, MLLMs are trained on single image-text pairs, whereas on the right, MLLMs are trained on multiple image-text pairs. Visual Prompt Generators (VPG) are used to create embeddings of images, which are then fed through the MLLM along with text embeddings.

single image-text pairs and multiple image-text pairs. Through our evaluation, we aim to elucidate the strengths and applications of MLLMs trained on these distinct dataset types.

MLLMs pretrained on Captioning Datasets. Recent advancements in NLP have sparked significant interest in models adept at handling individual image-text pairs [122, 161, 163–166]. These models demonstrate exceptional capabilities in comprehending and generating textual descriptions for given images, thereby enhancing tasks like image captioning, visual question answering, and image retrieval. By leveraging sophisticated architectures and multimodal learning techniques, these models effectively integrate visual and textual information to derive semantic meaning and context, demonstrating substantial potential for a wide array of applications in image understanding, multimedia analysis, and human-computer interaction.

LLaVA, known as the Large Language and Vision Assistant, represents a significant advancement in multimodal research, encompassing models such as LLaVA 1.5 [83] and LLaVA-NeXT [82]. These models surpass tasks involving natural instruction following and visual reasoning, with LLaVA 1.5 establishing new benchmarks across 12 datasets. The latest iteration, LLaVA-NeXT, enhances capabilities in reasoning, OCR, and world knowledge, surpassing benchmarks set by Gemini Pro 1.0 [167] in specific evaluations. LLaVA-NeXT achieves these milestones while maintaining a streamlined design and high data efficiency, requiring fewer than 1 million samples for visual instruction fine-tuning. Notably, it leads among open-source multimodal models with significantly reduced training costs. In our evaluations, we opted for the LLaVA-NeXT 34B variant.

PaliGemma, developed by Google, is another formidable MLLM equipped with a Transformer decoder and a Vision Transformer image encoder, boasting 3 billion parameters. Derived from Gemma-2B [168] and SigLIP-So400m/14 [169], it adheres to the PaLI-3 training protocol [170]. PaliGemma accepts inputs of images and text strings, generating outputs like image captions, answers to questions, object bounding box coordinates, or segmentation codewords. Pretrained on diverse datasets including WebLI [170], CC3M-35L [171], VQ²A-CC3M-35L/VQG-CC3M-35L (a subset of VQ2A-CC3M [172]), OpenImages [173], and WIT [174], PaliGemma surpasses visual semantic understanding and multilingual tasks. Rigorous data filters ensure training data safety, cleanliness, and privacy compliance by removing inappropriate or sensitive content using advanced filtering techniques.

Intern-VL-Chat-V1-5 [175] stands as an advanced vision-language model with 26 billion parameters, aimed at bridging the gap between open-source and commercial models. It utilizes the InternViT-6B [84] vision foundation model and InternLM2-20B [159] language model, enhanced by continuous learning with high-quality image-text data, a dynamic high-resolution strategy for detailed image analysis, and a diverse multilingual dataset pipeline. In evaluations across 18 multimodal benchmarks, InternVL 1.5 outperforms competitors in 8 benchmarks, including OCR tasks, demonstrating its capability to narrow the gap between open-source and commercial multimodal models.

InternLM-XComposer2 [85], with 7 billion parameters, surpasses generating and comprehending free-form text-image content. By integrating text and graphics from diverse inputs such as outlines and reference images, it facilitates flexible content production beyond traditional comprehension. Leveraging Partial LoRA (PLoRA) to strategically enhance image token parameters while preserving language understanding, InternLM-XComposer2 exhibits superior performance in various evaluations compared to existing multimodal models like GPT-4V [176] and Gemini Pro [167].

MLLMs pretrained on Interleaved Image-Text Datasets. The development of models capable of handling multiple image-text pairs has become a focal point in research [38, 52, 73, 74, 77, 78, 80, 81]. These frameworks demonstrate proficiency in analyzing and comprehending multiple instances of image-text pairs concurrently, enabling deeper understanding and interpretation of multimodal data. By leveraging advanced multimodal fusion techniques and attention mechanisms, these models seamlessly integrate information from diverse sources to extract nuanced semantics and context across multiple modalities. This expanded capability enhances applications ranging from image album summarization to cross-modal retrieval and interactive storytelling, enriching the depth and complexity of information processing and comprehension.

OpenFlamingo [78] introduces a novel approach to vision and language modeling, empowering autoregressive models to process sequences of mixed images and text for enhanced versatility, including few-shot learning and multi-round chatbot interactions. Unlike proprietary models such as Flamingo [38], CM3 [177], Kosmos-1 [87], PALME [178], and multimodal GPT-4 [176], OpenFlamingo provides an accessible open-source alternative, fostering research accessibility. By leveraging pretrained language models with cross-modal attention to vision encoders, OpenFlamingo achieves competitive performance across various models ranging from 3 billion to 9 billion parameters. Evaluations across seven datasets demonstrate that OpenFlamingo models achieve 85% to 89% of the performance of their proprietary counterparts, highlighting their effectiveness and adaptability.

Idefics [52, 80] comprises two variants: 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 such as CLIP-ViT-H-14 [49] and LLaMA-65B [129], it is available in two sizes (80 billion and 9 billion parameters) and outperforms benchmarks in tasks such as image captioning and visual question answering. Idefics2, with 8 billion parameters, enhances OCR capabilities, document understanding, and visual reasoning, handling images in native resolutions with the NaViT strategy [179] and incorporating new training data for improved OCR and document comprehension.

xGen-MM [79], developed by Salesforce AI Research, builds on the successful BLIP series aligned with Salesforce's XGen initiative for large foundational models. Trained on diverse datasets including high-quality image captions, xGen-MM models 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 [50] to academic VQA tasks, ensuring versatility and robustness. In our experiments, we utilized the xGen-MM mini base variant with a model size of 4.6 billion parameters.

Qwen-VL [81] expands on the Qwen language model, overcoming limitations of traditional LLMs by integrating visual understanding capabilities. These models, including Qwen-VL-Chat with 9.6 billion parameters, facilitate user interaction through both text and images. They surpass tasks such as image captioning and question answering, demonstrating superior performance and supporting multiple languages. Additionally, Qwen-VL models handle multiple images effectively, achieving robust performance across diverse benchmarks, particularly in fine-grained visual understanding.

MMICL [54], Multi-Modal In-Context Learning, aims to overcome limitations observed in current MLLMs when handling intricate prompts involving multiple images and text. MMICL, equipped with a model size of 12.1 billion parameters, introduces novel methodologies for processing multi-modal inputs, including a unique context scheme to enhance contextual learning. Leveraging the Multi-modal In-Context Learning (MIC) dataset, MMICL enhances its capability to interpret complex multi-modal prompts, addressing challenges such as understanding text-to-image correlations and relationships across multiple images. Moreover, MMICL mitigates language biases that can lead to erroneous interpretations in extensive textual contexts.

For our experimental setup, we utilized implementations from the HuggingFace repository. We employed half-precision for running Idefics1, MMICL, and full precision for OpenFlamingo variants and xGen-MM. InterVL-Chat was tested with 8-bit quantization, while other models underwent testing with 4-bit quantization. Our experiments were conducted on GPUs including Tesla T4, Quadro P4000, V100, or A40.

3.3. Experimental Analysis

We demonstrate the zero-shot and few-shot capabilities of MLLMs trained on interleaved image-text datasets or captioning datasets in Table 3.2. Additionally, we include qualitative examples in the Appendix.

Observation 1: Instruction tuning and In-Context Learning improve models' adherence to user instructions.

Given our questions, we expect the MLLMs to provide Yes/No responses. However, in the zero-shot setting, some models produced outputs with irrelevant information, resulting in notably low scores. Instruction tuning or providing demonstration examples through ICL often helps models follow the expected answer templates. For example, OpenFlamingo-3B and xGen-MM exhibit this improvement.

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. This behavior is consistently observed across evaluated MLLMs, regardless of model size. Notably, examples similar to the 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: More similar demonstration examples improve performance compared to more random examples.

Table 3.2 Accuracy performance of the evaluated MLLMs, varying by model size and pretraining strategies, assessed with 0-8 shots across three settings: Random (**R**), Similar (**S**), and Similar with Chain of Thought (**S+C**). In the **R** setting, few-shot demonstrations are randomly selected. In the **S** setting, few-shot examples are chosen based on visual and textual similarity. In the **S+C** setting, examples are selected based on visual and textual similarity and include a CoT description. Models with the suffix 'I' denote 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

[155] studied atomic foils with the CREPE benchmark [70], which is similar to the VALSE benchmark in measuring model performance changes when atomic foils completely alter sentence meanings. They found that increasing the number of random demonstration

examples provides almost no gain in this setup. Our results support this finding, showing that increasing random example count can sometimes deteriorate performance. However, using a higher number of similar examples enhances MLLM performance, as more similar examples help establish a link between the context and query.

> *Observation 4:* Chain-of-Thought reasoning impacts instruction-following abilities in OpenFlamingo variants and MMICL but enhances performance in other models.

CoT descriptions in demonstration examples help models reason about a given image-text pair, 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 can sometimes cause OpenFlamingo variants and MMICL to ignore the expected answer templates, generating reasoning chains without providing direct answers, leading to poor performance. In contrast, for higher-capacity models, CoT generally enhances performance.

> *Observation 5:* With ICL and CoT, lower-capacity models trained on interleaved image-text datasets perform comparably to or better than larger models trained on captioning datasets.

Except for Idefics-2, 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 better performance than larger models trained on captioning datasets. For example, Idefics-9B achieved 77.2% accuracy with 4-shot ICL and CoT, while Intern-VL-Chat-V1-5-26B achieved 76.7% overall accuracy.

> *Observation 6:* Models prefer textually similar demonstrations over visually similar ones, slightly improving performance.

Table 3.3 shows the performance changes of models pretrained on interleaved image-text datasets across different K values within the ICL setting. Increasing K provides a larger pool of visually similar examples. When N examples are selected from this pool based on textual similarity, the final demonstration examples exhibit higher textual similarity to

Table 3.3 Accuracy performance of MLLMs pretrained on interleaved image-text data, varying by model size, in the few-shot ICL setting. Demonstrations are selected based on their similarity to the query, with N textually similar examples chosen from a pool of K visually similar examples. The table presents performance across different K values, specifically 20, 50, and 100. Models with the suffix 'I' denote 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

the query image-text pair, albeit potentially lower visual similarity. The results indicate a marginal performance improvement with higher K, suggesting that models prefer textually similar examples.

Herein, we delve into a detailed examination of each task based on our experimental insights.

Existence task is fundamental in VALSE, assessing a model's ability to determine if an object is present in an image. All models showed higher accuracy on this task compared to others, indicating that MLLMs can effectively identify and represent objects. However, when CoT descriptions were introduced, all models except Idefics-9B experienced a decline in performance. This decline is attributed to hallucinated and irrelevant reasoning chains, leading to incorrect answers. Additionally, as shown in Table 3.3, an increase in textually similar examples significantly improves model performance more than in other tasks.

Plurality task challenges models to recognize objects and determine their plural forms. Results indicate that demonstration examples do not improve the models' understanding of pluralism, even though they correctly recognize the objects. CoT reasoning is useful here, as it provides reasoning chains that describe plural forms, helping models develop an understanding of the task.

Counting task evaluates a model's ability to identify both the objects and the number of their appearances in a scene. Models trained on captioning datasets outperform those trained on interleaved image-text data. However, combining few-shot ICL and CoT reasoning enhances performance, bringing these models closer to those trained on captioning data. Qualitative examples show that models are guided to count each occurrence, allowing for accurate comparison between actual and stated occurrences.

Spatial Relations task assesses models' ability to recognize interactions between objects. Zero-shot performance shows that all models struggle with this task, which requires understanding object interactions and relationships. Providing demonstration examples through ICL helps achieve a certain performance level, but increasing the number of examples does not lead to further improvement. Performance gains saturate with more examples. However, using few-shot ICL combined with CoT reasoning can achieve up to a 30% performance increase (Idefics-9B).

Action task evaluates how well models detect actions and actors in a scene. This task is challenging as it requires identifying dynamic interactions and context-specific activities within an image. Models trained on captioning data performed better than those trained on interleaved image-text datasets. Few-shot ICL successfully elevated performance to up to

73%. However, except for the Idefics model family, none of the models benefited from CoT descriptions, and increasing the number of demonstration examples did not always improve performance.

Coreference task tests a model's ability to resolve pronoun references within a visual context, ensuring that MLLMs can accurately associate pronouns with their corresponding entities in images. This task is challenging as it requires maintaining contextual relationships between pronouns and their antecedents. Results show that models trained on captioning datasets outperformed those trained on interleaved image-text datasets. The Idefics model family, in particular, benefited substantially from CoT descriptions, which provide explicit reasoning pathways to correctly link pronouns. However, increasing the number of demonstration examples did not significantly improve overall performance with CoT descriptions.

Foil-It! task evaluates a model's understanding of objects by replacing the target object with an irrelevant one to create a foil. This task requires models to accurately recognize objects and detect subtle contextual inconsistencies. Zero-shot performance of models trained on captioning data surpassed those trained on interleaved image-text datasets. Even with few-shot ICL and CoT techniques, these models could not be outperformed.

4. PROBING SEMANTIC UNDERSTANDING CAPABILITIES OF VIDEO-LANGUAGE MODELS

4.1. Introduction

In this study, we introduce an approach to assessing pretrained VidLMs by proposing a zero-shot evaluation test. VidLMs represent a significant advancement in the field, gaining attention from researchers for their ability to integrate visual and temporal data from videos, facilitating a deeper understanding of dynamic phenomena [88, 91, 92, 96, 102, 108–110, 180–186].

Building on prior research, we present a zero-shot evaluation test specifically designed to gauge the language understanding capabilities of VidLMs. This test aims to address the limitations of existing evaluation methodologies by focusing on tasks that require a robust comprehension of both visual and linguistic elements within videos.

Video-Language Models offer several advantages over ILMs. Firstly, they provide a more comprehensive understanding of context by analyzing both visual and textual information simultaneously, leading to a richer representation of concepts and allowing for more detailed interpretations. Secondly, VidLMs can capture temporal dependencies and dynamic interactions within a scene, enabling them to infer complex relationships over time. This temporal understanding enhances their ability to generate coherent and contextually relevant responses. Additionally, VidLMs can facilitate tasks such as video summarization, captioning, and content recommendation more effectively by leveraging both visual and textual cues. Overall, integrating video content into language models enhances their capabilities and enables more sophisticated applications across various domains.

Proficiency Test: A shirtless man opens the window / door hurriedly.
Main Test: A shirtless man opens / smashes the window hurriedly.

Proficiency Test: The man in a navy blue coat drags the man in the green coat off the ledge / ground.
Main Test: The man in a navy blue coat drags / tosses the man in the green coat off the ledge.

Proficiency Test: A girl, wearing a yellow top, forcefully pushes her body / away.
Main Test: A girl, wearing a yellow top, forcefully pushes / covers her body.

Proficiency Test: A man, dressed in a black outfit, aims his gun at a man in a brown shirt / SUV.
Main Test: A man, dressed in a black outfit, aims / discharges his gun at a man in a brown shirt.

Proficiency Test: A man with a face mask breathes oxygen with difficulty / goggles.
Main Test: A man with a face mask breathes / measures oxygen with difficulty.

Figure 4.1 Examples of cases from the **Situation Awareness (Action Replacement)** test.

Proficiency Test: The man in a tan coat clasps the woman with blonde hair / locks outside.
Main Test: The man in a tan coat / woman with blonde hair clasps the woman with blonde hair / man in a tan coat outside.

Proficiency Test: The woman in red looks upward at the man in a hat / wheelchair.
Main Test: The woman in red / man in a hat looks upward at the man in a hat / woman in red.

Proficiency Test: A gentleman dressed in a tan suit pulls a man in a green shirt by the arms / collar.
Main Test: A gentleman dressed in a tan suit / man in a green shirt pulls a man in a green shirt / gentleman dressed in a tan suit by the arms.

Proficiency Test: The girl with the ponytail suddenly pushes the boy with glasses / her.
Main Test: The girl with the ponytail / The boy with glasses suddenly pushes the boy with glasses / the girl with the ponytail.

Proficiency Test: A young man sitting in a car notices a person wearing a dinosaur costume near a van / tree and a parked car.
Main Test: A young man sitting in a car / person wearing a dinosaur costume notices a person wearing a dinosaur costume / young man sitting in a car near a van and a parked car.

Figure 4.2 Examples of cases from the **Situation Awareness (Actor Swapping)** test.

To address this gap, our test introduces a task-independent evaluation framework focusing on key visuo-linguistic phenomena, namely *Situation Awareness* (*SA*). This SA test is an integral part of the ViLMA benchmark [187], which aims to comprehensively evaluate the capabilities of pretrained VidLMs. Within ViLMA, *SA* plays a vital role in assessing the models' understanding of dynamic scenarios depicted in videos. By challenging VidLMs' ability to perform *Action Replacement* and *Actor Swapping*, the *SA* test highlights the importance of temporal comprehension and context awareness in these models' performance evaluation.

Action Replacement mirrors the concept of foils in the VALSE benchmark [1], wherein we manipulate the actual action portrayed in the video while maintaining the contextual integrity of the scene (see Figure 4.1). This allows us to gauge the VidLMs ' ability to accurately discern subtle alterations in depicted actions.

In contrast, *Actor Swapping* introduces a new dimension by focusing on scenes featuring multiple actors. Here, we systematically swap actors within the captions, simulating scenarios where different individuals engage in similar actions (see Figure 4.2). This challenges VidLMs to not only recognize actions but also identify actors within the scene accurately. Similar to our predecessors, we undertake both manual and automatic validation processes to ensure the fidelity and integrity of our evaluation test. Foils are crafted by modifying small segments of the captions, preserving the essence of the original context while introducing controlled variations [154].

Our evaluation framework incorporates proficiency tests designed to assess fundamental capabilities such as object detection within scenes, serving as precursors to the main evaluation tasks. These preliminary assessments offer insights into the VidLMs' foundational understanding before delving into more complex challenges posed by *SA*.

Through rigorous testing on these components, we aim to comprehensively evaluate VidLMs' aptitude in understanding detailed visuo-linguistic phenomena, thereby advancing the state-of-the-art in video-language comprehension research.

4.2. Action&Actor Recognition

In this section, we delve into details about the methodologies in creating *SA* test. Section 4.2.1. provides an in-depth examination of the data sources utilized in our evaluation framework, ensuring diversity and representability. In contrast, Section 4.2.2. describes our foiling method, detailing how controlled alterations are introduced to maintain contextual integrity. Section 4.2.3. states proficiency tests designed to assess fundamental capabilities before proceeding to main evaluation tasks. Section 4.2.4. describes the Evaluation Metric utilized in our model assessment, providing insights into the chosen metrics and their relevance to evaluating VidLMs' language understanding abilities. Following this, Section 4.2.5. details the selection of models tested against benchmarks, shedding light on the models chosen for comparison and their respective pretrained objectives. Finally, in Section 4.2.6., we present the implementation specifics of the models. By incorporating these sections into our comprehensive methodology, we ensure a thorough and systematic approach to evaluating VidLMs' language understanding capabilities.

4.2.1. Data Sources

For our SA subtests, we rely on the VidSitu dataset [15], a comprehensive repository of short video clips extracted from movies. This dataset is annotated, providing detailed information about various elements within each clip, including verbs, semantic roles, entity references, and event relationships. To generate captions for these SA subtests, we enlist the assistance of ChatGPT [188], an advanced language model developed by OpenAI. Our approach involves presenting ChatGPT with straightforward sentences obtained from the VidSitu dataset and instructing it to enhance their linguistic sophistication. Specifically, we task ChatGPT with elevating the vocabulary and sentence structures to imbue the text with a more refined and elegant tone while preserving the original meaning. The prompt fed to ChatGPT is given below:

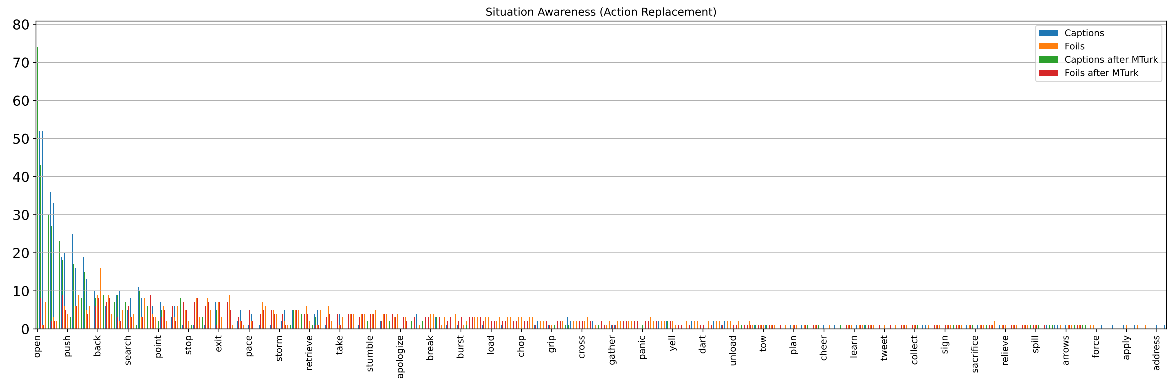


Figure 4.3 Caption and foil distribution of Situation Awareness main test, before and after Amazon Mechanical Turk validation process.

> *"I want you to act as an English spelling corrector and improver. I will speak to you in English and you will answer in the corrected and improved version of my text, in English. I want you to replace my simplified A0-level words and sentences with more beautiful and elegant, upper-level English words and sentences. Keep the meaning same, but make them more literary. I want you to only reply with the corrections, the improvements, and nothing else, do not write explanations. Your responses should be enumerated. Each sentence is separated by a dot (.). My sentences are: ..."*

Following this refinement process, we evaluate the readability of the improved captions using two widely recognized metrics: the Flesch-Kincaid [189] and Flesch Reading Ease [190] methods. The resulting scores, 4.54 and 83.27 respectively, indicate that the text is generally at a grade 5-6 reading level, reflecting a balance between complexity and accessibility.

4.2.2. Foiling Method

We generate alternative options for each caption by selecting the top 32 most likely words from *RoBERTa-base*[2] outputs. These alternatives undergo a two-phase evaluation process to determine their suitability. Initially, an *ALBERT*[3] model assesses them using Natural Language Inference (NLI) to gauge their alignment with the video content. Any options

[2] `https://huggingface.co/roberta-base`
[3] `https://huggingface.co/ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli`

identified as entailment are discarded, while those labeled as neutral or contradiction are retained. Subsequently, we evaluate their grammatical correctness using the GRUEN score, discarding options with less than an 80% score. Alternatives must pass both NLI and GRUEN evaluations, ensuring they are contextually appropriate and linguistically coherent. At the final stage, we submit the remaining foils to Amazon Mechanical Turk for manual validation. For the breakdown of foil distribution, refer to Figure 4.3.

4.2.3. Proficiency Tests

In our *SA* proficiency test, our primary focus lies in object identification, which is crucial for evaluating the model's ability to recognize actions and actors accurately. This emphasis on object identification is critical because it forms the foundation for comprehending actions within specific scenarios and pinpointing the individuals or entities involved, both of which are fundamental components of situational understanding. Our methodology revolves around foiling objects based on the transitivity of verbs: when a verb necessitates an object, we replace it with a counterfactual generated by RoBERTa [191], allowing us to gauge the model's grasp of the object's role in actions. Conversely, when a verb cannot directly take an object, we mask the subject, ensuring a comprehensive evaluation of the model's ability to identify actors. Object identification, within this framework, facilitates a holistic interpretation of scenes, aiding the model in grasping the broader context and interrelationships among elements in dynamic scenarios, perfectly aligning with the goals of the *SA* assessment.

4.2.4. Evaluation Metric

While pairwise ranking accuracy serves as our primary evaluation metric, it's important to consider its comparison with other metrics such as BLEU [111], ROUGE [113], or METEOR [112]. In some models, perplexity has been utilized to gauge the coherence and fidelity of generated text, whereas other models employ image-text similarity as a measure of semantic alignment between visual and textual inputs. We chose pairwise ranking accuracy for its suitability in comparing models pretrained with different objectives. This approach enables the comparison of all 16 VLMs, including those pretrained with Visual-to-Text Captioning and Natural Language Generation objectives, with acc_r scores presented for main tests (T) and corresponding proficiency tasks (P). A stricter combined score (P+T) is introduced, ensuring a model's success on the main test correlates with proficiency test performance.

4.2.5. Pretrained Models

Unimodal Models. We explore the performance of various decoder-only or encoder-decoder language models, namely GPT-2 [8], OPT [9], T5 [11] and BART [10], on the benchmark. Following a methodology akin to VALSE [1], we assess perplexity values for both captions and foils, opting for the text input with the lower perplexity score. Additionally, in our experimental setup involving GPT-2[4] with 124M parameters and OPT-6.7B[5], we utilize GPT-2 with 124M parameters and OPT-6.7B, respectively, aiming to examine their performance under specific parameter configurations.

Image-Language Models. In our investigation, we explore the capabilities of two leading Image-Language Models: CLIP [28] and BLIP-2 [163]. CLIP adopts a dual-encoder architecture, incorporating a contrastive loss function to optimize the training of image-caption pairs. Conversely, BLIP-2 represents a further development of BLIP [121], capitalizing on frozen pretrained image encoders and large language models to advance vision-language fusion. For these experiments involving CLIP and BLIP-2, we employed their largest versions, CLIP[6] and BLIP-2[7], respectively, along with OPT-6.7B, aiming to evaluate their performance under specific parameter configurations and compare them against other models.

Video-Language Models. Here we provide comprehensive information about the pretrained video-language models listed in Table 4.1.

[4] `https://huggingface.co/gpt2`
[5] `https://huggingface.co/facebook/opt-6.7b`
[6] `https://huggingface.co/openai/clip-vit-large-patch14`
[7] `https://huggingface.co/Salesforce/blip2-opt-6.7b`

ClipBERT [88] leverages the powerful BERT for text encoding and ResNet-50 [192] for video encoding. An intriguing departure from conventional approaches, ClipBERT's pretraining solely relies on images [47, 193]. However, it diverges in its inability to grasp temporal sequences; its video-text similarity computation is based on the average frame-text similarity score.

UniVL [109] features a sophisticated two-stream encoder-decoder architecture. A pretrained BERT handles textual inputs, while visual cues are processed through S3D and a transformer encoder. Modalities are fused using a cross-encoder. UniVL's unique pretraining on HowTo100M involves a generative task, distinguishing it from many other VidLMs.

VideoCLIP [91] stands out by using BERT for text encoding and S3D [194] for video encoding, both pretrained on HowTo100M. Like ClipBERT, it uses mean pooling to merge modalities, enabling effective cross-modal understanding.

FiT [102] employs BERT for text encoding and TimeSFormer [106] for video representation. Unlike many models, FiT is pretrained on both images (CC3M) and videos (W2), creating a cohesive video-text space through contrastive learning. The authors also developed the W2 dataset, enhancing resources for video understanding research.

CLIP4Clip [96] leverages the extensive knowledge within the CLIP [28] model for video-language retrieval. Through empirical studies, the authors explore whether image features alone suffice for robust video-text retrieval, the impact of post-pretraining with CLIP on extensive video-text datasets, methods for modeling temporal dependencies among video frames, and the role of hyperparameters in video-text retrieval systems.

VIOLET [108] uses a dual-stream encoder-only architecture, with a BERT-based text module and a Video Swin Transformer [195] for video frames. VIOLET models spatial and temporal dimensions through positional embeddings. Its training spans diverse data sources, including videos and images, with each module fine-tuned for optimal performance.

X-CLIP [180] tackles video-text retrieval with a multi-grained contrastive mechanism, encoding textual and visual inputs into coarse-grained and fine-grained representations. Its

Attention Over Similarity Matrix module allows selective focus on critical frames and words while minimizing irrelevant elements during retrieval.

MCQ [181] introduces a pretraining paradigm centered on Multiple Choice Questions (MCQ) for VLMs. Using a dual-encoder mechanism and BridgeFormer, it links local features from VideoFormer [196] and TextFormer [197]. By optimizing multiple-choice question answering through contrastive learning, MCQ enhances semantic associations between video-text representations and maintains efficiency in retrieval tasks, with the flexibility to remove the BridgeFormer module for downstream applications.

Singularity [110] demonstrates the effectiveness of single-frame training for VidLM tasks, such as video question answering and text-to-video retrieval. It integrates a vision encoder [196], a language encoder [6], and a multi-modal encoder with a cross-attention fusion mechanism. Additionally, it introduces a benchmark to address the overemphasis on temporal learning capabilities, highlighting a significant static appearance bias in current video-and-language datasets.

UniPerceiver [182] focuses on pretraining a unified framework for general perception tasks, emphasizing zero-shot and few-shot learning. It combines transformer capabilities with neural perceptrons for learning across multiple modalities, including texts, audio, and images. Through a common encoder-decoder structure, UniPerceiver leverages correlations between modalities during pretraining, enabling comprehensive perception across diverse data domains.

Merlot Reserve [92] advances video comprehension by integrating audio, subtitles, and video frames. It uses a training method that substitutes text and audio bits with a MASK token, selecting the appropriate masked-out segment. This strategy surpasses in various challenges, including Visual Commonsense Reasoning [198], TVQA [199], and Kinetics-600 [200].

VindLU [183] enhances VidLMs pretraining with a methodical approach, using image [201] and text encoders [6] trained on video and caption pairs through a visual-text contrastive objective. The framework incrementally incorporates components, assessing their significance, resulting in six essential steps that contribute to effective VidLMs pretraining.

InternVideo [184] addresses limitations in existing vision models, focusing on video-level understanding tasks. By combining generative and discriminative self-supervised learning, InternVideo enhances performance across diverse video applications, achieving state-of-the-art results on various datasets [23, 202].

mPLUG-2 [185] unifies multiple modalities, including language, image, and video, similar to UniPerceiver [182]. Through cross-modal transformer layers and a fusion module, mPLUG-2 generates visually-aware textual features and textually-aware visual features, adapted to generative tasks through a text decoder. During pretraining, it processes millions of image-text and video-text pairs, enriching its understanding across diverse modalities.

Otter [74] specializes in multi-modal in-context instruction tuning, using the MIMIC-IT [203] dataset to enhance its ability to process and respond to instructions across video and multiple image inputs. After fine-tuning on MIMIC-IT with multimodal instruction-response pairs, Otter shows improved instruction-following abilities compared to its precursor, OpenFlamingo [78].

Video-LLaMA [186] is a conversational VidLM designed for following instructions. Built on BLIP2 [163], it incorporates separate query-formers for video and audio, combined with a language model through a frozen prefix. Video-LLaMA's pretraining uses a subset of CC3M and the WebVid2M dataset, followed by fine-tuning on the Video-Chat instructions dataset [204], achieving state-of-the-art performance across various video tasks.

4.2.6. Implementation Details

We aim to use each model in its original form with official implementations, following a zero-shot approach. Using Huggingface [205] implementations, we maintain consistency across GPT-2, OPT, CLIP, BLIP2, and X-CLIP. Most VidLMs use a specific number of frames (K) for video input, with some variations: X-CLIP, InternVideo, and Video-LLaMA use $K = 8$; ClipBERT uses $K = 16$; and others use $K = 4$. Merlot Reserve partitions videos into segments with a 5-second interval, capturing the middle frame. For video-caption match scores in ILMs, mean pooling is done over scores obtained using $K = 8$ frames. Experiments are run on single Tesla T4, Quadro P4000, or V100 GPUs using half precision.

4.3. Experiments

In the following section, we delve into the performance analysis of three categories of language models: Unimodal Models, Image-Language Models, and Video-Language Models. We discuss their proficiency and main test scores, highlighting notable trends and disparities in their performance.

Unimodal models exhibit an intriguing contrast between proficiency and main tests, showing significant improvements in handling complex tasks. Moreover, we present our results in 4.1, providing a comprehensive overview of the performance metrics across different models. Additionally, to offer a visual representation of our findings, we include example figures in Figure 4.1 (Action Replacement) and Figure 4.2 (Actor Swapping), illustrating specific instances of task challenges and model responses.

ILMs demonstrate superiority over both unimodal models and VidLMs, with BLIP2 particularly standing out. VidLMs face challenges in detecting actors and actions within scenes, although Video-LLaMA presents a promising exception. We explore these findings in detail, shedding light on the capabilities and limitations of each model category in visuo-linguistic comprehension tasks.

Unimodal Models. Unimodal models exhibit a notable disparity in performance between proficiency and main tests. In proficiency tests, their scores are close to random, indicating poor performance in tasks emphasizing object identification. This suggests a struggle in grasping foundational elements crucial for situational awareness. However, in main tests, their performance improves significantly, exhibiting higher scores. This suggests an ability to handle the complexities of tasks like action replacement and actor swapping more effectively,

Table 4.1 The results of the Situation Awareness subtests are evaluated using the pairwise accuracy (acc_r) metric, with P, T, and P+T denoting the scores achieved on proficiency tests, main tests only, and combined tests, respectively.

Model	**Action Replacement**			**Actor Swapping**			**All**		
	P	T	P+T	P	T	P+T	P	T	P+T
Random	50.00	34.42	17.21	50.00	50.00	25.00	50.00	37.96	18.98
GPT-2$^{\Psi}$	43.03	67.47	31.68	49.76	63.77	31.88	44.57	66.63	31.72
OPT$^{\Psi}$	50.14	<u>71.88</u>	38.49	57.00	<u>69.57</u>	39.61	51.70	<u>71.35</u>	38.75
CLIP$^{\Psi}$	70.74	45.03	33.66	71.98	47.34	33.82	71.02	45.55	33.70
BLIP2$^{\Psi}$	72.30	**78.12**	**57.24**	<u>77.29</u>	66.18	<u>50.72</u>	73.44	**75.41**	**55.76**
ClipBERT$^{\Phi}$	53.41	55.11	29.69	56.52	63.29	39.61	54.12	56.97	31.94
UniVL$^{\Phi}$	53.98	44.46	23.86	49.28	54.11	25.12	52.91	46.65	24.15
VideoCLIP$^{\Phi}$	62.78	37.36	22.59	57.97	50.72	32.85	61.69	40.40	24.92
FiT$^{\Phi}$	68.47	38.64	27.56	74.40	44.93	34.78	69.81	40.07	29.20
CLIP4Clip$^{\Phi}$	<u>73.15</u>	46.59	35.51	76.33	57.49	44.93	<u>73.87</u>	49.07	37.65
VIOLET$^{\Phi}$	69.32	41.19	29.69	73.43	55.56	42.03	70.25	44.46	32.49
X-CLIP$^{\Phi}$	64.91	43.32	30.68	58.94	50.24	32.37	63.56	44.90	31.06
MCQ$^{\Phi}$	65.20	33.10	22.44	73.43	50.72	39.61	67.07	37.10	26.34
Singularity$^{\Phi}$	67.05	38.78	27.70	74.88	48.31	38.65	68.83	40.94	30.19
UniPerceiver$^{\Phi}$	52.13	29.12	14.35	49.28	**86.47**	44.44	51.48	42.15	21.19
Merlot Reserve$^{\Phi}$	68.89	30.97	21.16	76.33	51.69	39.61	70.58	35.68	25.36
VindLU$^{\Phi}$	69.46	39.63	29.40	74.40	48.31	37.68	70.58	41.60	31.28
InternVideo$^{\Phi}$	70.88	39.20	28.12	73.91	47.34	34.30	71.57	41.05	29.53
mPLUG-2$^{\Psi}$	47.60	34.70	18.32	56.50	46.40	32.40	49.60	37.40	21.50
Otter$^{\Psi}$	58.10	39.63	23.86	59.42	62.32	34.78	58.76	50.98	29.32
Video-LLaMA$^{\Psi}$	**77.56**	67.61	<u>53.55</u>	**80.19**	64.73	**55.56**	**78.15**	66.96	<u>54.01</u>

Φ: Image-Text similarity used
Ψ: Perplexity used

highlighting their capacity to integrate visual and linguistic information in understanding dynamic scenarios.

Image-Language Models. ILMs emerge as superior performers compared to both unimodal models and VidLMs. Particularly, BLIP2 [163] stands out with exceptional performance across both proficiency and main tests. The proficiency tests reveal that ILMs surpass in foundational tasks like object identification, surpassing the performance of unimodal models. Moreover, their robust performance extends to main tests, where they continue to outperform other models. Notably, while most VidLMs exhibit a decrease in scores from proficiency to main tests, BLIP2 demonstrates an opposite trend with higher scores in main tests. This suggests that BLIP2 maintains its effectiveness in handling the complexities of real-world scenarios, presenting its capability to integrate visual and linguistic information efficiently.

Video-Language Models. VidLMs generally exhibit poor performance compared to random chance, suggesting significant challenges in detecting actors and actions within scenes. Despite the advanced architecture and training methodologies of VidLMs, their proficiency in foundational tasks like object identification remains lacking. However, Video-LLaMA, leveraging a foundation from BLIP, emerges as an exception with the highest score among VidLMs. This success indicates the effectiveness of BLIP2's performance and highlights the potential benefits of leveraging strong foundations in model development. Nevertheless, the overall performance of VidLMs indicates considerable room for improvement in comprehending and interpreting dynamic scenarios depicted in videos. Further advancements in model architectures and training strategies are necessary to enhance their capabilities in visuo-linguistic understanding tasks.

5. CONCLUSION

This thesis has delved into the comparative analysis of Video-Language Models and Multimodal Large Language Models, focusing on their ability to bridge the semantic gap between visual content and natural language. Through empirical evaluation, this research has highlighted the strengths and limitations of these models in comprehending and articulating visual inputs. This thesis has contributed to a deeper understanding of the models' performance in real-world applications by examining the advancements in deep learning and multimodal architectures.

The primary contributions of this thesis are multifaceted. First, we investigate the zero-shot and few-shot capabilities of MLLMs trained on interleaved image-text datasets versus captioning datasets. We find that instruction tuning and In-Context Learning significantly improve models' ability to follow user instructions, particularly enhancing performance when demonstration examples are similar to the query image-text pairs. Using more similar demonstration examples consistently enhances MLLM performance compared to random examples, fostering better contextual understanding and improving task-specific outcomes. Additionally, while Chain-of-Thought reasoning aids in complex reasoning tasks like counting and coreference, it sometimes detracts from models' adherence to expected answer templates, particularly in lower-capacity variants. Interestingly, lower-capacity models trained on interleaved image-text datasets can achieve comparable or better performance than larger models trained on captioning datasets when augmented with ICL and CoT. Finally, models show a preference for demonstration examples that are textually similar to the query, suggesting that textual coherence plays a crucial role in enhancing model performance across different settings. These findings highlight the efficacy of ICL and CoT in improving MLLM performance across various tasks and dataset configurations.

Secondly, the development of a zero-shot foiling benchmark for VidLMs has been a pivotal contribution. This benchmark is specifically designed to assess the models' ability to recognize actions and actors within a scene. It includes a proficiency test alongside the main test, providing a range of difficulty levels for evaluation. Experiments revealed that current VidLMs fail to adequately identify actions and actors, with their performance only slightly better than random chance. This finding indicates the models' significant shortcomings in temporal reasoning and action recognition.

In outlining the future directions of this study, we propose optimizing the identified strategies and exploring additional methods to bolster model robustness and reasoning capabilities in MLLMs. Enhancing sophisticated reasoning mechanisms promises to render these models more versatile across diverse tasks and environments. A crucial focus will be on analyzing the performance nuances of models trained on interleaved image-text datasets under various pretraining strategies. Moreover, investigating why models exhibit a preference for textual over visual similarity in examples presents an intriguing avenue for exploration, potentially refining model training strategies accordingly. Addressing challenges with chain-of-thought reasoning through the integration of fine-grained reasoning data also stands out as a critical area for improvement. These efforts aim to push the boundaries of multimodal learning, fostering advancements that could broaden the applicability and effectiveness of MLLMs across different domains.

Furthermore, the insufficient performance of cutting-edge VidLMs on benchmarks demanding temporal reasoning should encourage researchers to build more advanced techniques. Developing video-language modeling algorithms to succeed at spatio-temporal tasks is critical for future advancement in this discipline.

In conclusion, this thesis has shed light on critical aspects of VidLMs and MLLMs performance, offering valuable insights and identifying areas for future enhancement. The findings highlight the importance of continued innovation in multimodal architectures to achieve more robust and contextually aware language models capable of effectively bridging the gap between visual content and natural language.

REFERENCES

[1] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL:2022:long*, pages 8253–8280. acl, Dublin, Ireland, **2022**. doi:10.18653/v1/2022.acl-long.567.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, **2017**.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587. **2014**.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, **2020**.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, **2017**.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, **2018**.

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

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, **2019**.

[9] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, **2022**.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, **2019**.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, **2020**.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, **2014**.

[13] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, R Zemel, and Yoshua Bengio. Neural image caption generation with visual attention. In *Proc. ICML*, volume 37, pages 2048–2057. **2015**.

[14] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, **2015**.

[15] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5589–5600. **2021**.

[16] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433. **2015**.

[17] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, **2014**.

[18] REN Meng-ye, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Litoral Revista De La Poesía Y El Pensamiento*, 6:8–31, **2015**.

[19] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*, **2015**.

[20] Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553. **2016**.

[21] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer, **2010**.

[22] Tom F. H. Runia, Cees G. M. Snoek, and Arnold W. M. Smeulders. Real-world repetition estimation by div, grad and curl. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. **2018**.

[23] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850. **2017**.

[24] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, **2018**. ISBN 978-1-57735-800-8.

[25] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216. **2019**.

[26] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*, **2020**.

[27] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. **2021**.

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

[29] Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35:20450–20468, **2022**.

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, **2021**.

[31] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, **2021**.

[32] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133. **2022**.

[33] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650. **2022**.

[34] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, **2021**.

[35] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173. **2021**.

[36] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040. **2022**.

[37] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning. *arXiv preprint arXiv:2201.12723*, **2022**.

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

[39] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, **2022**.

[40] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, **2019**.

[41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, **2019**.

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, **2019**.

[43] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647. **2023**.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, **2015**.

[45] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, **2022**.

[46] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *Advances in Neural Information Processing Systems*, 36, **2024**.

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, **2014**.

[48] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 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. **2015**.

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

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

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

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

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

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

[55] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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. **2018**.

[56] Zhanzhan Cheng, Jing Lu, Yi Niu, Shiliang Pu, Fei Wu, and Shuigeng Zhou. You only recognize once: Towards fast video text spotting. In *Proceedings of*

the 27th ACM International Conference on Multimedia, page 855–863. ACM, **2019**.

[57] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*. **2021**.

[58] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, **2021**.

[59] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975. **2022**.

[60] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913. **2017**.

[61] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223. **2017**.

[62] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204. **2019**.

[63] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can

read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326. **2019**.

[64] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. In *European Conference on Computer Vision*. **2020**.

[65] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617. **2018**.

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

[67] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*. **2022**.

[68] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA, **2013**.

[69] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, **2015**. doi:10.1007/s11263-015-0816-y.

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

[71] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, **2023**.

[72] Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. The bla benchmark: Investigating basic language abilities of pre-trained multimodal models. *arXiv preprint arXiv:2310.15061*, **2023**.

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

[74] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, abs/2305.03726, **2023**.

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

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

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

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

[79] Salesforce AI Research. xgen-mm-phi3-mini-instruct model card, **2024**.

[80] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, **2024**.

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

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

[83] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, **2023**.

[84] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, **2023**.

[85] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition

and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, **2024**.

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

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

[88] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*. **2021**.

[89] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, **2020**.

[90] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755. **2020**.

[91] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, **2021**.

[92] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, **2021**.

[93] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968. **2022**.

[94] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, **2022**.

[95] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963. **2022**.

[96] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, **2022**.

[97] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, **2021**.

[98] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 413–430. Springer, **2022**.

[99] Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy III. Touch-gs: Visual-tactile supervised 3d gaussian splatting. *arXiv preprint arXiv:2403.09875*, **2024**.

[100] Reza Pebdani Babadian, Karim Faez, Mahmood Amiri, and Egidio Falotico. Fusion of tactile and visual information in deep learning models for object recognition. *Information Fusion*, 92:313–325, **2023**.

[101] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. **2019**.

[102] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738. **2021**.

[103] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, **2011**.

[104] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, **2022**.

[105] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, **2022**.

[106] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4. **2021**.

[107] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *Advances in Neural Information Processing Systems*. **2022**.

[108] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, **2021**.

[109] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, **2020**.

[110] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, **2022**.

[111] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. **2002**.

[112] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. **2005**.

[113] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. **2004**.

[114] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575. **2015**.

[115] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12487–12496. **2019**.

[116] Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. Interpretable video captioning via trajectory structured localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6829–6837. **2018**.

[117] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631. **2018**.

[118] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: Multimodal memory modelling for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7512–7520. **2018**.

[119] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, **2020**.

[120] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *Advances in Neural Information Processing Systems*, 35:34405–34420, **2022**.

[121] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. **2022**.

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

[123] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36, **2024**.

[124] Yi-Syuan Chen, Yun-Zhu Song, Cheng Yu Yeo, Bei Liu, Jianlong Fu, and Hong-Han Shuai. Sinc: Self-supervised in-context learning for vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15430–15442. **2023**.

[125] Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F Yang, and Kai-Wei Chang. Metavl: Transferring in-context learning ability from language models to vision-language models. *arXiv preprint arXiv:2306.01311*, **2023**.

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

[127] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, **2023**.

[128] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030, **2022**.

[129] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,

Faisal Azhar, et al. Llama: Open and efficient foundation language models (2023). *arXiv preprint arXiv:2302.13971*, **2023**.

[130] Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. How do in-context examples affect compositional generalization? *arXiv preprint arXiv:2305.04835*, **2023**.

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

[132] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, **2021**.

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

[134] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, **2022**. doi:10.18653/v1/2022.emnlp-main.759.

[135] Hritik Bansal, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff, and Dan Roth. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. *arXiv preprint arXiv:2212.09095*, **2022**.

[136] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen,

et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, **2022**.

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

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

[139] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, **2022**.

[140] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, **2023**.

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

[142] Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. Let's think frame by frame: Evaluating video chain of thought with video infilling and prediction. *arXiv preprint arXiv:2305.13903*, **2023**.

[143] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*, **2023**.

[144] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, **2023**.

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

[146] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, **2024**.

[147] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962. **2023**.

[148] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, **2023**.

[149] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, **2022**.

[150] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, **2022**.

[151] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, **2022**.

[152] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, **2023**.

[153] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, **2022**.

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

[155] Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *arXiv preprint arXiv:2310.00647*, **2023**.

[156] Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*, **2022**.

[157] Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models, **2023**.

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

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

[160] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`, **2023**.

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

[162] AI@Meta. Llama 3 model card. **2024**.

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

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

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

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

[167] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, **2023**.

[168] Google Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, **2024**.

[169] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, **2023**.

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

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

[172] Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for VQA are image captions. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *NAACL:2022:main*, pages 1947–1963. acl, Seattle, United States, **2022**. doi:10.18653/v1/2022.naacl-main.142.

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

[174] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 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. **2021**.

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

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

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

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

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

[180] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647. **2022**.

[181] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16167–16176. **2022**.

[182] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for

generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16804–16815. **2022**.

[183] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10739–10750. **2023**.

[184] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, **2022**.

[185] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. *ArXiv*, abs/2302.00402, **2023**.

[186] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, **2023**.

[187] Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models, **2023**.

[188] OpenAI. Chatgpt: A large language model. `https://openai.com`, **2021**. Accessed on April 25th, 2023.

[189] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count

and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, **1975**.

[190] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, **1948**.

[191] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, **2019**.

[192] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. **2016**.

[193] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, **2017**.

[194] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321. **2018**.

[195] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211. **2022**.

[196] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, **2020**.

[197] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, **2019**.

[198] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731. **2019**.

[199] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, **2018**.

[200] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, **2018**.

[201] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ArXiv*, abs/2106.08254, **2021**.

[202] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308. **2017**.

[203] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, C. Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *ArXiv*, abs/2306.05425, **2023**.

[204] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, **2023**.

[205] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, **2019**.

6. APPENDIX

In this section, we present example model responses from our MLLM evaluation. Figures 6.1 to 6.7 illustrate the evaluation process of few-shot ICL. Here, demonstration examples are chosen based on their similarity to the query and are provided along with their ground truth answers before presenting the actual query image-text pair to the model. Similarly, Figures 6.8 to 6.14 demonstrate the evaluation of CoT reasoning combined with ICL. In this setup, CoT descriptions accompany the demonstration examples, offering detailed reasoning chains that 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 6.1 Example model predictions on instances from the *Existence* task, using 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 ❌

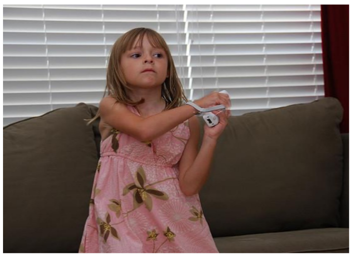
Idefics-9B I: Answer: Yes ❌

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 ✅

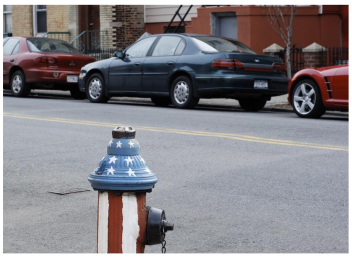
xGen-MM-4.6B-7B: Answer: Yes ✅

Qwen-VL-Chat-9.6B: Yes ✅

MMICL-12.1B: Yes ✅

Figure 6.2 Example model predictions on instances from the *Plurality* task, using 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 6.3 Example model predictions on instances from the *Counting* task, using 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` ✗

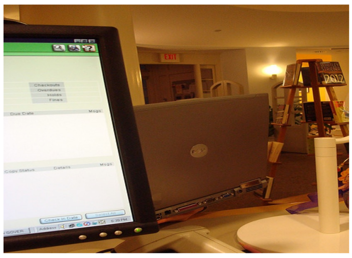
xGen-MM-4.6B-7B: `Answer: No` ✗

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` ✓

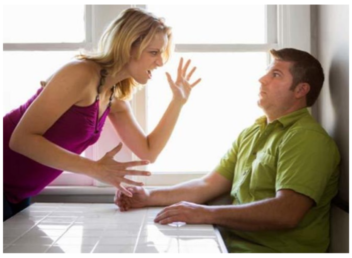
Figure 6.4 Example model predictions on instances from the *Spatial Relations* task, using demonstrations selected based on both visual and textual similarity (setting **S**).

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

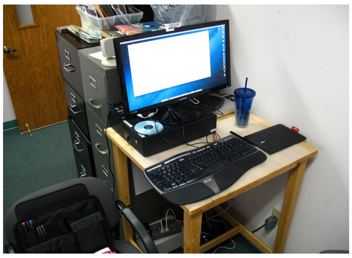
Does the following sentence describe the image?
A man confronts a man.

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 6.5 Example model predictions on instances from the *Actions* task, using 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 6.6 Example model predictions on instances from the *Coreference* task, using demonstrations selected based on both visual and textual similarity (setting **S**).

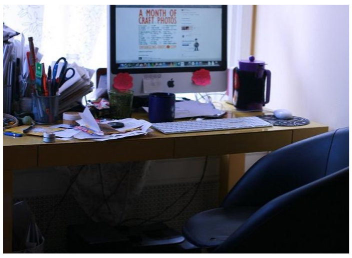
```
Does the following sentence describe the image?
A large commercial airplane parked on the
runway.
Answer: Yes
```

```
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
```



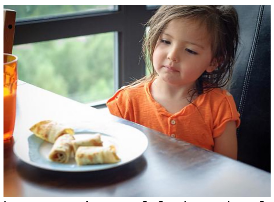
```
Does the following sentence describe the image?
A cluttered desk with a black chair next to it.
Answer: Yes
```

```
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 6.7 Example model predictions on instances from the *Foil-It!* task, using 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 6.8 Example model predictions on instances from the *Existence* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (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 ✅

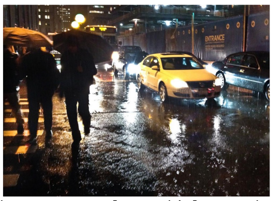
xGen-MM-4.6B-7B: The image shows a skateboarder in the middle of performing an action on his board. He is not turning it at all, but rather appears to be balancing or preparing for another trick while standing upright with one foot forward and holding onto both sides simultaneously (as indicated by "at 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 6.9 Example model predictions on instances from the *Plurality* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (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 6.10 Example model predictions on instances from the *Counting* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (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

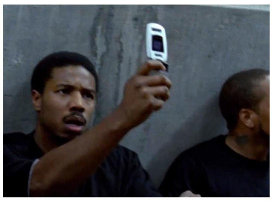
Idefics2-8B: The image shows a woman carrying bananas on her head. She is walking down a street, and the bananas are visible on her head. The woman is wearing a hat, and the bananas are piled on top of her head. The sentence provided states, "A lady walks down the street with bananas on her head. " This 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 6.11 Example model predictions on instances from the *Spatial Relations* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (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…

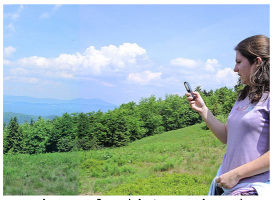
Qwen-VL-Chat-9.6B: The image shows a man lifting a barbell off the ground. The man is standing and holding the bar with both hands, while his torso is upright. The focus of the image is on the man and the barbell, so it is not accurate to say that the man is "lifting his torso." Instead, he is engaged 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 6.12 Example model predictions on instances from the *Actions* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (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 6.13 Example model predictions on instances from the *Coreference* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (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 6.14 Example model predictions on instances from the *Foil-It!* task, where demonstrations are chosen based on both visual and textual similarity, and Chain-of-Thought (CoT) reasoning is integrated (setting **S+C**).